Estadística aplicada uni y multivariante

Andrés Muñoz Serrano

Consejería de Agricultura y Pesca

Tomo III

JUNTA DE ANDALUCIA

ESTADÍSTICA APLICADA UNI Y MULTIVARIANTE

TOMO II

Andrés Muñoz Serrano
Departamento de Genética
Universidad de Córdoba

ESTADÍSTICA APLICADA UNI Y MULTIVARIANTE (2 volúmenes). Tomo II

© *Edita:* JUNTA DE ANDALUCÍA. ***Consejería de Agricultura y Pesca.***

Publica: Viceconsejería. Servicio de Publicaciones y Divulgación.

Autor: Muñoz Serrano, Andrés

I.S.B.N.: 84-8474-070-6 (Tomos I y II, obra completa)

Depósito Legal: SE. 3.996 (Obra completa)

Fotocomposición e impresión: J. de Haro Artes Gráficas, S.L. Parque Ind. P.I.S.A. Mairena del Aljarafe • Sevilla

CAPÍTULO 11

Regresión

CAPÍTULO 11

Regresión

Introducción.-

Con frecuencia se miden dos tipos de variables en cada individuo o unidad experimental. Un tipo de variable es la denominada ***variable respuesta*** o ***variable dependiente*** y la otra es la ***variable explicativa*** o ***variable independiente***. Por consiguiente, se necesita de otros análisis que expresen de una manera más precisa la naturaleza de las relaciones entre estos dos tipos de variables.

Cuando se tiene solo variables dependientes, los diferentes análisis parten del modelo esencial de que la observación o medida *i-ésima* de la variable *Y* es la suma de una constante o media de la población (μ) y una componente aleatoria o error (ε), lo que se ha expresado con el modelo lineal

$$Y_i = \mu + \varepsilon_i$$

al que se le añadían más componentes o factores de acuerdo con el modelo que mejor representara el problema establecido.

Cuando se tienen variables dependientes y variables independientes, se va a considerar que la *i-ésima* medida u observación de la variable *Y* es, también, la suma de una constante o media de la población y del error, pero en este caso la población de *Y* viene determinada por el valor *i-ésimo* de la otra variable, llamada variable *concomitante*, variable *relacionada*, *covariable* o variable *independiente*. Por tanto, la modificación que se introduce en el modelo anterior consiste en hacer explícito, en el nuevo modelo, que existe una *dependencia* entre los valores de las dos variables en cuestión, o dicho de otro modo, la variabilidad de la variable *Y* viene determinada en cierta medida por la variabilidad de la variable *X*, siendo este nuevo modelo

$$Y_i = \mu_{Y.X_i} + \varepsilon_i$$

que quiere decir que Y_i es función de la media poblacional de *Y* en la población determinada por el valor *i-ésimo* de *X*, más el error ε.

La diferencia entre los dos modelos consiste en que antes se tenía un sólo parámetro, μ, mientras que ahora se tiene i parámetros, μ_i, que dependen de los valores X_i asociados con ellos, es decir, que $\mu_{Y.Xi}$ es la media de Y dado el valor fijo X_i.

Este modelo permite considerar una posible dependencia entre las variables pero no especifica el tipo de relación existente. Para ello, se necesita asumir una de las infinitas expresiones de proporcionalidad que puede tomar $\mu_{Y.Xi}$.

Si se asume que en este modelo existe una relación de proporcionalidad del tipo

$$\mu_{Y.X_i} = \beta X_i$$

en el que $\mu_{Y.Xi}$ es igual al valor X_i por una constante de proporcionalidad, β, se tiene que sustituyendo esta última expresión en el modelo, quedaría

$$Y_i = \beta X_i + \varepsilon_i$$

Este modelo indica que X_i está medida sin error apreciable, mientras que la respuesta observada (Y_i) tienen un error

$$\varepsilon_i = Y_i - \mu_{Y.X_i}$$

que son los errores experimentales aleatorios e independientes con media cero y varianza constante.

Como se ha dicho en un párrafo anterior, existen prácticamente infinitas relaciones de proporcionalidad, algunas de las más comunes, y que se estudiaran más adelante en el capítulos 14, son

$$\begin{aligned} \mu_{Y.X_i} &= \beta^{X_i} \\ \mu_{Y.X_i} &= X_i^{\beta} \\ \exp^{(\mu_{Y.X_i})} &= X_i^{\beta} \end{aligned}$$

A la constante de proporcionalidad, β, se le denomina en estadística *coeficiente de regresión*, lo que nos lleva al concepto de *regresión* como la relación entre dos (o más) variables expresada como una *función lineal*. Como los valores de Y se obtienen de varias poblaciones, cada una determinada por el valor correspondiente de X, a la variable Y se le denomina variable *dependiente*, pues todo valor de Y depende de la población concreta muestreada, y a la variable X se le denomina variable *independiente*. Si se desea puntualizar que el coeficiente de regresión es de la variable Y sobre la variable X se escribe $b_{Y.X}$.

El coeficiente de regresión más utilizado es $Y_i = \beta X_i + \varepsilon_i$, en el que la relación de proporcionalidad es lineal. La mayor utilidad de este coeficiente de regresión lineal no es solo por su ajuste a multitud de procesos naturales, sino porque muchas

relaciones curvilíneas, como las vistas en la página anterior, se pueden linealizar por transformaciones logarítmicas y utilizar, entonces, la regresión lineal, como se estudiará más adelante.

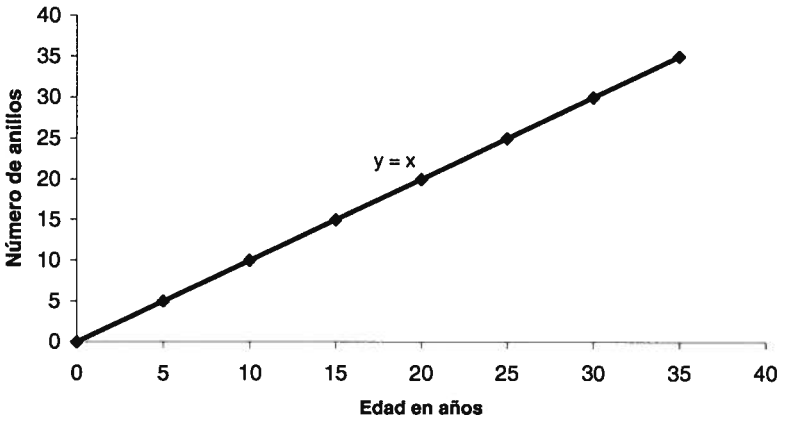
Relación funcional/Relación causal.-

En cualquier análisis de los dos tipos de variables se espera que la relación funcional hallada represente algún mecanismo básico asociado con los factores y variables investigados, sin embargo pocas veces el nivel de conocimientos es suficiente como para determinar una relación *causa-efecto*, y en todo caso no es la Estadística la encargada de determinar este tipo de relación. A causa de la incertidumbre en las variables y en los mecanismos básicos, hay que hacer constar que el supuesto de la existencia de una relación funcional y el que se haya encontrado una función que se ajusta bien a los datos observados, no presupone la existencia de una relación causal entre las variables. Esta relación causal, en el caso de existir, sólo la puede determinar el especialista en la materia en la que se realizó el experimento; el análisis estadístico es solamente un instrumento de ayuda en el análisis e interpretación de los datos.

Se puede poner como ejemplo un caso histórico en el que, en cierta localidad alemana, se observó que el aumento anual de la población humana de dicha localidad era una función lineal del aumento de la población de cigüeñas. Está claro que el crecimiento de la población humana no tiene su causa en la población de cigüeñas, sino que, más bien, ambas son efecto de una causa común no tenida en cuenta en este experimento.

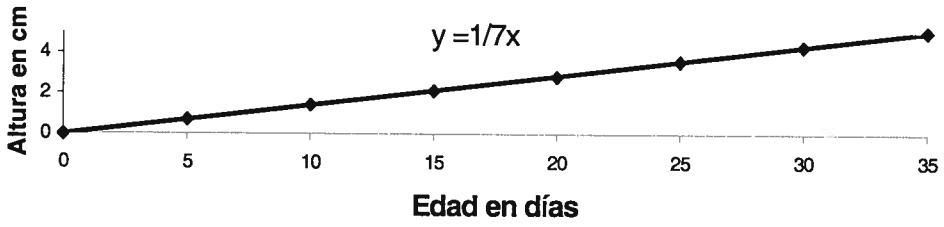
Regresión lineal.-

El tipo más simple de regresión sigue la ecuación $\mu_Y = X$, cuya línea se representa en esta gráfica



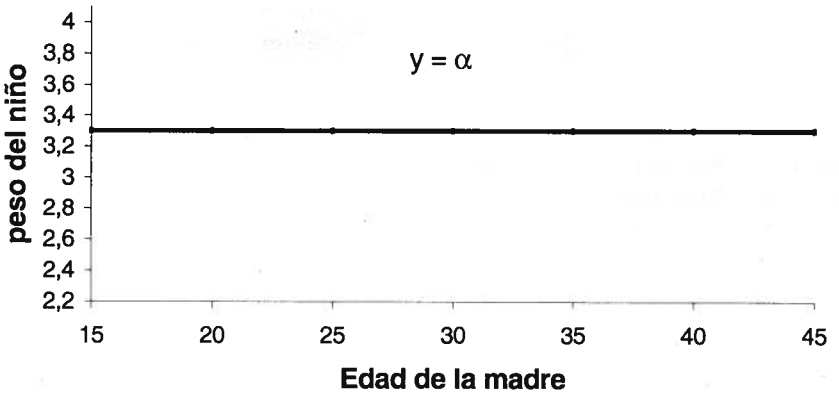
El ejemplo para esta función puede ser la relación entre el número de anillos de crecimiento de un árbol como una función de la edad (en años) del árbol. Sea cual sea la edad del árbol, el número de anillos tendrá el mismo valor. La línea descrita por la función pasa por el centro de coordenadas, lo cual significa que un árbol de cero años de edad tiene cero anillos, lo cual coincide con lo que se sabe de la fisiología del crecimiento de los árboles. Esto permite predecir con seguridad el número de anillos una vez conocida la edad de un árbol.

En la figura siguiente se muestra otra relación funcional dada por la ecuación $\mu_Y=(1/7)X$



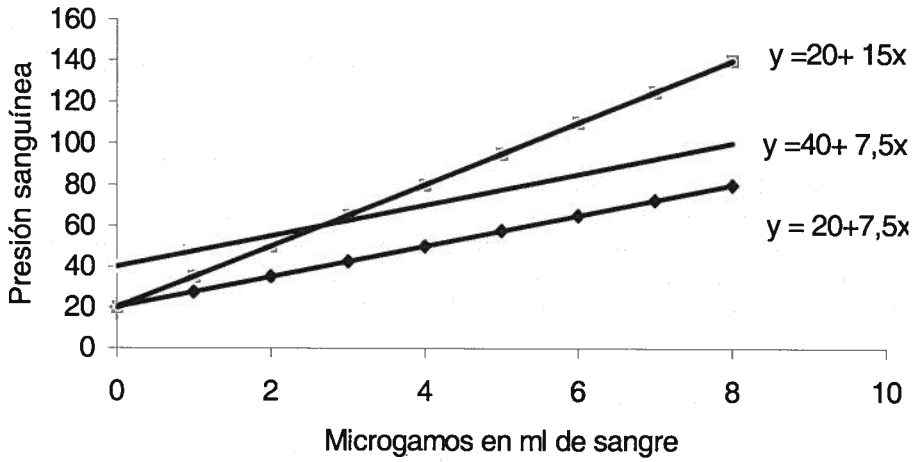
Este ejemplo puede ser el de la altura de una planta en centímetros en relación a su edad en días. Se tiene que la variable independiente está multiplicada por el coeficiente 1/7, es decir, que para un incremento de siete unidades de *X* habrá un incremento de una unidad de *Y*, por tanto, el promedio de altura de una planta de siete días será de un *cm*, y la altura promedio de una planta de 14 días será de dos *cm*. Este coeficiente de proporcionalidad, por tanto, nos está indicando la pendiente de la recta, es decir, cuanto mayor sea este coeficiente más inclinada será la recta y cuanto menor sea este coeficiente más horizontal será la recta. A este coeficiente, que nos indica la pendiente de la recta, es lo que en estadística se conoce como coeficiente de *Regresión*, que se simboliza con una β, el parámetro y como una *b* el estadístico. Nótese, así mismo, que en el ejemplo anterior el valor de este coeficiente era de uno, es decir, pendiente de la recta igual a uno que es la tangente de un ángulo de 45°, como el que forma dicha recta con el eje *X*. En este segundo ejemplo también para *X*=0, la variable dependiente también es igual a cero, lo que resulta razonable, pues una planta con cero días de edad tendrá cero *cm* de altura.

Otra situación puede ser la de la función $\mu_Y = \alpha$, es decir, una recta paralela al eje de las *X* como la siguiente



La pendiente de la recta en este caso es cero. Este ejemplo puede ser el del peso de los niños al nacimiento en función de la edad de la madre. En este caso se observa que el valor medio que tome *Y* es independiente del valor que toma *X*, por lo que se tiene una recta paralela al eje de las *X*. Esta recta no pasa, lógicamente, por el centro de coordenadas, es decir, no es una línea superpuesta al eje *X* pues el peso promedio al nacimiento es de 3.3 *Kg*, independientemente de la edad de la madre. Aparece, pues, un nuevo parámetro, α, que representa el valor de μ_Y para $X=0$, lo que se denomina *ordenada en el origen*

Otra situación puede ser la descrita en la siguiente figura



en donde se ilustra el efecto de dos compuestos sobre la presión sanguínea en dos especies de animales domésticos. La relación descrita en este gráfico puede expresarse mediante la fórmula o ecuación lineal $\mu_Y = \alpha + \beta X$.

La recta superior de la última gráfica representa la relación $\mu_Y = 20 + 15X$, que es el efecto medio del compuesto *A* sobre el animal *P*. La cantidad del compuesto se mide en microgramos, y la presión sanguínea en milímetros de *Hg*. A partir de la

ecuación dada, es fácil calcular la presión sanguínea esperada en el animal después de una dosis de 9 μg del compuesto, la presión sanguínea sería Y=20+15*9=155 *mm Hg*.

Como se puede ver, cuando la variable independiente es igual a cero, la variable dependiente no es igual a cero, sino que tiene el valor α, es decir, 20 *mm Hg* en la primera ecuación, que es el promedio de la presión sanguínea normal de la especie *P* en ausencia de compuesto. No sería razonable pensar que el animal no tiene presión en ausencia de compuesto. A este valor de α es lo que se conoce como *ordenada en el origen*.

Las otras dos funciones de la gráfica anterior muestran los efectos al variar los dos parámetros α y β. En la línea inferior $\mu_Y = 20 + 75X$, la ordenada en el origen es la misma, pero la pendiente se ha reducido a la mitad. Dicha función representa el efecto medio de un compuesto diferente, el *B*, sobre el mismo organismo *P*. Obviamente, cuando no se administra compuesto, la presión sanguínea debería ser la misma en el origen, dado que el organismo estudiado en este caso es el mismo al anterior. Sin embargo, compuestos diferentes es probable que ejerzan efectos hipertensos diferentes, como queda reflejado por la diferente pendiente de las rectas. La tercera relación describe el efecto del compuesto *B*, pero el experimento se realiza ahora sobre una especie diferente, la *Q*, cuya presión sanguínea normal es de 40 *mm Hg*, De esta forma, la ecuación para el efecto del compuesto *B* sobre la especie *Q* se escribe como $\mu_Y = 40 + 75$. Esta pendiente es la misma que la que produce el mismo compuesto sobre la otra especie, como consecuencia de que al ser el mismo compuesto, el efecto es el mismo sobre la presión sanguínea.

Como es fácil ver, la ecuación general es $\mu_Y = \alpha + \beta X$ o $Y=\alpha+X\beta+\varepsilon$, siendo las ecuaciones de las tres gráficas primeras, respectivamente: $\mu_Y = 0 + 1X$, $\mu_Y = 0 + 1/7\ X$ y $\mu_Y = 33 + 0X$.

Se ha comenzado con la línea recta o regresión lineal por simplicidad. A menudo se escoge una recta como aproximación porque se ajusta razonablemente bien en el intervalo de *X* en cuestión, aún cuando se sepa que la verdadera forma no es recta; tal es el caso del segundo ejemplo, en el que se vio que el crecimiento de una planta en *cm* por día, se podía explicar con una recta si los días son pocos, porque es obvio que una planta no crece linealmente toda su vida.

Estima de la recta de regresión por mínimos cuadrados.-

En cualquier situación real de observaciones bivariantes no se tendrán datos que se extiendan a la perfección a lo largo de una línea (recta o no). Aún en ejemplos en que sí debiera haber una recta perfecta, como puede ser medir la temperatura en escala *Fahrenheit* y en escala *Celsius*, habrá variación debido a los errores de medida de *Y* y errores debidos a factores ambientales impredecibles; esto impide una relación perfecta aún cuando exista una relación funcional entre variables.

Si se representan los *n* pares de valores (*X, Y*) de una experiencia en un sistema de coordenadas rectangulares, se tendrá una nube de puntos denominada

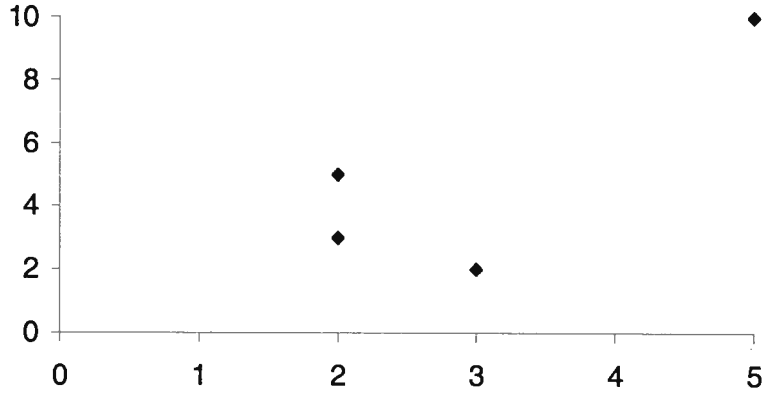
diagrama de dispersión. Puesto que las observaciones reales no van a caer exactamente sobre una línea recta, el problema consiste en trazar una línea recta a través de estos puntos que represente la tendencia lineal de la nube de puntos. Para poder trazar esta recta hay que tener un criterio.

El criterio adoptado para trazar esta recta es el mismo que se adoptó en la Introducción para el cálculo de la varianza, este es el criterio de *mínimos cuadrados.* Este criterio viene a decir que la suma de los cuadrados de las distancias verticales de todos los puntos a la recta trazada sea mínima, o dicho de otro modo, que esta recta es la que pasa más cerca de todos los puntos.

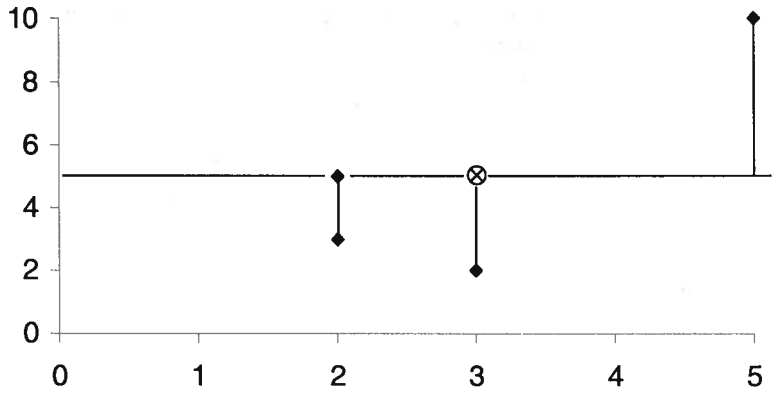
Para entender mejor las explicaciones que siguen supóngase una población de cuatro pares de valores

X	Y
3	2
2	3
2	5
5	10

si se representa en un eje rectangular se tendría

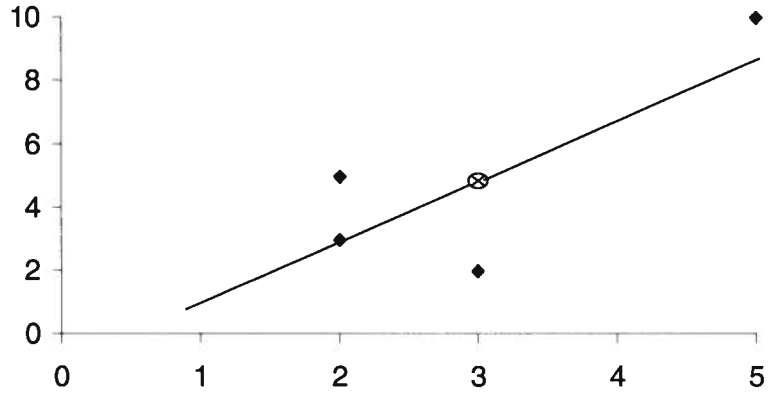


Si se traza una línea horizontal que pase por el punto $(\overline{X}, \overline{Y})$ se tendría



La suma de las distancias verticales de todos los puntos a esta línea es cero, pues es la suma de las diferencias de todos los valores de *Y* con su media (ver el epígrafe *Varianza* de la Introducción). Esta suma de las desviaciones a cualquier otra línea horizontal que no pase por $\overline{Y}$ no será cero, y la suma de los cuadrados de estas desviaciones será mayor que la suma de los cuadrados de la recta horizontal que se tracen por $(\overline{X},\overline{Y})$. Esta suma de cuadrados no es sino el numerador de la varianza de la variable *Y*, es decir la $SC_{(Y)}$ que se ha estado utilizando en todos los capítulos anteriores.

Si ahora se hace girar esta recta buscando la tendencia de los puntos y tomando como centro de giro el punto $(\overline{X},\overline{Y})$, las suma de las distancia seguirán siendo cero, pero la suma de los cuadrados de las distancias irán disminuyendo hasta llegar a una recta para la cual la suma de los cuadrados de las distancias verticales de todos los puntos sea mínima. Esta recta sería la siguiente



Si se sigue girando la recta, esta suma de cuadrados comienza de nuevo a aumentar.

Por lo tanto, si existe alguna tendencia lineal, la suma de los cuadrados de las distancias a la recta que es mínima, será siempre menor que la $SC_{(Y)}$. En el caso de

que no exista relación entre ambas variables, la suma de los cuadrados de las distancias a la recta que es mínima será igual a $SC_{(Y)}$, pues esta recta será horizontal, como ya se ha visto anteriormente.

La recta de regresión tiene, por tanto, las siguientes propiedades:

1) El punto $(\overline{X}, \overline{Y})$ se encuentra sobre la recta de regresión muestral.
2) La suma de las desviaciones a la recta de regresión es 0
3) La suma de cuadrados de los residuos es mínima. Esto es, si se reemplaza la recta de regresión calculada por cualquier otra recta, la suma de cuadrados del nuevo conjunto tendrá un valor mayor.

Modelo y supuestos de la regresión.-

Por definición, la verdadera regresión de *Y* con respecto a *X* consiste en las medias de las poblaciones de valores de *Y* (μ_Y), donde una población está determinada por el valor de *X*. Una línea de regresión no tiene que ser recta. En el muestreo es necesario suponer la forma de la línea.

El modelo matemático que nos define una observación es

$$Y_i = \mu_{Y \cdot X_i} + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

como la media paramétrica de *Y* es

$$\mu = \alpha + \beta \overline{X}$$

se tiene que α es

$$\alpha = \mu - \beta \overline{X}$$

que sustituyendo en el modelo lineal se tiene

$$Y_i = \alpha + \beta X_i + \varepsilon_i = \mu - \beta \overline{X} + \beta X_i + \varepsilon_i = \mu + \beta (X_i - \overline{X}) + \varepsilon_i$$

donde α y β son los parámetros que hay que estimar.

El modelo de regresión puede ser el *Modelo I* o con los valores de *X* fijos, y el *Modelo II* o con los valores de *X* aleatorios.

El *Modelo I* o de efectos fijos, es el mas común por ser el más adecuado a las situaciones experimentales que normalmente se presenta. Este modelo se basa en los siguientes supuestos paramétricos.

Supuesto 1. La variable independiente *X* se mide sin error, es decir, las *X* son fijas. Esto quiere decir que es el experimentador el que selecciona los valores de *X*, no

hay variación muestral aleatoria. En cambio, los valores de *Y* deben ser aleatorios con distribución normal de media $\mu_{Y.X}$ y varianza σ^2_Y. Nótese que esta suposición solamente implica que *Y* es una variable aleatoria que depende de *X*, y no toma en cuenta la forma lineal de la ecuación. La selección de los *X* puede inducir un conjunto específico de valores o simplemente valores que se encuentran dentro de un intervalo deseado. La respuesta a un compuesto para subir la presión sanguínea, por ejemplo, puede medirse mediante una serie de diluciones concretas, tal como las de la cuarta gráfica del presente capítulo; si se pretende saber los valores esperados, se usaran los mismos *X* en la repetición del muestreo.

Existe una relación muy estrecha entre este *Modelo I* (de *X* fijos) y el *Modelo I* o de efectos fijos del *ANOVA*, como se verá más adelante.

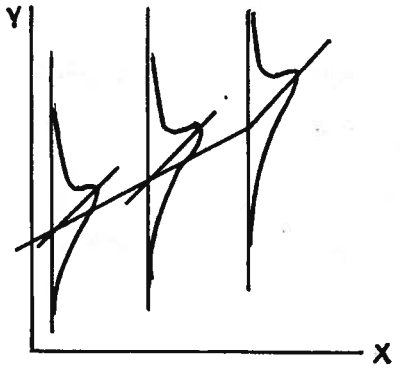
Los valores de la variable independiente, por ejemplo, horas de luz, niveles de temperatura, cantidades de tratamientos, etc, pueden estar espaciados a la misma distancia o de otra forma que el investigador crea conveniente en la realización del experimento.

Supuesto 2. El valor esperado de la variable *Y* se describe por la función, $\mu_Y = \alpha + \beta X$, ya descrita anteriormente. Ya se sabe que esto quiere decir que las medias poblacionales de los valores de *Y* (μ_Y) son una función de *X* y se extienden sobre una línea recta.

Si se realiza una observación aleatoria de *Y* para $X = X_i$ correctamente identificada, entonces el experimentador puede proceder a usar esta información correctamente. Pero si se registra incorrectamente el valor de X_i entonces los cálculos subsiguientes son incorrectos porque la población de *Y* está identificada incorrectamente.

Supuesto 3. Para cualquier valor dado de *X*, los valores de *Y* están distribuidos normal e independientemente. Esto equivale a suponer que la variable aleatoria no observable ε, o error de medida de la *Y*, es normal con media cero. No ocurre lo mismo con la *X*, ya que es una variable no aleatoria susceptible de ser manipulada por el investigador.

La varianza observada de todos los valores de *Y* es, entonces, la suma de la varianza inherente al material biológico con el que se este experimentando más la varianza debida al error de medida. Por ejemplo, si se aplica una misma dosis de un preparado que aumenta la presión arterial a una serie de individuos está claro que la respuesta a la dosis de los diferentes individuos no va a ser la misma, sino que se obtendrá una distribución de frecuencias alrededor de un valor esperado consecuencia del error de medida y de la variabilidad individual en la respuesta a dicho fármaco, tal como se muestra en esta figura

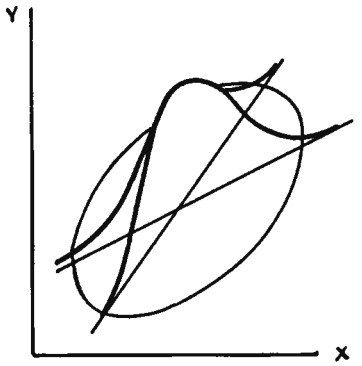


Naturalmente, es importante procurar el mínimo en los errores de medida.

No será normal que los experimentos de regresión tenga más de un valor de *Y* para cada valor de *X*. De hecho, los cálculos básicos que se verán dentro de un momento serán para un valor de *Y* por cada valor de *X*. Sin embargo, incluso en este caso, la hipótesis básica del *Modelo I* de regresión es que el valor único de *Y* correspondiente al valor dado de *X* es una muestra de una población de valores distribuida normal e independientemente.

Supuesto 4. Este supuesto ya se ha citado anteriormente, y es el mismo del *ANOVA*, este es el de la *homocedasticidad*, es decir, que la varianza de *Y* a lo largo de la línea de regresión es única, que es, en definitiva, la varianza de ε. Por tanto, la varianza a lo largo de la línea de regresión es constante e independiente de la magnitud de *X*. Nótese que esto equivale a suponer que la media de *Y* se modifica con el valor de *X*, pero la varianza se mantiene constante.

Para el *Modelo II*, tanto *X* como *Y* son aleatorios, es decir, que *X* no es fija y dependiente del control del experimentador, como se muestran en esta figura



Este es el problema clásico de regresión bivariante, en el que se supone normalidad pero en el que se tendrá dos distribuciones, una para cada variable, que probablemente serán diferentes. El muestreo aleatorio es de individuos en los cuales

se efectúan pares de medidas. Por ejemplo, se toman ratas y se mide el peso total del individuo y el peso del hígado; la elección de cuál es la variable dependiente y cuál es la variable independiente viene determinada por el problema. Son posibles dos líneas de regresión la de *Y* respecto a *X* y la de *X* respecto a *Y*. Para problemas de este tipo tal vez sean más adecuados los métodos relativos a la *correlación*, aunque muchas veces nos interesará describir la relación funcional entre tales variables, para lo cual se tiene que recurrir a técnicas especiales de regresión para el modelo II, que se estudiarán más adelante.

En la regresión lineal se supone que los ε se distribuyen normal e independientemente con una varianza común. Si es válido el supuesto de una varianza común, se puede aplicar el cuadrado medio del error residual (ε) para hacer inferencias probabilísticas válidas respecto a una media poblacional ($\mu_{Y.Xi}$), independientemente del valor de *X*. Este cuadrado medio se calcula a partir de las desviaciones respecto de la recta de regresión, también llamados residuos (ver siguiente epígrafe). Si las varianzas no son homogéneas, entonces es necesaria una regresión ponderada o una transformación de los datos de tal manera que las varianzas sean homogéneas. Este es el caso del análisis *probit* (ver más adelante) para porcentajes de mortalidad, donde la varianza es binomial.

Una vez estimados α y β, es posible estimar la media de una población de *Y* sin haber observado uno solo de los individuos. Por ejemplo, en un estudio semejante al de la cuarta gráfica que se propuso al principio del presente capítulo (página 395, gráfica superior), se puede estimar la media de la población de los *Y* para una concentración del compuesto *A* de *X*=45 $\mu g/ml$ en un individuo de la especie *P*, usando la ecuación de regresión ya conocida

$$\hat{Y}_{4.5} = 20+15\times 4.5 = 87.5$$

Se ha usado la anotación de $\hat{Y}$ y no de $\overline{Y}$ para distinguir lo que es una estima poblacional sin muestra de lo que es una estima poblacional con muestra, respectivamente, es decir, la primera es la estima de la media de la población de *Y* para *X*=45, mientras que lo segunda es la estima de la media de *Y*, para todo valor de *X*.

Las estimas de α y β se simbolizan como $\hat{\alpha}$ y $\hat{\beta}$ o más comúnmente como *a* y *b*.

Estima de la recta de regresión.-

Según el modelo, para estimar la recta de regresión se tiene que estimar α y β por medio de los estadísticos *a* y *b*, siguiendo el criterio de *mínimos cuadrados*, es decir, haciendo mínimo

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

Como

$$\hat{Y} = a+bX$$

sustituyendo en la expresión que hay que minimizar, se tiene

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2$$

Por cálculo diferencial de la expresión anterior se obtiene el sistema de ecuaciones *normales* siguiente

$$n\,a+\Sigma_i X_i\; b=\Sigma_i Y_i$$
$$\Sigma_i X_i\, a+\Sigma_i X_i^2\, b=\Sigma_i X_i Y_i$$

que resolviéndolas se obtiene

$$b = \frac{\Sigma_i X_i Y_i - \frac{\Sigma_i X_i \;\; \Sigma_i Y_i}{n}}{\Sigma_i X_i^2 - \frac{(\Sigma_i X_i)^2}{n}}$$
$$a = \overline{Y} - b\overline{X}$$

Por tanto, para la estima del coeficiente de regresión, *b*, se tiene en el denominador una expresión ya conocida, es la suma de cuadrados de la variable independiente, *X*, mientras que en el numerador se tiene una expresión muy semejante a la del denominador, esta es la *suma de productos* de ambas variables.

La *suma de los productos cruzados*, en analogía a la suma de los cuadrados, es la suma de los productos de las desviaciones de las observaciones de cada variable respecto de su media correspondiente. La fórmula definición es

$$SP = \Sigma_i (X_i - \overline{X})\,(Y_i - \overline{Y})$$

y la fórmula de cálculo práctico es

$$SP = \Sigma_i X_i Y_i - \frac{\Sigma_i X_i \Sigma_i Y_i}{n}$$

La semejanza entre una suma de los productos y una suma de cuadrados, tanto en la fórmula de definición como la práctica, es evidente. Si se reemplaza *Y* por *X* en la fórmula de la suma de los productos se obtiene la suma de cuadrados de *X*.

Cuando la *SP* se divide por los grados de libertad, recibe el nombre de *covarianza* por analogía con la varianza. Una covarianza es una medida de la variación conjunta de dos variables y puede ser positiva (si la variación es directa) o negativa (si

la variación es inversa). Es simétrica en X o en Y, y las variables no necesitan especificarse como dependientes o independientes.

Nótese que si la SP (o la covarianza) es negativa, la pendiente de la recta será negativa, es decir, que a medida que se incrementa X disminuye Y.

Por lo tanto el coeficiente de regresión es igual a la suma de los productos dividido por la suma de cuadrados de X. O bien es igual a la covarianza de XY dividido por la varianza de X.

Dado que la recta de regresión pasa por el punto $(\overline{X}, \overline{Y})$, una vez conocido b se puede hallar la ordenada en el origen, a, tal como se ha visto en la solución de las ecuaciones normales.

Por tanto, resumiendo, para la determinación de la recta de regresión se necesita realizar los cálculos de ΣX, ΣX^2, ΣY, ΣY^2 y ΣXY, cantidades que nos proveen las calculadoras de bolsillo cuando no nos proveen directamente de b y de a.

Ejemplo.-

Se tiene diez líneas de gallinas ponedoras que se caracterizan, entre otras cosas, por su diferente peso medio. Se toman 50 gallinas de cada línea, de manera que el peso medio, X, de cada grupo sea el de la línea; y se le mide, además, el consumo de pienso Y en 350 días. Se quiere saber la función que relaciona ambas variables.

	X	Y	X^2	Y^2	XY
	2.11	40.73	4.4521	1658.93	85.940
	1.99	41.45	3.9601	1718.10	82.486
	2.31	42.23	5.3361	1783.37	97.551
	2.67	45.13	7.1289	2036.72	120.497
	2.13	41.77	4.5369	1744.73	88.970
	2.31	42.82	5.3361	1833.55	98.914
	2.08	39.50	4.3264	1560.25	82.160
	2.31	43.31	5.3361	1875.76	100.046
	2.35	45.04	5.5225	2028.60	105.844
	2.22	42.36	4.9284	1794.37	94.039
Σ	22.48	424.34	50.8636	18034.39	956.448

$$SP = 956.4478 - \frac{22.48 \times 424.34}{10} = 2.5315$$

$$SC_{(X)} = 50.8636 - \frac{(22.48)^2}{10} = 0.3286$$

$$b = \frac{2.5315}{0.3285} = 7.704$$

$$a = 42.434 - 7.706 \times 2.248 = 25.115$$

El valor de *b* nos indica que para un aumento de una unidad del peso de la gallina, aumenta 7.704 unidades el consumo de pienso en 350 días. Por lo que la línea de regresión que explica la variabilidad del consumo de pienso en función del peso de las gallinas es

$$\hat{Y} = 25.115 + 7.704X$$

Fuentes de variación en la regresión.-

El modelo de regresión lineal, considera una observación como la suma de una media $\mu_{Y.X}=\alpha+\beta X$ y una componente aleatoria o error ε.

En términos de la regresión muestral, una observación Y está compuesta de una media muestral $\hat{Y}$ determinada por la recta de regresión y una desviación muestral o residuo $e = Y - \hat{Y}$ respecto a la media.

Como *a* es

$$a = \overline{Y} - b\overline{X}$$

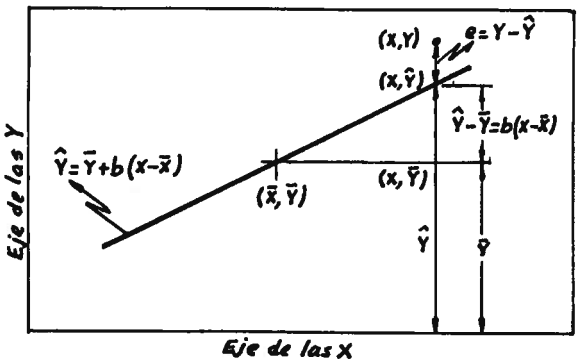
se puede sustituir en la expresión

$$\hat{Y} = a + bX$$

obteniendo

$$\hat{Y} = (\overline{Y} - b\overline{X}) + bX = \overline{Y} + b(X - \overline{X})$$

Estas componentes quedan expresadas gráficamente de la siguiente forma



Como se ve en la gráfica anterior, la desviación de $\hat{Y}$ con respecto a $\overline{Y}$ es debida a la regresión, esto es

$$\hat{Y} - \overline{Y} = b(X - \overline{X})$$

Mientras que el residuo

$$e = Y - \hat{Y}$$

es una estimación de la desviación aleatoria o error, ε, que también puede expresarse como $e_{Y.X}$.

De manera que la desviación total de una observación de la variable Y con respecto a la media es

$$Y - \overline{Y} = (\hat{Y} - \overline{Y}) + (Y - \hat{Y}) = b(X - \overline{X}) + e_{Y.X}$$

Es decir, la desviación total de la variable dependiente con respecto a su media es igual a la *desviación debida a la regresión* más la *desviación debida al error*. Por tanto

Desviación Total = Desviación Regresión + Desviación Error

Puesto que esta igualdad es cierta para todas las observaciones, se pueden sumar todas las observaciones y elevar al cuadrado ambos miembros de la anterior ecuación, con lo que se tiene

$$\Sigma_i (Y_i - \overline{Y})^2 = \Sigma_i \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})\right]^2$$

desarrollando el cuadrado de la suma

$$\Sigma_i (Y_i - \overline{Y})^2 = \Sigma_i (Y_i - \hat{Y}_i)^2 + \Sigma_i (\hat{Y}_i - \overline{Y})^2 + 2\Sigma_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})$$

el doble producto es cero, por lo que

$$\Sigma_i (Y_i - \overline{Y})^2 = \Sigma_i (Y_i - \hat{Y}_i)^2 + \Sigma_i (\hat{Y}_i - \overline{Y})^2$$

Por tanto, la suma de cuadrados total de Y puede descomponerse de acuerdo con estas fuentes de variación

$$SC_{(total)} = SC_{(error)} + SC_{(regresión)}$$

Las expresiones prácticas para el cálculo de estas sumas de cuadrados son

$$SC_{(regresión)} = \Sigma_i (\hat{Y}_i - \overline{Y})^2 = \Sigma_i (a_i + bX_i - \overline{Y})^2 = \Sigma_i (\overline{Y} - b\overline{X} + bX_i - \overline{Y})^2 =$$
$$= b^2 \Sigma_i (X_i - \overline{X})^2 = b^2 SC_{(X)}$$

como

$$b = \frac{SP_{(XY)}}{SC_{(X)}}$$

se tiene

$$SC_{(\text{regresión})} = b^2 SC_{(X)} = b\frac{SP_{(XY)}}{SC_{(X)}} sc_{(X)} = b\,SP_{(XY)} = \frac{SP^2_{(XY)}}{SC_{(X)}}$$

Con respecto al error se tiene

$$\begin{aligned} SC_{(\text{error})} &= \Sigma_i (Y_i - \hat{Y})^2 = \Sigma_i (Y_i - a - b X_i)^2 = \Sigma_i (Y_i - \overline{Y} + b\overline{X} - bX_i)^2 = \\ &= \Sigma_i [(Y_i - \overline{Y}) - b(X_i - \overline{X})]^2 = \\ &= \Sigma_i (Y_i - \overline{Y})^2 + b^2 \Sigma_i (X_i - \overline{X})^2 - 2b\Sigma_i (X_i - \overline{X})(Y_i - \overline{Y}) = \\ &= SC_{(Y)} + b^2 SC_{(X)} - 2b\,SP_{(XY)} = \\ &= SC_{(Y)} + b\frac{SP_{(XY)}}{SC_{(X)}} SC_{(X)} - 2b\,SP_{(XY)} = \\ &= SC_{(Y)} - b\,SP_{(XY)} = SC_Y - \frac{SP^2_{(XY)}}{SC_{(X)}} \end{aligned}$$

Por tanto la varianza del error es

$$\begin{aligned} S^2_{Y.X} &= \frac{SC_{(error)}}{n-2} = \frac{SC_{(Y)} - b\,SP_{(XY)}}{n-2} = \\ &= \frac{SC_{(Y)} - \dfrac{SP^2_{(XY)}}{SC_{(X)}}}{n-2} \end{aligned}$$

La suma de cuadrados atribuible a la media o término de corrección es

$$n\overline{Y}^2 = \frac{(\Sigma_i Y_i)^2}{n}$$

Y la suma de cuadrados total es, lógicamente

$$SC_{(Y)} = \Sigma_i Y_i^2 - \frac{(\Sigma_i Y_i)^2}{n}$$

La varianza de las estimaciones de *a*, *b*, $\overline{Y}$ y $(\hat{Y}.X_o)$ son

$$S_a^2 = \frac{S_{(Y.X)}^2 \Sigma_i X_i^2}{n\, SC_{(X)}}$$

$$S_b^2 = \frac{S_{(Y.X)}^2}{SC_{(X)}}$$

$$S_{(\overline{Y})} = \sqrt{\frac{S_{Y.X}^2}{n}}$$

$$S_{\hat{Y}.X_o}^2 = S_{Y.X}^2 \left(\frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}} \right);\ X_o \text{ Interpolado}$$

$$S_{e.X_o}^2 = S_{Y.X}^2 \left[1 - \left(1 + \frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}} \right) \right]$$

$$S_{\hat{Y}.X_o}^2 = S_{Y.X}^2 \left(1 + \frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}} \right);\ X_o \text{ extrapolado}$$

Resumiendo, la distribuciones de los diferentes parámetros son

$$a \sim N\left(\alpha,\ \frac{S_{Y.X}^2 \Sigma_i X_i^2}{n\, SC_{(X)}} \right)$$

$$b \sim N\left(\beta,\ \frac{S_{Y.X}^2}{SC_{(X)}} \right)$$

$$\hat{Y}.X_o \sim N\left[\mu_{Y.X_o} = \alpha + \beta X_o,\ S_{Y.X}^2 \left(\frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}} \right) \right]$$

Puesto que la *SC total* es la misma independientemente del modelo al que se ajuste, existirá una interdependencia entre las sumas de cuadrados de regresión y del error, y consecuentemente entre *b* y *SC error*. A mayor valor de *b*, menor *SC error*.

Prueba de ajuste.-

En el epígrafe anterior se ha realizado la descomposición de la suma de cuadrados de la variable dependiente en el análisis de regresión, de manera análoga a como se realizó con el *ANOVA*. Esta descomposición de la variación total puede usarse para realizar diferentes pruebas de hipótesis de los diferentes parámetros que se han descrito hasta el momento.

Al igual que en el *ANOVA* se puede realizar un análisis de la varianza de la descomposición de la suma de cuadrados, de la siguiente manera

FV	*gl*	*SC*	*CM*	F_o	*ECM*
Regresión	1	$b\,SP_{(XY)}$	$b\,SP_{(XY)}$	$\frac{b\,SP_{(XY)}}{S^2_{Y.X}}$	$\sigma^2+b^2CM_{(X)}$
Error	*n*-2	$SC_{(Y)}-bSP_{(XY)}$	$S^2_{Y.X}$		σ^2
Total	*n*-1	$SC_{(Y)}$			

Los grados de libertad de la componente debida a la regresión es el número de parámetros estimados menos uno. Como se han estimado α y β, se tiene un grado de libertad.

El cuadrado medio debido a la regresión mide la cantidad de variación de *Y* que es debida a *X*.

De manera que para las hipótesis

$$H_o : \beta = 0$$
$$H_1 : \beta \neq 0$$

el estadístico de prueba será

$$F_o = \frac{CM_{(\text{regresión})}}{S^2_{Y.X}}$$

La regla de decisión consiste en rechazar H_o si $F_o \geq F_{(1,n-2;\alpha)}$. Resulta evidente que si F_o es grande lo es porque una porción grande y significativa de la varianza de *Y* ha quedado explicada por la regresión sobre *X*; por lo que se concluirá que el ajuste a la recta es bueno, es decir, que una parte significativa de la variabilidad de *Y* es debida a la variabilidad de *X*.

Coeficiente de determinación y coeficiente de alineación o factor de mejoramiento.-

Como se estudió en el Capítulo 5, la razón de la suma de cuadrados debida al modelo por la suma de cuadrados total, es el denominado *coeficiente de determinación* que se simboliza como r^2 o más comúnmente como R^2

$$R^2 = \frac{\frac{SP^2}{SC_{(X)}}}{SC_{(Y)}}$$

Por tanto, este coeficiente indica la proporción de la suma de cuadrados total de *Y* que es atribuible al ajuste del modelo, es decir, a la regresión. Es una medida de bondad de ajuste al modelo.

El campo de variación de este coeficiente esta entre el 0 y el +1. No puede ser negativo, independientemente de que lo sea *b*, pues todas las sumas de cuadrados son positivas y el numerador es una expresión elevada al cuadrado. Y no puede ser mayor de uno puesto que la suma de cuadrados explicada de cualquier variable tiene que ser menor que su suma de cuadrados total, o, en caso extremo, si explica toda la variación de una variable, puede ser tan grande como la suma de cuadrados total, pero no mayor.

Este coeficiente de determinación puede indicar ajustes casi perfectos (valores cercanos al uno) introduciendo en el modelo más variables (ver más adelante la regresión múltiple) aunque estas variables sean superfluas y realmente no mejoren el ajuste; para evitar este problema, se puede ajustar el coeficiente de determinación para el número (*m*) de coeficientes de regresión estimados, siendo este ajuste

$$R^2_{adj} = 1 - \left(\frac{(1 - r^2)(n-1)}{n-m-1} \right)$$

Este R^2 ajustado tiende a estabilizarse en un cierto valor cuando se introducen variables adecuadas al ajuste. Aunque para un solo coeficiente de regresión, ambos valores son próximos.

Hay que hacer notar que el coeficiente de determinación ajustado puede tener valores negativos.

Puesto que el valor del coeficiente de determinación oscila entre 0 y +1, se le puede restar a 1 con el fin de poder calcular la fracción de la suma de cuadrados explicada y no explicada, es decir, la fracción de la suma de cuadrados total de *Y* explicada por la variabilidad de *X* (debida a la regresión) es

$$SC_{(\text{regresión})} = \frac{SP^2}{SC_{(Y)}} = r^2 SC_{(Y)}$$

Y la suma de cuadrados no explicada por *X* es

$$SC_{(\text{error})} = SC_{(Y)} - \frac{SP^2}{SC_{(Y)}} = (1 - r^2) \; SC_{(Y)}$$

La cantidad $1\text{-}r^2$ se denomina *coeficiente de indeterminación* o de no determinación y expresa la proporción de la varianza de una variable que no ha sido explicada por la otra variable. La raíz cuadrada de este coeficiente de indeterminación

$$\sqrt{1 - r^2}$$

se denomina *coeficiente de alineación* o *factor de mejoramiento* y mide la falta de asociación entre las variables *X* e *Y*.

Tanto el coeficiente de *determinación* como de *indeterminación* como el coeficiente de *alineación* son cantidades o proporciones indicativas de lo que denomina su nombre, pero no son estadísticos que permitan hacer inferencias.

Sin embargo, una estima insesgada de la varianza no explicada por la regresión con la que si se puede realizar inferencias es el *CM* del residuo o error, que se simboliza como $S^2_{Y.X}$, es decir, la fracción de la suma de cuadrados total de *Y* no explicada por la variabilidad de *X* dividida por sus grados de libertad (*n*-2). A su raíz cuadrada se le denomina *error típico o desviación típica de Y para X fijo* o bien se le denomina *desviación típica o error típico de Y manteniendo constante X.*

Ejemplo.-

Siguiendo con el mismo ejemplo de los pollos.

	X	*Y*	X^2	Y^2	*XY*
	2.11	40.73	4.4521	1658.93	85.940
	1.99	41.45	3.9601	1718.10	82.486
	2.31	42.23	5.3361	1783.37	97.551
	2.67	45.13	7.1289	2036.72	120.497
	2.13	41.77	4.5369	1744.73	88.970
	2.31	42.82	5.3361	1833.55	98.914
	2.08	39.50	4.3264	1560.25	82.160
	2.31	43.31	5.3361	1875.76	100.046
	2.35	45.04	5.5225	2028.60	105.844
	2.22	42.36	4.9284	1794.37	94.039
Σ	22.48	424.34	50.8636	18034.39	956.4478

$$SP = 956.4478 - \frac{22.48 \times 424.34}{10} = 2.5315$$

$$SC_{(X)} = 50.8636 - \frac{(22.48)^2}{10} = 0.3286$$

$$SC_{(Y)} = 18034.39 - \frac{(424.34)^2}{10} = 27.9464$$

$$b = \frac{2.5315}{0.3286} = 7.704$$

$$a = 42.434 - 7.704 \times 2.248 = 25.115$$

$$b\,SP_{(XY)} = 7.704 \times 2.5315 = 19.5027$$

$$SC_{(Y)} - b\,SP_{(XY)} = 27.9464 - 19.5027 = 8.4437$$

FV	*gl*	*SC*	*CM*	F_o
Regresión	1	19.5027	19.5027	18.477***
Error	8	8.4437	1.0555	
Total	9	27.9464		

Por lo que se puede concluir que la variación de *X* contribuye a la variación de *Y*, es decir, que *b* es significativamente diferente de cero. Y la recta de regresión que define esta función lineal es

$$Y = 25.115 + 7.707X$$

También se puede observar (como se hizo en el *ANOVA*) que el coeficiente de determinación, es decir, la proporción de la suma de cuadrados total de *Y* atribuible a la suma de cuadrados de *Y* explicada por la variación de *X* es

$$r^2 = \frac{19.5027}{27.9464} = 0.6978$$

es decir, del 70%.

El valor del coeficiente de determinación ajustado en este ejemplo, es

$$r^2_{adj} = 1 - \left(\frac{(1 - 0.6978)9}{8} \right) = 0.66$$

En este caso, al haber un solo coeficiente de regresión, ambos valores son próximos.

Continuemos ahora con los errores típicos de diversos estadísticos de regresión, su empleo para comprobar hipótesis y el cálculo de límites de confianza.

Pruebas de hipótesis e intervalos de confianza para el coeficiente de regresión.-

La varianza debida a la regresión es el cociente de la varianza no explicada divida por la suma de cuadrados de *X*, y el *error típico del coeficiente de regresión* es la raíz cuadrada de este valor

$$S_b = \sqrt{\frac{S^2_{Y.X}}{SC_{(X)}}}$$

Con este error típico se puede probar diversas hipótesis y estimar intervalos de confianza para *b*.

Se puede probar cualquier hipótesis. Por ejemplo, para probar la hipótesis nula de que el valor muestral de *b* proviene de una población con un valor paramétrico del coeficiente de regresión, β_o, se realiza por medio de la *t* siguiente

Cola derecha	Cola izquierda	Dos colas
$H_o: \beta \leq \beta_o$	$H_o: \beta \geq \beta_o$	$H_o: \beta = \beta_o$
$H_1: \beta > \beta_o$	$H_1: \beta < \beta_o$	$H_1: \beta \neq \beta_o$

$$t_o = \frac{b - \beta_o}{S_b}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n-2;\, \alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-2;\, \alpha)}$ para las hipótesis de una cola.

Ejemplo.-

Siguiendo con el mismo ejemplo de los pollos, probar si $\beta>7$.

Si $\beta=0$

$$S_b = \sqrt{\frac{1.0555}{0.3286}} = 1.7922$$

$$H_o: \beta \leq 7$$
$$H_1: \beta > 7$$
$$t_o = \frac{7.704 - 7}{1.7922} = 0.3928\ ns$$
$$t_{(8,\, 0.05)} = 1.8595$$

Se acepta la hipótesis nula, por lo que se puede concluir que, en la población en donde se tomó esta muestra, por cada unidad de aumento de peso de los pollos, aumenta siete veces el pienso consumido en 350 días, pero no más de siete veces.

También se puede probar si es significativa la pendiente

$$H_o: \beta = 0$$
$$H_1: \beta \neq 0$$
$$t_o = \frac{7.704 - 0}{1.7922} = 4.2986\ **$$
$$t_{(8,\, 0.01/2)} = 3.3554$$

Se rechaza la hipótesis nula, por lo que se puede concluir que el coeficiente de regresión de la población de donde se tomó esta muestra es significativamente diferente de cero.

Constátese, así mismo, que el valor de t_o para la hipótesis de $\beta=0$ es la raíz cuadrada del valor de F del *ANOVA* para la prueba de bondad de ajuste.

Siguiendo métodos similares a los ya estudiados resulta fácil calcular los límites de confianza para β, de la siguiente manera

$$LC_{(\beta)} = b \pm S_b t_{(n-2,\, \alpha/2)}$$

Ejemplo.-

Para los datos de los pollos de carne, los límites de confianza al 95% son

$$LC_{(\beta)} = 7.704 \pm 1.7922 \times 2.306 = 7.704 \pm 4.1328$$
$$L_i = 3.571$$
$$L^s = 11.837$$

Por lo que se concluye que β está entre 3.571 y 11.837 unidades de alimento por unidad de peso del pollo con un 95% de confianza. Como se ve, estos límites o intervalos de confianza no incluyen al cero, por lo que se puede concluir, así mismo, que $\beta>0$.

Estos límites de confianza permitirían trazar otras dos rectas que se cruzarían entre si y con la recta de regresión en el punto $(\overline{X}, \overline{Y})$.

Pruebas de hipótesis e intervalos de confianza de la estimación, por interpolación, de un valor *Y*.-

El *error típico de un valor estimado* $\hat{Y}$, para un valor dado de X_o dentro del rango de las X muestreadas en el experimento es

$$S_{\hat{Y}.X_o} = \sqrt{S^2_{Y.X}\left(\frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}}\right)}$$

Como se ve, la varianza del error se ve incrementada en un factor que es proporcional a la distancia que exista entre la X_o y la media $\overline{X}$, por lo que cuanto mayor sea la distancia entre X_o y su media mayor será el error típico de la estima de $\hat{Y}$ y mayor, por tanto, el intervalo de confianza, que sería

$$LC_{(\hat{Y}.X_o)} = \hat{Y} \pm S_{(\hat{y}.X_o)} t_{(n-2,\alpha/2)}$$

Si se estiman los límites de confianza de $\hat{Y}$ para todos los puntos X_o del experimento, y se representan estos puntos en la gráfica donde se tiene la recta, se obtendrá dos curvas cóncavas a ambos lados de la recta. Es decir, a medida que nos

alejamos de la media, menos fiable son las estimas de Y como consecuencia de la incertidumbre de la verdadera pendiente de β.

Ejemplo.-

Para los datos de los pollos, el intervalo de confianza al 95% del valor estimado $\hat{Y}$ para X_o=2 es

$$\hat{Y} = 25.115+7.704\times 2 = 40.523$$

$$S_{(\hat{Y}.2)} = \sqrt{1.0555\left[\frac{1}{10}+\frac{(2-2.248)^2}{0.3286}\right]} = 0.5506$$

$$LC_{(\hat{Y}.2)} = 40.523 \pm 0.5506\times 2.306 = 40.523 \pm 1.2697$$

$$L_i = 39.2533$$

$$L^s = 41.7927$$

Se pueden realizar, también, pruebas de hipótesis para contrastar la hipótesis nula de que $\hat{Y}$ es una estima de $\mu_{Y.X_o}$ por medio de la siguiente prueba t.

Cola derecha	Cola izquierda	Dos colas
$H_o: \hat{Y} \le \mu_{Y.X}$	$H_o: \hat{Y} \ge \mu_{Y.X}$	$H_o: \hat{Y} = \mu_{Y.X}$
$H_1: \hat{Y} > \mu_{Y.X}$	$H_1: \hat{Y} < \mu_{Y.X}$	$H_1: \hat{Y} \ne \mu_{Y.X}$

$$t_o = \frac{\hat{Y}-\mu_{Y.X}}{S_{\hat{Y}.X_o}}$$

Esta t_o se distribuye como la t de *Student* por lo que se puede contrastar con la $t_{(n-2;\,\alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-2;\,\alpha)}$ para las hipótesis de una cola.

Ejemplo.-

El ejemplo anterior se podría haber planteado como la prueba de Y=40 para X_o=2.

$$H_o: \hat{Y} = 40$$

$$H_1: \hat{Y} \ne 40$$

$$t_o = \frac{40.523-40}{0.5506} = 0.9499 ns$$

$$t_{(8;\,0.05/2)} = 3.306$$

Se acepta la hipótesis nula, por lo que se puede concluir que la media de la población para X=2 es igual a 40.

En el ejemplo *SAS* del epígrafe *Pruebas de hipótesis e intervalos de confianza de la predicción, por extrapolación, de un valor Y* se vera la resolución, por el *SAS*, de este ejemplo.

Pruebas de hipótesis e intervalos de confianza para la media de la variable dependiente.-

Un caso especial de valores estimados es el de la estima de $\overline{Y}$ en el punto $\overline{X}$. El *error típico de la media muestral observada* $\overline{Y}$, se calculó en el capítulo 3, la expresión que lo mide es

$$S_{(\overline{Y})} = \sqrt{\frac{S^2_{(Y)}}{n}}$$
$$gl = n-1$$

Este error típico se utilizó en el capítulo 4 para estimar los intervalos de confianza de la media.

Pero ahora se puede explicar parte de la variación de Y en función de la variación de X, quedando como varianza no explicada $S^2_{Y.X}$. Se puede, por tanto, utilizar el error típico

$$S_{\hat{Y}.X_o} = \sqrt{S^2_{Y.X}\left(\frac{1}{n}+\frac{(X_o-\overline{X})^2}{SC_{(X)}}\right)} =$$

$$S_{(\overline{Y})} = \sqrt{\frac{S^2_{(Y.X)}}{n}}$$
$$gl = n-2$$

y utilizar este para el calculo del intervalo de confianza de $\overline{Y}$, de la siguiente manera

$$LC_{(\overline{Y})} = \overline{Y} \pm S_{(\overline{Y})}\ t_{(n-2,\alpha/2)}$$

Ejemplo.-

Para los datos de los pollos, los límites de confianza al 95% para $\overline{Y}$, son

$$S_{(\overline{Y})} = \sqrt{\frac{1.0555}{8}} = 0.3632$$

$$LC_{(\overline{Y})} = 42.434 \pm 0.3632 \times 2.306 = 42.434 \pm 0.8375$$

$$L_i = 41.5965$$

$$L^s = 43.2715$$

Este intervalo es notablemente mas estrecho que el intervalo de confianza para la misma media de la variable *Y*, sin considerar la influencia de la variable *X*, a pesar de que aquí se tiene un grado de libertad menos. Recuérdese que este intervalo hubiera sido

$$S_{(\overline{Y})} = \sqrt{\frac{27.9464}{10}} = 1.6717$$

$$LC_{(\overline{Y})} = 42.434 \pm 1.6717 \times 2.2622 = 42.434 \pm 3.7818$$

$$L_i = 38.6522$$

$$L^s = 46.2158$$

Se pueden realizar también pruebas de hipótesis para contrastar la hipótesis nula de que $\overline{Y}$ es una estima de la media poblacional μ_o, por medio de la siguiente prueba *t*.

Cola derecha	Cola izquierda	Dos colas
$H_o : \mu_Y \leq \mu_o$	$H_o : \mu_Y \geq \mu_o$	$H_o : \mu_Y = \mu_o$
$H_1 : \mu_Y > \mu_o$	$H_1 : \mu_Y < \mu_o$	$H_1 : \mu_Y \neq \mu_o$

$$t_o = \frac{\overline{Y} - \mu_o}{S_{(\overline{Y})}}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n-2;\, \alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-2;\, \alpha)}$ para las hipótesis de una cola.

Ejemplo.-

El problema anterior se podría haber planteado como la prueba de $\overline{Y} = 42$

$$H_o: \overline{Y} = 42$$
$$H_1: \overline{Y} \neq 42$$
$$t_o = \frac{42.434 - 42}{0.3632} = 1.1949 ns$$
$$t_{(8;\,0.05/2)} = 2.3060$$

Se acepta la hipótesis nula, luego se puede concluir que la media poblacional de Y no es diferente de 42.

Pruebas de hipótesis e intervalos de confianza de la ordenada en el origen.-

Otro caso especial es el *error típico y límites de confianza de* α, el parámetro de la ordenada en el origen. Este se estima igual que para cualquier $\hat{Y}$, teniendo en cuenta que ahora el valor de X_o=0. El error típico de *a* será

$$S_{(\hat{Y}.X=0)} = \sqrt{S^2_{Y.X}\left(\frac{1}{n} + \frac{0 - \overline{X}^2}{SC_{(X)}}\right)} =$$
$$S_a = \sqrt{\frac{S^2_{Y.X} \Sigma_i X_i}{n\, SC_{(X)}}}$$

por lo tanto, el intervalo de confianza para *a*, sería

$$LC_{(\alpha)} = a \pm t_{(n-2,\,\alpha/2)}\, S_{(a)}$$

Ejemplo.-

Para los datos de los pollos, el intervalo de confianza al 95% de la ordenada en el origen es

$$a = 25.115$$
$$S_{(\hat{Y}.0)} = \sqrt{1.0555\left[\frac{1}{10} + \frac{(0 - 2.248)^2}{0.3286}\right]} = 4.0420$$
$$LC_{(\alpha)} = 25.115 \pm 2.306 \times 4.042 = 25.115 \pm 9.3209$$
$$L_i = 15.7941$$
$$L^s = 34.4359$$

Se pueden realizar también pruebas de hipótesis para contrastar la hipótesis nula de que α es una estima de la media poblacional, μ_Y, cuando no esta presente la influencia de la variable *X*. Otras veces es muy útil probar la hipótesis nula de que la recta pasa por el origen de coordenadas. Estos contrastes se hacen por medio de la

siguiente prueba *t*.

Cola derecha	Cola izquierda	Dos colas
$H_o: \alpha \leq \mu_o$	$H_o: \alpha \geq \mu_o$	$H_o: \alpha = \mu_o$
$H_1: \alpha > \mu_o$	$H_1: \alpha < \mu_o$	$H_1: \alpha \neq \mu_o$

$$t_o = \frac{a - \mu_o}{S_a}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n-2;\, \alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-2;\, \alpha)}$ para las hipótesis de una cola.

Ejemplo.-

Si los pollos consumen más pienso cuanto más pesan es razonable pensar que esta recta de regresión tendría que pasar por el origen de coordenadas, es decir, que gallinas de peso cero consuman cero pienso. Como esta hipótesis no puede probarse experimentalmente, según el ajuste a la recta en este rango de las *X*, es de esperar que sea cierta esta hipótesis.

$$H_o: \alpha = 0$$
$$H_1: \alpha \neq 0$$
$$t_o = \frac{25.115}{4.042} = 6.213^{***}$$
$$t_{(8;\, 0.001/2)} = 5.0410$$

Se puede concluir que esta recta no pasa por el centro de coordenadas, o lo que es lo mismo, que sin la influencia de *X* el valor medio de *Y* es diferente de cero.

Análisis de los residuo ($e_{Y.X}$).-

Como se ha visto anteriormente, los valores determinados por la ecuación de regresión, son estimaciones de parámetros poblacionales, es decir, de $\mu_{Y.X} = a + bX_i = \hat{Y}$. Las diferencias entre éstos valores estimados y los valores observados son estimaciones de la variación de *Y* no explicada por la variación de *X*, esto es lo que se ha denominado *residuo*.

Los residuos ($e_{Y.X}$) pueden ser particularmente útiles cuando se representan respecto de *X*. Si tienden a ser del mismo signo en ambos extremos de la gráfica y de signos opuestos en el medio, entonces queda comprobado que la respuesta no es lineal. Si sus magnitudes cambian de manera regular, por ejemplo, aumentando con *X*, entonces hay evidencia de heterogeneidad de la varianza. Los valores extremos o alejados pueden detectarse de la manera que se describe más adelante.

Una dificultad con los residuos es que no todos se estiman con la misma precisión. Sin embargo, se puede estimar errores típicos de los residuos y dividiendo cada residuo por su error típico se obtienen residuos típicos o residuos *studentizados*. Estos residuos studentizados se ajustan a la distribución *t* con *n*-2 grados de libertad, por lo que podría utilizarse como prueba de hipótesis. Sin embargo, esto no es posible porque la elección del residuo no es aleatorio, como es el supuesto de la prueba *t*, pero si se puede utilizar para detectar residuos excesivamente grandes (desviaciones sospechosamente grandes), pues residuos studentizados mayores de 2.5 son relativamente raros y habría que investigar esta desviación inusualmente grande.

El error típico del residuo es,

$$S_{eY.X} = \sqrt{S^2_{Y.X}\left[1-\left(\frac{1}{n}+\frac{(X_o-\overline{X})^2}{SC_{(X)}}\right)\right]}$$

Ejemplo.-

Calcúlese los valores estimados de *Y*, los residuos, el error típico de ambos y el valor studentizado de los residuos, del mismo ejemplo de las gallinas que se viene desarrollando en epígrafes anteriores

X	Y	$\hat{Y}$	$S_{\hat{Y}.X_o}$	$e_{Y.X}$	$e^2_{Y.X}$	$S_{eY.X}$	$e_{Y.X} / S_{eY.X}$
2.11	40.73	41.37	0.41	-0.64	0.41	0.94	-0.68
1.99	41.45	40.45	0.57	1.00	1.01	0.86	1.17
2.31	42.23	42.91	0.34	-0.68	0.46	0.97	-0.70
2.67	45.13	45.68	0.82	-0.55	0.31	0.61	-0.90
2.13	41.77	41.52	0.39	0.25	0.06	0.95	0.26
2.31	42.82	42.91	0.34	-0.09	0.01	0.97	-0.09
2.08	39.50	41.14	0.44	-1.64	2.69	0.93	-1.77
2.31	43.31	42.91	0.34	0.40	0.16	0.97	0.41
2.35	45.04	43.22	0.37	1.82	3.31	0.96	1.90
2.22	42.36	42.22	0.33	0.14	0.02	0.97	0.15
22.48	424.34	424.34		0	8.44		

Tal como era de esperar por el criterio minimocuadrático seguido para la estima de la recta, las sumas de los residuos vale cero, y la suma de los cuadrados de los residuos (que es la más pequeña posible) es igual a la de la $SC_{(error)}$ de la prueba de bondad de ajuste, con una ligera variación debido a errores de redondeo.

Como se observa en la penúltima columna, los errores típicos de los residuos oscilan alrededor de los mismos valores, no hay ningún error típico excesivamente grande. Y en la ultima columna se comprueba que ningún valor studentizado del error sobrepasa en valor absoluto el 2.5, por lo que se puede concluir que no hay ningún dato excesivamente alejado de su valor esperado.

Análisis de las *influencias*.-

Los valores extremos pueden, a veces, pasar desapercibidos al análisis de los residuos y sus valores studentizados del epígrafe anterior, porque las estimaciones minimocuadráticas de la recta de regresión tienden a situarse en valores intermedios de las observaciones extremas. Por tanto, las estimas de los residuos de dichas observaciones pueden no ser especialmente grandes, interfiriendo en la búsqueda de valores extremos. Esta dificultad puede ser soslayada si se calculan las estimas y estadísticos de la observación cuestionada por medio de una recta de regresión estimada con todos los pares de valores menos con el punto cuestionado. Esto no requiere la repetición de todos los cálculos, pues basta con restarle a los sumatorios básicos (ΣX, ΣX^2, ΣY, ΣY^2 y ΣXY) el valor correspondiente de la observación eliminada. En todo caso, los paquetes estadísticos proveen de estos estadísticos.

Dichos estadísticos se anotan con el subíndice *-i* refiriéndose este subíndice al subíndice de la observación omitida. Por ejemplo, $S^2_{b\text{-}2}$, es la varianza de la regresión estimada sin el segundo par de valores, de las parejas de valores que se tienen. Estos estadísticos nos indican la potencial *influencia* de una observación concreta.

Residuos *studentizados*.-

El primer estadístico es una versión de los *residuos studentizados*, en el que los residuos se dividen, como en el caso anterior, por los errores típicos de los residuos, pero utilizando para el cálculo de dichos errores típicos, la varianza del error calculada con todos los pares de valores menos con el par en cuestión, es decir, $S^2_{Y.X\text{-}i}$.

Este residuo studentizado se calcularía

$$\frac{Y_i - \hat{Y}_i}{S_{eY.X_{i,-i}}}$$

siendo

$$S_{eY.X_{i,-i}} = \sqrt{S^2_{Y.X\text{-}i}\left[1\left(\frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}}\right)\right]}$$

Este residuo studentizado se distribuye con la misma *t* y con los mismos *gl* que el del epígrafe anterior, pero es más sensible que aquel, siendo el criterio el mismo, es decir, rechazar por extremos las observaciones que de un residuo studentizado mayor de 2.5.

Diferencias de ajustes.-

Un segundo estadístico es el que se puede denominar *diferencias de ajustes*, consiste en la diferencia entre dos valores estimados de la misma *Y*, la primera estimación es la realizada con la ecuación de la recta calculada con todos los pares de valores, es decir, $\hat{Y}_i$, y la segunda es la estimación realizada con la ecuación de la recta calculada con todos los pares de valores menos con el par en cuestión, es decir, $\hat{Y}_{i,-i}$. Esta diferencia es dividida por el error típico de un valor estimado, pero utilizando la varianza del error calculada con todos los pares de valores menos el par en cuestión, esto es, con la simbolizada como $S^2_{Y.X-i}$.

El cálculo de estas diferencias de ajuste sería

$$DA = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{S_{\hat{Y}.X_{i,-i}}}$$

siendo

$$S_{\hat{Y}.X_{i,-i}} = \sqrt{S^2_{Y.X-i}\left(\frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}}\right)}$$

Este estadístico es un buen indicador de la *influencia*. Se sugiere, que un buen criterio para detectar observaciones influyentes, es el de las observaciones que superen, en valor absoluto, el valor

$$2\sqrt{\frac{m+1}{n}}$$

siendo *m* el número de variables independientes y *n* el número de pares de valores.

***D* de *Cook*.-**

Un tercer estadístico es la *D* de *Cook*, que es semejante al anterior, pero en este, a diferencia del anterior, se utiliza la varianza del error de todas las observaciones, es decir, el error típico de un valor estimado; y se eleva al cuadrado y divide por dos, para resaltar más los valores extremos.

El cálculo de esta *D* de *Cook* sería

$$D = \frac{\left(\frac{\hat{Y}_i - \hat{Y}_{i,-i}}{S_{\hat{Y}.X_i}}\right)^2}{2}$$

Medida tipificada de lo extrema que es una observación.-

Otro estadístico sería h_i,

$$h_i = \frac{1}{n} + \frac{(X_o - \overline{X})^2}{SC_{(X)}}$$

que indica la influencia de cada observación. Este es una medida tipificada de lo extrema que es una observación, en el espacio de las X, con respecto al centro.

La suma de todos los h_i vale m+1, siendo m el número de variables independientes, por tanto, el valor esperado de h_i es $(m+1)/n$; si una observación supera dos veces este valor, se puede considerar que tiene una gran influencia.

Proporción de las varianzas generalizadas.-

Otro estadístico es la *proporción de las varianzas generalizadas*. La varianza generalizada de una muestra de pares de valores, es la razón entre la varianza del error y la suma de cuadrados de la variable independiente por el número de observaciones, esto es

$$PVG = \frac{S^2_{Y.X}}{n SC_{(X)}}$$

Este estadístico es el resultado de dividir la varianza generalizada sin la *i-ésima* observación y la varianza generalizada con todas las observaciones, es decir

$$PVG = \frac{\frac{S^2_{Y.X-I}}{n SC_{(X-I)}}}{\frac{S^2_{Y.X}}{n SC_{(X)}}}$$

Si el valor de este estadístico es superior a uno indica que la inclusión de dicha observación tiene como resultado incrementar la precisión, mientras que un valor inferior a uno tiene como resultado una disminución de la precisión. Se sugiere que valores que excedan a la unidad en

$$\frac{3(m+1)}{n}$$

pueden ser considerados como excesivos.

Diferencias tipificadas de las *b* y de las *a*.-

Otro estadístico es la diferencia tipificada de la regresión calculada con todos y

sin el *i-ésimo* valor. Y lo mismo para la ordenada en el origen.

También puede ser de utilidad comparar la suma de cuadrados residual (suma de los cuadrados de todos los residuos) con la suma de cuadrados residual estimada, esto es, la suma de los cuadrados de los residuos obtenidos restándole al valor observado el valor estimando con arreglo a la ecuación calculada con todos los demás valores. La suma de cuadrados residual estimada se espera sea mayor que la original, y será mucho mayor cuanto más valores extremos haya.

Ejemplo.-

Calcúlese todos estos estadísticos de *influencia* en el mismo ejemplo de los pollos.

b_{-i}	a_{-i}	$\hat{Y}_{i,-i}$	$e^2_{Y,X_{i,-i}}$	$S^2_{Y,X-i}$	$e_{Y,X} / S_{e_{Y,X-i}}$	DA
7.385	25.909	41.491	0.579	1.139	-0.778	-0.283
8.835	22.429	40.011	2.072	0.999	1.724	0.792
7.850	24.865	42.997	0.589	1.131	-0.766	-0.242
9.697	20.790	46.681	2.406	1.083	-2.492	-1.195
7.807	24.854	41.484	0.082	1.195	0.282	0.098
7.724	25.081	42.923	0.011	1.204	-0.100	-0.033
6.675	27.631	41.514	4.056	0.734	-2.606	-1.015
7.620	25.259	42.862	0.201	1.180	0.438	0.137
7.054	26.367	42.944	4.394	0.661	2.767	0.934
7.718	25.068	42.202	0.025	1.202	0.152	0.045
			14.415			

D	h_i	PVG	$(b-b_{-i})/S_b$	$(a-a_{-i})/S_a$
0.043	0.158	1.278	0.178	-0.196
0.297	0.303	1.358	-0.631	0.664
0.031	0.112	1.207	-0.081	0.062
0.733	0.642	2.865	-1.112	1.070
0.005	0.142	1.321	-0.058	0.064
0.001	0.112	1.284	-0.011	0.009
0.358	0.186	0.854	0.574	-0.622
0.010	0.112	1.259	0.047	-0.036
0.273	0.132	0.721	0.363	-0.310
0.001	0.102	1.269	-0.008	0.012
	2.001			

Indiquemos y comentemos los cálculos para la primera observación. Si de los diez pares de valores de este ejemplo, eliminamos el primero, se tendría

$$SP_{(XY)} = 870.5074 - \frac{20.37 \times 383.61}{9} = 2.2701$$

$$SC_{(X)} = 46.4115 - \frac{20.37^2}{9} = 0.3074$$

$$SC_{(Y)} = 16375.46 - \frac{383.61^2}{9} = 24.7231$$

$$b = \frac{2.2701}{0.3074} = 7.3848$$

$$a = 42.6233 - 2.2633 \times 7.3848 = 25.909$$

$$b\,SP_{(xy)} = 7.3848 \times 2.2701 = 16.7642$$

$$SC_{(Y)} - b\,SP_{(XY)} = 24.7431 - 16.7642 = 7.9589$$

La estima de la primera observación con la ecuación de regresión calculada con todas las demás observaciones, es

$$\hat{Y} = 25.909 + 7.3848 \times 2.11 = 41.4909$$

El error de esta estima es

$$e_{Y.X_{i-i}} = 40.73 - 41.4909 = -0.7609$$

La varianza del error es

$$S^2_{Y.X-i} = \frac{SC_{(Y)} - b\,SP_{(XY)}}{n-2} = \frac{24.7231 - 16.7642}{7-2} = 1.5918$$

El error típico del residuos es

$$S_{eY.X-i} = \sqrt{S^2_{Y.X-i}\left[1 - \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SC_{(X)}}\right)\right]} =$$

$$= \sqrt{1.5918\left[1 - \left(\frac{1}{10} + \frac{(2.11 - 2.248)^2}{0.3286}\right)\right]} =$$

$$= 1.1577$$

Por tanto el nuevo residuo studentizado es

$$\frac{e_{Y.X}}{S_{eY.X-i}} = \frac{-0.6407}{1.1577} = -0.553$$

Como se observa no supera el 2.5, por lo que se puede afirmar que el valor de esta

observación no es estadísticamente excesivo.

Para *la diferencia de ajuste* de esta observación se necesita el error típico de un valor estimado, utilizando como varianza del error la calculada con todos los demás pares de valores, este error típico es

$$S_{\hat{Y}.X_{i,-i}} = \sqrt{S^2_{Y.X-i}\left(\frac{1}{n}+\frac{(X_i-\overline{X})^2}{SC_{(X)}}\right)} =$$

$$= \sqrt{11398\left(\frac{1}{10}+\frac{(2.11-2.248)^2}{0.3286}\right)} =$$

$$= 0.4243$$

Por lo tanto, la diferencia de ajuste es

$$DA = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{S_{\hat{Y}.X_{i,-i}}} = \frac{41.3707 - 41.4909}{0.4243} = -0.2833$$

Como se observa, no supera en valor absoluto la cifra:

$$2\sqrt{\frac{m+1}{n}} = 2\sqrt{\frac{1+1}{10}} = 0.894$$

por lo que es una observación no *influyente*.

La *D* de *Cook* es

$$D = \frac{\left(\frac{\hat{Y}_i-\hat{Y}_{i,-i}}{S_{\hat{Y}.X_i}}\right)^2}{2} = \frac{\left(\frac{-0.1202}{0.408}\right)^2}{2} = 0.0434$$

La h_i es

$$h_i = \frac{1}{10}+\frac{(2.11-2.248)^2}{0.3286} = 0.1579$$

como se ve no supera el valor:

$$2\frac{m+1}{n} = 2\frac{2}{10} = 0.4$$

por lo que esta observación está poco alejada del centro del espacio de las *X*.

La proporción de varianzas generalizadas es

$$PVG = \frac{\frac{1.1398}{9\times0.3074}}{\frac{1.0555}{10\times0.3286}} = 1.2827$$

Como este valor no es inferior a la unidad, la inclusión de esta pareja de valores no disminuye la precisión.

Para la diferencia tipificada entre el coeficiente de regresión obtenido con todos los pares de valores y el obtenido con todos menos con el *i-ésimo*, al igual que con la ordenada en el origen, se realiza exactamente igual que las pruebas de hipótesis que vimos anteriormente para ambos parámetros.

Como se ve, la suma de los residuos estimados es algo mayor que la suma de cuadrados del error, pero no varias unidades mayor.

Antes de seguir adelante veamos la resolución por el *SAS* del ejemplo que hemos estado realizando desde el comienzo de este Capítulo.

Archivo del programa SAS (C11-1.SAS).-

```
title 'Regresión, residuos e influencias';
options ls=75 ps=60;
data ejereg;
infile 'c11-1.dat';
input peso alimento;
title 'Valores estimados, Residuos e Influencia de los datos';
proc reg;
  model alimento = peso / R  influence;
run;
title 'prueba para b=0 (es la salida anterior por defecto)';
peso0:test peso = 0;
run;
title 'prueba para b=7 (se puede probar cualquier valor)';
   peso7:test peso = 7;
run;
```

Archivo de datos (C11-1.DAT) .-

```
2.11    40.73
1.99    41.45
2.31    42.23
2.67    45.13
2.13    41.77
2.31    42.82
2.08    39.50
2.31    43.31
2.35    45.04
2.22    42.36
```

Archivo de resultados (C11-1.LST) .-

Valores estimados, Residuos e Influencia de los datos

Model: MODEL1
Dependent Variable: ALIMENTO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	19.50448	19.50448	18.488	0.0026
Error	8	8.43976	1.05497		
C Total	9	27.94424			

Root MSE	1.02712	R-square	0.6980
Dep Mean	42.43400	Adj R-sq	0.6602
C.V.	2.42051		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	25.113672	4.04125572	6.214	0.0003
PESO	1	7.704772	1.79189594	4.300	0.0026

Obs	Dep Var ALIMENTO	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
1	40.7300	41.3707	0.408	-0.6407	0.943	-0.680
2	41.4500	40.4462	0.565	1.0038	0.858	1.170
3	42.2300	42.9117	0.343	-0.6817	0.968	-0.704
4	45.1300	45.6854	0.823	-0.5554	0.615	-0.904
5	41.7700	41.5248	0.388	0.2452	0.951	0.258
6	42.8200	42.9117	0.343	-0.0917	0.968	-0.095
7	39.5000	41.1396	0.443	-1.6396	0.927	-1.769
8	43.3100	42.9117	0.343	0.3983	0.968	0.411
9	45.0400	43.2199	0.373	1.8201	0.957	1.902
10	42.3600	42.2183	0.329	0.1417	0.973	0.146

Obs	-2-1-0 1 2	Cook's D	Rstudent	Hat Diag H	Cov Ratio	Dffits
1	* |	0.043	-0.6551	0.1580	1.3771	-0.2837
2	| **	0.297	1.2025	0.3026	1.2865	0.7921
3	* |	0.031	-0.6801	0.1117	1.2937	-0.2412
4	* |	0.732	-0.8922	0.6420	2.9415	-1.1948
5	|	0.006	0.2421	0.1424	1.4978	0.0986
6	|	0.001	-0.0887	0.1117	1.4671	-0.0314
7	*** |	0.357	-2.1211	0.1859	0.5945	-1.0136
8	|	0.011	0.3890	0.1117	1.4088	0.1379
9	| ***	0.274	2.4031	0.1317	0.4516	0.9357
10	| |	0.001	0.1364	0.1024	1.4474	0.0461

Obs	INTERCEP Dfbetas	PESO Dfbetas
1	-0.1895	0.1719
2	0.6826	-0.6481
3	0.0595	-0.0781
4	1.0564	-1.0978
5	0.0603	-0.0538
6	0.0078	-0.0102
7	-0.7465	0.6890
8	-0.0340	0.0446
9	-0.3919	0.4589
10	0.0107	-0.0070

Sum of Residuals	0
Sum of Squared Residuals	8.4398
Predicted Resid SS (Press)	14.4151

```
          prueba para b=0 (es la salida anterior por defecto)

Dependent Variable: ALIMENTO
Test: PESO0     Numerator:      19.5045  DF:     1   F value:  18.4882
                Denominator:    1.05497  DF:     8   Prob>F:    0.0026

Dependent Variable: ALIMENTO
Test: PESO7     Numerator:       0.1632  DF:     1   F value:   0.1547
                Denominator:    1.05497  DF:     8   Prob>F:    0.7044
```

Con el error típico de las predicciones (*Std Err Predict*) se puede estimar los intervalos de confianza de las estimaciones y de las predicciones, pero el *SAS* también provee de estos intervalos, lo cual se vera en el ejemplo *SAS* del siguiente epígrafe.

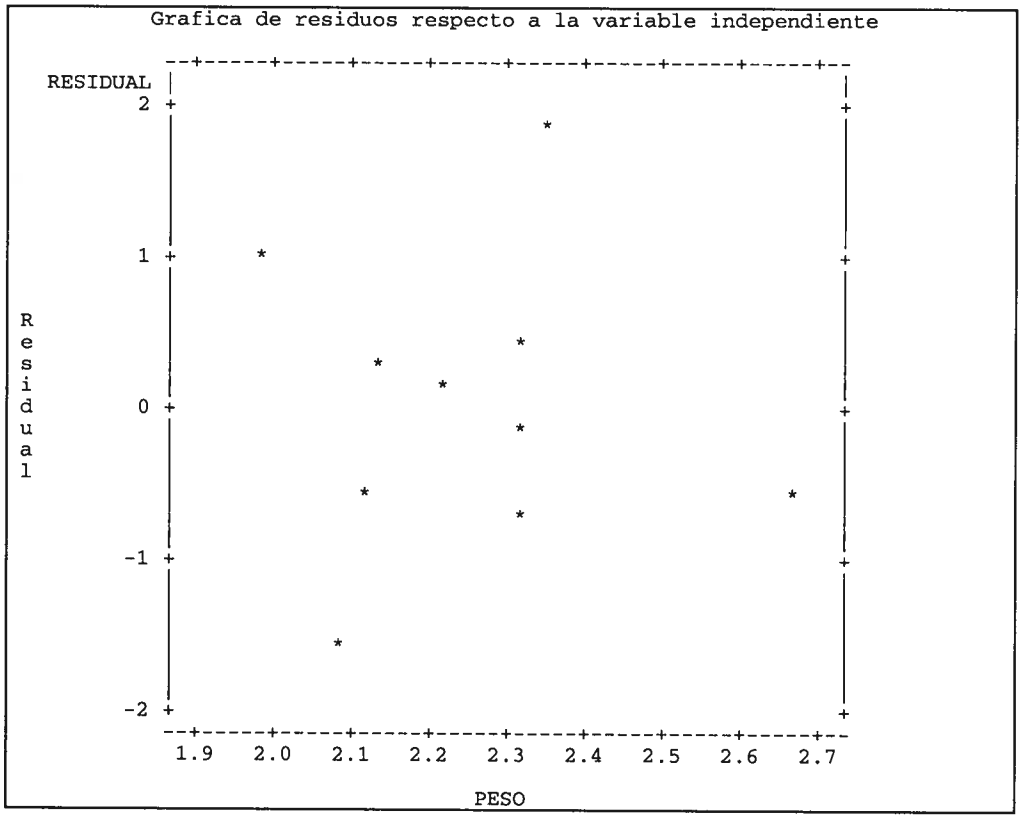
Como se ha dicho en el epígrafe *Análisis de los residuos*, los residuos ($e_{Y.X}$) pueden ser particularmente útiles cuando se representan respecto de X. Si tienden a ser del mismo signo en ambos extremos de la gráfica y de signos opuestos en el medio, entonces queda comprobado que la respuesta no es lineal. Si sus magnitudes cambian de manera regular, por ejemplo, aumentando con X, entonces hay evidencia de heterogeneidad de la varianza.

Si se quiere hacer dicha representación se haría con el siguiente programa *SAS*.

Archivo del programa *SAS* (C11-2.SAS).-

```
title 'Grafica de residuos respecto a la variable independiente';
option ls=64 ps=40;
data ejereg;
infile 'c11-1.dat';
input peso alimento;
proc reg;
  model alimento=peso/p noprint;
  plot residual.*peso='*';
run;
```

Archivo de resultados (C11-2.LST).-



Como se ve, hay cinco residuos mayor de cero y cinco residuos menores de cero. Eso ya se sabía viendo la columna encabezada por $e_{Y.X}$ en la tabla del *Análisis de los residuos*. La ventaja de la gráfica es que si hay una desviación significativa se puede visualizar la tendencia de esta desviación, como se vera en el ejemplos del epígrafe *Curvas polinómicas*.

Varios valores de *Y* por cada valor de *X*.-

Se puede dar el caso en el que se realicen varias medidas de la variable dependiente para cada valor de la variables independiente, por lo que se obtiene una distribución de los valores de *Y* para cada valor de *X*.

En el ejemplo de las gallinas, que se ha visto anteriormente, había tres valores de pienso consumido en 350 días para un peso promedio de 2.31. Estas o más repeticiones se pueden producir para los otros pesos, existiendo, por tanto, ocho pesos diferentes y varios valores repetidos de cantidad de pienso consumido para cada peso. Este diseño es igual al que vimos en el capítulo 5 y que se analizó por medio de un análisis de varianza de una vía. Es decir, los diferentes pesos pueden ser considerados

como los diferentes niveles del factor peso o como los diferentes tratamientos y se puede estudiar el efecto que el peso promedio del individuo tiene sobre el consumo de pienso para establecer si existen diferencias en la cantidad de pienso consumido por los diez grupos. Y como este factor es cuantitativo se puede, también, establecer una línea de regresión de cantidad de pienso sobre el peso, tal como se ha hecho en este capítulo.

Recuérdese que la desviación de una observación cualquiera respecto de la media general, en el modelo de una vía es

$$Y_{ij} - \overline{Y}_{..} = (\overline{Y}_{i.} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{i.})$$

y que la desviación de una observación respecto a la media general en el modelo de regresión lineal es

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

Por lo tanto se puede hacer ambos análisis. En una primera aproximación se puede realizar un *ANOVA* de una vía. Si este análisis da claramente no significativo, indicaría que las medias de la variable dependiente no son significativamente diferentes entre ellas y, por tanto, la recta que ajustemos tendrá una pendiente, estadísticamente, no diferente de cero. Si el análisis de la varianza da significativo, puede ser porque las medias de los grupos (la variable dependiente) aumentan o disminuyen conforme aumenta o disminuye el valor de la variable independiente, por lo que habrá una regresión significativa de la *Y* en la *X*. Esta regresión puede ser significativa aunque el análisis de la varianza haya dado no significativo pero próximo a los niveles críticos. Para nuestro ejemplo, esta significación en el *ANOVA* puede ser porque aumenta o disminuye el consumo de pienso cuando aumenta o disminuye el peso promedio. Si el *ANOVA* ha dado significativo y la regresión no es significativamente diferente de cero, es porque el cambio de las medias de la variable *Y* para los diferentes grupos de la variable *X* no es recta, sino que es curvilínea y habría que hacer el ajuste a un polinomio (ver más adelante).

El calculo del coeficiente de regresión y la prueba de ajuste se hace de la misma manera que se ha visto a lo largo de este capítulo, la única precaución es la de repetir el valor de *X* para los diferentes valores de *Y* de cada grupo.

Pero también se puede realizar ambos análisis conjuntamente, si se observa que el miembro de la izquierda de la desviación de una observación con respecto a la media general en el modelo de regresión, que se ha expuesto más arriba, es así cuando solo hay una medida de *Y* para cada valor de *X*, pero si se tiene varias medidas de *Y* para cada valor de *X* este miembro de la izquierda sería la desviación de la media de cada grupo con respecto a la media general, es decir,

$$Y_i - \overline{Y} = \overline{Y}_{i.} - \overline{Y}_{..}$$

y como se ve, esto es la desviación debida al efecto del *i-ésimo* nivel del factor, tal como se expresa en el primer miembro de la parte derecha de la expresión de la

desviación de una observación cualquiera respecto de su media general, en el modelo de una vía, que se ha anotado unos párrafos más arriba. Por tanto, este miembro a su vez es igual a

$$\overline{Y}_{i.} - \overline{Y}_{..} = (\hat{Y}_{i.} - \overline{Y}_{..}) + (\overline{Y}_{i.} - \hat{Y}_{i.})$$

quedando una expresión compuesta de la siguiente manera

$$Y_{ij} - \overline{Y}_{..} = (\overline{Y}_{i.} - \overline{Y}_{..})[= (\hat{Y}_{i.} - \overline{Y}_{..}) + (\overline{Y}_{i.} - \hat{Y}_{i.})] + (Y_{ij} - \overline{Y}_{i.})$$

que permitirá realizar un análisis teniendo en cuenta, a la vez, la descomposición de la suma de cuadrados a la que conduce ambas desviaciones.

Este análisis puede quedar de la siguiente manera

FV	*gl*	*SC*	*CM*	F_o
Tratamientos	*t*-1	$SC_{(T)}$	$\frac{SC_{(T)}}{t-1}$	$\frac{CM_{(T)}}{CM_{(E)}}$
Regresión	1	$b\,SP_{(XY)}$	$b\,SP_{(XY)}$	$\frac{b\,SP_{(XY)}}{S^2_{Y.X}}$
Residuo	*t*-2	$SC_{(T)}$-$bSP_{(XY)}$	$\frac{SC_{(T)} - b\ SP_{(XY)}}{t-2}$	$\frac{S^2_{Y.X}}{CM_{(E)}}$
Error	*N*-*t*	$SC_{(Y)}$-$SC_{(T)}$	$\frac{SC_{(E)}}{N-t}$	
Total	*N*-1	$SC_{(Y)}$		

siendo

$$SC_{(T)} = \Sigma_i \frac{Y^2_{i.}}{n_i} - \frac{Y^2_{..}}{N}$$

$$S^2_{Y.X} = \frac{SC_{(T)} - b\,SP_{(XY)}}{t-2}$$

Las pruebas de hipótesis para *Tratamientos* y para *Regresión* tienen el mismo sentido que el estudiado anteriormente. Pero la prueba de hipótesis del *Residuo* o de la *desviación de la regresión* es nueva, si esta prueba es no significativa indica que la variabilidad de *Y* es explicada suficientemente como una función lineal de *X*, en el caso de que la prueba fuera significativa, indicaría, si la prueba de la regresión es significativa, que existe una gran heterogeneidad aleatoria alrededor de la línea de regresión, y si la prueba de la regresión es no significativa, indicaría que la relación funcional es curvilinea.

Ejemplo.-

Siguiendo con el ejemplo de los pollos visto anteriormente, ahora se tiene ocho líneas de pollos que se caracterizan, entre otras cosas, por su diferente peso medio. Se toman 50 gallinas de cada línea, de manera que el peso medio, X, de cada grupo sea el de la línea; y le medimos, además, el consumo de pienso Y en 350 días. Para aumentar la precisión se hacen cuatro repeticiones. Se quiere saber si existe diferencia en el consumo de pienso entre las líneas, y en el caso positivo, se quiere saber que parte de la variabilidad del consumo de pienso de las líneas es debida a la regresión lineal con el peso de las líneas y si esta regresión lineal es suficiente para explicar esta variabilidad.

Línea	1	2	3	4	5	6	7	8
Peso	1.99	2.09	2.13	2.18	2.22	2.31	2.36	2.68
Consumo	41.46	39.51	41.78	40.73	42.37	42.23	45.04	45.13
	40.96	39.14	41.59	40.10	41.87	43.32	44.50	44.99
	41.73	40.01	42.14	40.91	42.68	42.82	42.82	45.27
	41.28	39.60	41.82	40.78	42.27	42.37	44.99	45.18
ΣY	165.43	158.26	167.33	162.52	169.19	170.74	177.35	180.57
ΣX	7.96	8.36	8.52	8.72	8.88	9.24	9.44	10.72

$$\overline{Y} = 42.23$$

$$\overline{X} = 2.24$$

$$SC_{(T)} = 57165.336 - \frac{1351.39^2}{32} = 94.8693$$

$$SC_{(X)} = 162.536 - \frac{71.84^2}{32} = 1.2552$$

$$SC_{(Y)} = 57170.932 - \frac{1351.39^2}{32} = 100.4657$$

$$SP_{(XY)} = 3043.1604 - \frac{71.84 \times 1351.39}{32} = 9.2898$$

$$b = \frac{9.2898}{1.2552} = 7.401$$

$$a = 42.23 - 7.401 \times 2.24 = 25.6518$$

FV	*gl*	*SC*	*CM*	F_o
Líneas	7	94.869	13.553	58.17***
Regresión	1	68.755	68.755	15.79**
Residuo	6	26.114	4.352	18.68***
Error	24	5.596	0.233	
Total	31	100.465		

Como se observa, existe diferencias significativas entre líneas para el consumo de pienso. Parte de estas diferencias es achacable a que este consumo es función lineal del peso de los individuos, tal como muestra la significación de la prueba de la *Regresión*, aunque queda otra variabilidad no explicada por la regresión tal como muestra la significación de la componente *Residuo*, que se estudiará en el siguiente epígrafe.

Dado que existe una regresión significativa del consumo en el peso, podemos calcular la función lineal que relaciona ambas variables, siendo esta

$$Y = 25.6518 + 7.401\,X$$

Archivo del programa *SAS* (C11.3.SAS).-

```
Title 'Anova y Regresión';
options ls=75 ps=60;
data regres;
infile 'c11-3.dat';
input linea $ peso alimento @@;
title 'MODEL=líneas LINEA=Residuo';
proc glm;
class linea;
  model alimento = peso linea / ss1;
run;
title 'PESO= Regresión';
   test h=peso e=linea;
run;
title 'Análisis de regresión sin tener en cuenta las líneas';
proc reg;
  model alimento = peso;
run;
title 'Análisis de varianza sin tener en cuenta la regresión';
proc anova;
class linea;
  model alimento = linea;
run;
```

Archivo de datos (C11.3.DAT).-

```
L1 1.99 41.46  L1 1.99 40.96  L1 1.99 41.73  L1 1.99 41.28
L2 2.09 39.51  L2 2.09 39.14  L2 2.09 40.01  L2 2.09 39.60
L3 2.13 41.78  L3 2.13 41.59  L3 2.13 42.14  L3 2.13 41.82
L4 2.18 40.73  L4 2.18 40.10  L4 2.18 40.91  L4 2.18 40.78
L5 2.22 42.37  L5 2.22 41.87  L5 2.22 42.68  L5 2.22 42.27
L6 2.31 42.23  L6 2.31 43.32  L6 2.31 42.82  L6 2.31 42.37
L7 2.36 45.04  L7 2.36 44.50  L7 2.36 42.82  L7 2.36 44.99
L8 2.68 45.13  L8 2.68 44.99  L8 2.68 45.27  L8 2.68 45.18
```

Archivo de resultados (C11-3.LST).-

MODEL=líneas LINEA=Residuo

General Linear Models Procedure

Dependent Variable: ALIMENTO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	94.8690969	13.5527281	58.12	0.0001
Error	24	5.5965750	0.2331906		
Corrected Total	31	100.4656719			

R-Square	C.V.	Root MSE	ALIMENTO Mean
0.944294	1.143470	0.48290	42.2309

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PESO	1	68.7550295	68.7550295	294.84	0.0001
LINEA	6	26.1140674	4.3523446	18.66	0.0001

PESO= Regresión

Tests of Hypotheses using the Type I MS for LINEA as an error term

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PESO	1	68.7550295	68.7550295	15.80	0.0073

Análisis de regresión sin tener en cuenta las líneas

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	68.75503	68.75503	65.046	0.0001
Error	30	31.71064	1.05702		
C Total	31	100.46567			

Root MSE	1.02812	R-square	0.6844
Dep Mean	42.23094	Adj R-sq	0.6738
C.V.	2.43451		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	25.615487	2.06816518	12.386	0.0001
PESO	1	7.401091	0.91766766	8.065	0.0001

Análisis de varianza sin tener en cuenta la regresión

Analysis of Variance Procedure

Dependent Variable: ALIMENTO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	94.8690969	13.5527281	58.12	0.0001
Error	24	5.5965750	0.2331906		
Corrected Total	31	100.4656719			

R-Square	C.V.	Root MSE	ALIMENTO Mean
0.944294	1.143470	0.48290	42.2309

Source	DF	Anova SS	Mean Square	F Value	Pr > F
LINEA	7	94.8690969	13.5527281	58.12	0.0001

Valores de regresión y valores ajustados.-

Los valores determinados por la ecuación de regresión, *valores de regresión*, de Y son estimaciones de parámetros poblacionales, es decir, de $\mu_{Y.X} = a + bX_i$. Las diferencias entre éstos y los valores observados son estimaciones de la variación de Y no explicada por la variación de X. En la siguiente tabla presentamos los residuos observados para el mismo ejemplo (el primero) de los pollos.

X	$\overline{Y}$	$\hat{Y}$	$e_{Y.X}$	$\overline{Y} + e_{Y.X}$
1.99	41.36	40.37	-0.99	42.35
2.09	39.56	41.12	1.56	38.00
2.13	41.83	41.41	-0.42	42.25
2.18	40.63	41.78	1.15	39.48
2.22	42.30	42.08	-0.22	42.52
2.31	42.68	42.75	0.07	42.61
2.36	44.34	43.12	-1.22	45.56
2.68	45.14	45.49	0.35	44.79
17.96	337.84	338.12	0.28	

A los valores ajustados (última columna de tabla anterior), se les ha eliminado la contribución debida a la regresión. Es como si el consumo de pienso de cada línea se moviera paralelamente a la recta de regresión muestral para cada valor de X y se midiese entonces como un nuevo valor ajustado de Y, es decir, en este ejemplo son los consumos ajustados, que son los esperados si todas las gallinas de las ocho líneas tuvieran el mismo peso del cuerpo. Estos se obtienen sumándole los residuos a $\overline{Y}$=93.275

$$Y_{adj} = \overline{Y} + e_{Y.X} = -b(X - \overline{X})$$

Las comparaciones entre medias ajustadas son muy útiles. En el ejemplo anterior, pongamos por caso, puede ser interesante saber cual es el consumo de pienso debido a la *línea* quitándole la influencia del peso del cuerpo, como si las gallinas de las ocho líneas pesaran lo mismo. Ese consumo estimado o ajustado es el de la ultima columna de la tabla anterior. Ahora se puede hacer un análisis de varianza de estos valores ajustados, lo que nos dará si existe diferencia para el consumo del pienso entre las líneas como si todas pesaran lo mismo. Esta prueba se hizo en el ejemplo del epígrafe anterior cuando se probó el *Residuo*, que al dar significativo, indica que el consumo de pienso de las diferentes líneas es diferente, aunque todas las gallinas pesaran lo mismo.

Esta prueba la hace el paquete estadístico *SAS* con el procedimiento *GLM* tal como se vio en el ejemplo *SAS* del epígrafe anterior. En el capítulo 13 se hace un ejemplo en el epígrafe igual a este.

Prueba de homogeneidad de dos o más líneas de regresión.-

A menudo el experimentador obtiene dos o más líneas de regresión a partir de datos análogos y desea saber si las relaciones funcionales descritas por las ecuaciones de regresión son las mismas o diferentes. Por ejemplo, se puede haber establecido la regresión de la concentración sanguínea de colesterol sobre la edad en una muestra de individuos y se puede desear, ahora, comparar esta ecuación de regresión con la de otra u otras muestras sometidas a una dieta diferente. Por lo que lo que se desea es contrastar la *homogeneidad* de los *b*, es decir, determinar si pueden considerarse o no estimaciones de un β común. El diseño básico de tal tipo de prueba es el del análisis de varianza. Existirán *t* muestras representando los grupos de tratamiento y el control, si lo hay. Existe, sin embargo, un nuevo aspecto de bastante importancia: en los análisis de varianza realizados hasta el momento lo han sido para una variable, la *Y*. En este ejemplo, *Y* sería la concentración de colesterol. Sin embargo, ahora además, para cada lectura de *Y* tenemos una lectura de *X*, la edad del individuo. De manera que son posibles dos análisis de varianza separados, uno para cada variable y también un análisis conjunto de la *covarianza de X e Y*. Es decir, tenemos un análisis de la covarianza, que aunque se estudiará más detenidamente en un próximo capítulo, se verá su utilidad en la comparación o prueba de homogeneidad de varios coeficientes de regresión.

Para comprobar la igualdad entre *t* coeficientes de regresión se necesita la suma de cuadrados debida a los diferentes coeficientes de regresión, esta es

$$SC_{(\text{entre b})} = \Sigma_i \, SC_{(X_i)} (b_i - \bar{b})^2$$

teniendo *t*-1 grados de libertad y siendo

$$\bar{b} = \frac{\Sigma_i \, SP_{(X_i Y_i)}}{\Sigma_i \, SC_{(X_i)}}$$

Es decir, el coeficiente de regresión medio es igual a la suma de las *SP* de los *t* tratamientos o grupos dividido por la suma de las *SC* de los *t* grupos. Para calcular la suma de cuadrados debida a los diferentes coeficientes de regresión, se le resta a cada coeficiente de regresión el coeficiente medio, se eleva al cuadrado y se multiplica por la suma de cuadrados de la variable independiente del grupo, sumándose al final todas estas cifras.

La formula practica para la suma de cuadrados debida a los diferentes coeficientes de regresión es

$$SC_{(\text{entre b})} = \Sigma_i \, SC_{(\text{regresión}_i)} - \bar{b} \, \Sigma_i \, SP_{(X_i Y_i)}$$

El término de error para contrastar si el cuadrado medio debido a los diferentes coeficientes de regresión es significativo, es la suma de las sumas de cuadrados residuales de los *t* grupos, siendo sus grados de libertad la suma de los grados de libertad de cada coeficiente de regresión

Global

$$\bar{b} = \frac{2.228 + 3.087 + 2.882}{35.58 + 27.68 + 41.08} = \frac{8.197}{104.34} = 0.0786$$

$$SC_{(\text{entre b})} = 0.1395+0.3443+0.2022 - (0.0786 \times 8.197) = 0.0420$$

$$SC_{(\text{residuo})} = 0.506+0.5120+0.195 = 1.213$$

$$SC_{(Y)} = 536.59 - \frac{126.61^2}{30} = 2.2536$$

FV	*gl*	*SC*	*CM*	F_o
Entre b	2	0.0420	0.021	0.415*ns*
Error	24	1.213	0.0505	
Total	26	1.2536		

Se concluye que en las tres dietas existe la misma relación funcional entre el peso inicial y el peso final.

Archivo del programa *SAS* (C11-4.SAS).-

```
title 'Homogeneidad de coeficiente de regresión';
options ls=75 ps=60;
data homoregr;
infile 'c11-4.dat';
input dieta $ X Y @@;
title 'Coeficiente de regresión en cada dieta';
proc sort; by dieta;
proc reg;
model y=x;
by dieta;
run;
title 'Homogeneidad de coeficientes de regresión';
proc glm;
  class dieta;
  model Y = dieta X dieta*X;
run;
```

Archivo de datos (C11-4.DAT).-

```
A 14.8  4.22  B 15.8  3.78  C 19.8  4.07
A 18.9  3.84  B 19.9  4.65  C 19.9  4.02
A 19.2  4.43  B 21.2  4.04  C 19.2  4.28
A 19.7  4.46  B 15.7  3.71  C 16.7  4.02
A 21.4  4.74  B 19.4  4.20  C 19.4  4.02
A 19.8  4.41  B 19.8  4.46  C 19.8  4.36
A 17.9  4.42  B 17.9  4.33  C 15.9  4.09
A 16.2  4.43  B 19.2  4.46  C 15.2  3.70
A 18.7  4.61  B 18.7  4.08  C 20.7  4.41
A 16.4  4.01  B 18.4  4.42  C 15.4  3.94
```

Archivo de resultados (C11-4.LST).-

Coeficiente de regresión en cada dieta

---------- DIETA=A ----------

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.13952	0.13952	2.207	0.1757
Error	8	0.50569	0.06321		
C Total	9	0.64521			

Root MSE	0.25142	R-square	0.2162
Dep Mean	4.35700	Adj R-sq	0.1183
C.V.	5.77047		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.211064	0.77542867	4.141	0.0032
X	1	0.062619	0.04214983	1.486	0.1757

---------- DIETA=B ----------

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.34428	0.34428	5.363	0.0492
Error	8	0.51353	0.06419		
C Total	9	0.85781			

Root MSE	0.25336	R-square	0.4013
Dep Mean	4.21300	Adj R-sq	0.3265
C.V.	6.01379		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	2.138643	0.89929014	2.378	0.0447
X	1	0.111525	0.04815667	2.316	0.0492

---------- DIETA=C ----------

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.20219	0.20219	8.132	0.0214
Error	8	0.19890	0.02486		
C Total	9	0.40109			

Root MSE	0.15768	R-square	0.5041
Dep Mean	4.09100	Adj R-sq	0.4421
C.V.	3.85429		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	2.814165	0.45051176	6.247	0.0002
X	1	0.070156	0.02460131	2.852	0.0214

Homogeneidad de coeficientes de regresión

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.04056813	0.20811363	4.10	0.0078
Error	24	1.21812853	0.05075536		
Corrected Total	29	2.25869667			

R-Square	C.V.	Root MSE	Y Mean
0.460694	5.338192	0.22529	4.22033

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DIETA	2	0.05232051	0.02616026	0.52	0.6037
X	1	0.67380185	0.67380185	13.28	0.0013
X*DIETA	2	0.04202125	0.02101062	0.41	0.6657

Si se tiene el caso particular de sólo dos rectas de regresión y se quiere hacer el contraste de homogeneidad tal como se ha visto más arriba, este equivale a la prueba *t*

$$t = \frac{b_1 - b_2}{\sqrt{S_p^2\left(\frac{1}{SC_{(X_1)}} + \frac{1}{SC_{(X_2)}}\right)}}$$

se distribuye como una *t* de *Student* con $gl = n_1+n_2-4$.

Siendo S_p las sumas de cuadrados residuales combinadas de las dos regresiones independientes, dividido por los grados de libertad combinados

$$S_p^2 = \frac{[SC_{(Y_1)} - SC_{(\text{regresión.1})}] + [SC_{(Y_2)} - SC_{(\text{regresión.2})}]}{n_1 - 2 + n_2 - 2}$$

Bibliografía

Dagnelie, P. 1970. THÉORIE ET MÉTHODES STATISTIQUES. Ed J. Duculot, S.A. Gembloux.

Freund, R.J., and Littel, R.C. 1991. *SAS* SYSTEM FOR REGRESION. SAS Institute Inc., Cary, NC, USA.

*Infante Gil, S. y Zárate De Lara, G.P.*1984. METODOS ESTADISTICOS. Ed. TRILLAS. México.

Lite, TM, y Jackson Hills, F. 1987. METODOS ESTADISTICOS PARA LA INVESTIGACION EN LA AGRICULTURA. Ed TRILLAS. México.

Ostle, B. 1965. ESTADISTICA APLICADA. Ed. Limusa-Wiley. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

Littel, R.C., Freund, R.J. and Spector, P.C. 1991. *SAS* FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. 1990. *SAS* PROCEDURE GUIDE. Cary, NC, USA.

CAPÍTULO 12

Predicciones y Ajustes

CAPÍTULO 12

Predicciones y Ajustes

Pruebas de hipótesis e intervalos de confianza de la predicción, por extrapolación, de un valor *Y*.-

Entre los usos de la regresión está la predicción de valores de *Y* para valores de *X* fuera de su campo de variación. Estos pueden ser valores futuros o bien pueden ser valores de *X* que son posibles observar pero es imposible o poco práctico medir el correspondiente valor de *Y*.

Lo que se ha hecho en el epígrafes *Pruebas de hipótesis e intervalos de confianza de la estimación, por interpolación, de un valor*, era estimar un valor de *Y*, esto es, $\hat{Y}$, entre dos valores de *X* observados, es decir, se ha estimado $\hat{Y}$ por *interpolación*. Mientras que ahora de trata de predecir un valor futuro cuya *X* aún no existe o, si existe, es inasequible como consecuencia del valor de *Y*.

Es importante que no se confundan los dos tipos de predicción. Si se estima, en el ejemplo anterior, la cantidad de pienso consumido en 350 en gallinas de peso 5.0, se está estimando el consumo medio de pienso esperado en gallina de un peso dado. Pero se podría haber planteado la cuestión como la predicción del consumo de pienso de una nueva gallina de peso superior al de las gallinas actuales.

Al predecir un valor o al estimar una media $\mu_{Y.X}$ para un *X* fuera del intervalo observado, esto es, al *extrapolar*, se supone que la relación se mantiene lineal y este supuesto puede no ser acertado, especialmente si se usa la recta como una aproximación y se extrapolan puntos muy alejados del rango muestreado. Para extrapolar con cierta garantía hay que revisar la recta de regresión a medida que se acumulan experiencias.

La predicción se hace igual que la estima de $\hat{Y}$, es decir,

$$\hat{Y}_i = a + bX_i$$

pero el error típico de la predicción es, ahora

$$S_{\hat{Y}.X_i} = \sqrt{S^2_{Y.X}\left(1+\frac{1}{n}+\frac{(X_i-\overline{X})^2}{SC_{(X)}}\right)}$$

Por lo que se puede realizar pruebas de hipótesis para contrastar la hipótesis nula de que $\hat{Y}$ es una estima de $\mu_{Y.Xi}$ por medio de la siguiente prueba *t*.

$$H_o : \hat{Y} = \mu_{Y.X}$$
$$H_1 : \hat{Y} \neq \mu_{Y.X}$$
$$t_o = \frac{\hat{Y} - \mu_{X.Y}}{S_{\hat{Y}.X_i}}$$

Esta t_o se distribuye como la *t* de *Student* con *n*-2 grados de libertad.

El intervalo de confianza sería

$$LC_{(\hat{Y}.X_i)} = \hat{Y} \pm t_{(n-2,\,\alpha/2)} S_{(\hat{y}.X_i)}$$

Ejemplo.-

Supóngase que se está en el año 1988. Se tiene el censo de cabras en España durante los años 1984 a 1988 en miles de cabezas, y se quiere saber cual será el censo de la cabaña caprina en 1989.

Año	*Cabras*	
1984	2533	$SC_{(X)} = 10$
1985	2584	$b = 253.60$
1986	2850	$a = -500749$
1987	2888	
1988	3649	$Y = -500749 + 253.6\ X$
9930	14504	

Se observa que la cantidad de cabras va aumentando. Como el valor de *X* se ha tomado con el número real del año (podríamos haber puesto, sencillamente, desde el año 1 al 5), la ordenada en el origen nos da la cantidad estimada de cabras en España el año cero.

El censo de cabras estimado para el año próximo es

$$\hat{Y}_{1989} = -500749+253.60\times 1989 = 3661.4$$

y el error típico de esta predicción es

$$S_{\hat{Y}.X_{1989}} = \sqrt{51685.7333\left(1+\frac{1}{5}+\frac{(1989-1986)^2}{10}\right)} = 329.4542$$

Por lo que el intervalo de confianza al 95% es

$$LC_{(\hat{Y}.1989)} = 3661.4 \pm 3.182 \times 329.4542 = 3661.4 \pm 1048.3233$$
$$L_i = 2613.1$$
$$L^s = 4709.7$$

Cuando llegó el año 1989 el censo de cabras en España era de 3780, lo cual cae dentro de la predicción que se hizo (L_i<3661<L^s) aunque la estimación puntual dio 3661-3780=-119 mil cabras por defecto.

Para predecir una media futura de k datos se usa también el valor de la regresión, aunque el error típico adecuado para la predicción viene dado por una ecuación semejante a la anterior pero sustituyendo el 1 del interior del paréntesis por $1/k$, de manera que queda

$$S_{\hat{Y}.X_I} = \sqrt{S^2_{Y.X}\left(\frac{1}{k}+\frac{1}{n}+\frac{(X_i-\overline{X})^2}{SC_{(X)}}\right)}$$

Ejemplo.-

Sigamos con el ejemplo de los pollos. Supongamos que se han conseguidos gallinas de peso 3 y se va a realizar una experiencia con 5 grupos de dichas gallinas y se quiere estimar la cantidad de pienso que consumirán en 350, con un 95% de confianza.

El error típico en este caso es

$$S_{\hat{Y}.X_3} = \sqrt{1.0555\left(\frac{1}{5}+\frac{1}{10}+\frac{(3-2.248)^2}{0.3286}\right)} = 1.4605$$

Por lo que los límites de confianza al 95% son

$$\hat{Y}_3 = 25.115+7.704\times 3 = 48.227$$
$$LC_{(\hat{Y}.3)} = 48.227 \pm 2.306 \times 1.4605 = 48.227 \pm 3.3679$$
$$L_i = 44.8591$$
$$L^s = 51.5949$$

Si la experiencia se realizara con un solo grupo de gallinas de peso medio 3, el error típico se calcularía como en el ejemplo de las cabras.

$$S_{\hat{Y}.X_3} = \sqrt{1.0555\left(1+\frac{1}{10}+\frac{(3-2.248)^2}{0.3286}\right)} = 1.7255$$

Por lo que los límites de confianza al 95% son

$$\hat{Y}_3 = 25.115+7.704\times 3 = 48.227$$
$$LC_{(\hat{Y}.3)} = 48.227 \pm 2.306\times 1.7255 = 48.227 \pm 3.979$$
$$L_i = 44.2480$$
$$L^s = 52.2060$$

El *SAS* provee estos intervalos (*Predict*) de confianza de la predicción, así como los intervalos de la estimación por interpolación (*Means*).

Archivo del programa *SAS* (C12-1.SAS).-

```
title .Limites de confianza para valores interpolados y extrapolados.;
options ls=75 ps=60;
data limites;
infile 'c12-1.dat';
input peso alimento n;
proc reg;
weight n;
  model alimento = peso / clm cli;
run;
```

Archivo de datos (C12-1.DAT).-

```
2.11    40.73    1
1.99    41.45    1
2.31    42.23    1
2.67    45.13    1
2.13    41.77    1
2.31    42.82    1
2.08    39.50    1
2.31    43.31    1
2.35    45.04    1
2.22    42.36    1
2         .      1
3         .      5
3         .      1
```

Archivo de resultados (C12-1.LST).-

```
Model: MODEL1
Dependent Variable: ALIMENTO

                          Sum of          Mean
  Source            DF    Squares        Square      F Value      Prob>F
  Model              1   19.50448      19.50448       18.488      0.0026
  Error              8    8.43976       1.05497
  C Total            9   27.94424
```

Root MSE	1.02712	R-square	0.6980
Dep Mean	42.43400	Adj R-sq	0.6602
C.V.	2.42051		

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	25.113672	4.04125572	6.214	0.0003
PESO	1	7.704772	1.79189594	4.300	0.0026

Obs	Weight	Dep Var ALIMENTO	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean
1	1.0000	40.7300	41.3707	0.408	40.4294	42.3121
2	1.0000	41.4500	40.4462	0.565	39.1433	41.7491
3	1.0000	42.2300	42.9117	0.343	42.1201	43.7033
4	1.0000	45.1300	45.6854	0.823	43.7876	47.5832
5	1.0000	41.7700	41.5248	0.388	40.6311	42.4186
6	1.0000	42.8200	42.9117	0.343	42.1201	43.7033
7	1.0000	39.5000	41.1396	0.443	40.1184	42.1608
8	1.0000	43.3100	42.9117	0.343	42.1201	43.7033
9	1.0000	45.0400	43.2199	0.373	42.3604	44.0793
10	1.0000	42.3600	42.2183	0.329	41.4604	42.9761
11	0	.	40.5232	0.550	39.2539	41.7925
12	0	.	48.2280	1.386	45.0316	51.4243
13	0	.	48.2280	1.386	45.0316	51.4243

Obs	Lower95% Predict	Upper95% Predict	Residual
1	38.8220	43.9195	-0.6407
2	37.7429	43.1494	1.0038
3	40.4144	45.4090	-0.6817
4	42.6503	48.7205	-0.5554
5	38.9933	44.0564	0.2452
6	40.4144	45.4090	-0.0917
7	38.5603	43.7189	-1.6396
8	40.4144	45.4090	0.3983
9	40.7002	45.7395	1.8201
10	39.7314	44.7051	0.1417
11	37.8360	43.2104	.
12	44.8607	51.5953	.
13	44.2497	52.2063	.

Sum of Residuals	0
Sum of Squared Residuals	8.4398
Predicted Resid SS (Press)	14.4151

NOTE: The above statistics use observation weights or frequencies.

Predicción de X a partir de Y. Calibración lineal.-

En algunas ocasiones nos encontramos con problemas en los que conocemos el valor de *Y* para un individuo y deseamos calcular el valor correspondiente de *X* para dicho individuo. Por ejemplo, *X* puede ser la concentración de cualquier elemento químico o molécula orgánica en un líquido o en un alimento, e *Y* puede ser la cantidad de absorción de rayos de cierta longitud de onda (por ejemplo, rayos infrarrojos) de dicho líquido o alimento. El experimentador prepara cierta cantidad de muestras con cantidades conocidas de *X* que cubran todo el rango posible, y mide *Y* para cada muestra. Con estos datos se obtiene una recta de calibración que le permitirá obtener la concentración de cierto compuesto o molécula por la cantidad de absorción medida en un espectrofotómetro, sin necesidad de realizar el análisis químico de laboratorio.

Se podría pensar que este problema es bastante simple, bastaría con hallar una nueva regresión con X como variable dependiente. Sin embargo esto es incorrecto, debido a que los supuestos iniciales eran que X estaba medida sin error y que Y es la variable aleatoria distribuida normalmente.

Dado que

$$\hat{Y}_i = a + bX_i$$

podemos calcular X_i reordenando esta ecuación y obteniendo

$$\hat{X}_i = \frac{(Y_i - a)}{b}$$

Para determinar los límites de confianza para esta X, se utiliza los límites de confianza de Y prevista del epígrafe anterior, teniendo en cuenta, también, el error típico del coeficiente de regresión. Como quiera que todos estos requisitos da una fórmula un poco complicada se subdivide en las siguientes cantidades

$$D = b^2 - t^2_{(\alpha;n-2)} S^2_b$$

$$H = \frac{t_{(\alpha;n-2)}}{D} \sqrt{S^2_{Y.X}\left[D\left(1+\frac{1}{n}\right)+\frac{(Y_i-\overline{Y})^2}{SC_{(X)}}\right]}$$

Siendo los límites de confianza:

$$LC = \overline{X} + \frac{b(Y_i - \overline{Y})}{D} \pm H$$

Como se ve, estos límites de confianza no cumplen una característica que hasta el momento cumplen todos los límites de confianza que se han estudiado, a saber, que no son simétricos con respecto al valor de $\hat{X}_i$, sino que son simétricos respecto al valor

$$\overline{X} + \frac{b(Y_i - \overline{Y})}{D}$$

que es muy cercano a $\hat{X}_i$ pero no idéntico a él.

Ejemplo.-

El ejemplo del siguiente epígrafe incluye un ejemplo de calibración.

Ajuste de la recta por el origen.-

En algunos problemas de regresión, la característica de los datos o consideraciones teóricas exigen que la recta pase por el origen de coordenadas, es decir, que cuando X=0, Y debe ser igual a cero. En tales casos, el ajuste satisfactorio es

$$Y = \beta X + \varepsilon$$

para el cual el estimador mínimo cuadrado de β es

$$b = \frac{\Sigma_i X_i Y_i}{\Sigma_i X_i^2}$$

teniendo como característica, este ajuste mínimo cuadrático, que la suma de las desviaciones respecto a esta recta no vale cero.

La descomposición de la suma de cuadrados total sería

$$SC_{\text{regresión}} = b\Sigma_i X_i Y_i = \frac{(\Sigma_i X_i Y_i)^2}{\Sigma_i X_i^2}$$

$$SC_{\text{error}} = \Sigma_i Y_i^2 - b\Sigma_i X_i Y_i = \Sigma_i Y_i^2 \frac{(\Sigma_i X_i Y_i)^2}{\Sigma X_i^2}$$

$$SC_{\text{total}} = \Sigma_i Y_i^2$$

Dado que no se ha hecho ajuste para la media (no se le ha restado el término de corrección), la suma de cuadrados total tiene n grados de libertad y, por lo tanto, el residuo tendrá n-1 grados de libertad, quedando el cuadrado medio del error

$$S_{y.x}^2 = \frac{\Sigma_i Y_i^2 - b\Sigma_i X_i Y_i}{n-1}$$

y la varianza de la regresión

$$S_b^2 = \frac{S_{X.Y}^2}{\Sigma_i X_i^2}$$

Con estas ecuaciones se pueden reproducir todas las pruebas de hipótesis e intervalos de confianza vistos anteriormente, con la particularidad de que los grados de libertad de la pruebas serán de n-1.

Este modelo no deber adoptarse si no es después de una inspección cuidadosa de los datos, ya que puede surgir complicaciones. Si no hay observaciones cercanos a X=0 se debe hacer el ajuste a una recta con ordenada en el origen y realizar la prueba de hipótesis de que esta ordenada en el origen es cero (tal como se ha visto

anteriormente), para en caso de confirmarse la hipótesis, realizar el nuevo ajuste por el origen. Podría ocurrir, que aunque teóricamente para X=0, Y=0, el ajuste a una recta que pase por el origen, sea un mal ajuste, porque en realidad el buen ajuste sea una línea curva. Línea que puede ser más curva cuanto más próxima se encuentra del origen, mientras que es ligeramente curva en el rango de X que han sido medidas, por lo que una recta con ordenada en el origen es una aceptable aproximación, pero no es fiable para extrapolar hacia el origen.

Ejemplo.-

La predicción del contenido proteico de la leche se puede realizar usando técnicas muy fáciles y rápidas de aplicar como la absorbancia de rayos de longitud de onda de infrarrojo cercano. Con éste objetivo, se toman muestras de leche de muy diferente origen, con objeto de cubrir todo el rango de concentración de proteína, se analiza en el laboratorio la cantidad de proteína por unidad de volumen de cada muestra y se mide la cantidad de absorbancia, de dichas muestras, en un espectrofotómetro. Los datos de 20 de dichas muestras son

Proteína	*Absorbancia*	*Proteína*	*Absorbancia*
2.2364	0.2421	1.7331	0.4440
2.7789	0.2590	2.4463	0.4375
1.9522	0.2166	1.7045	0.4662
2.6812	0.1704	2.4691	0.5275
2.1099	0.1779	2.3606	0.5253
2.1808	0.1786	2.3931	0.5094
1.4619	0.1487	2.4587	0.5396
1.2254	0.1306	2.5617	0.5744
1.7761	0.1379	2.8031	0.5854
1.0286	0.1345	3.1150	0.5963

$$SP = 16.359 - \frac{7.0 \times 43.48}{20} = 1.138$$

$$SC_{(X)} = 100.24 - \frac{(43.48)^2}{20} = 5.73$$

$$SC_{(Y)} = 3.08 - \frac{(7.0)^2}{20} = 0.629$$

$$b = \frac{1.138}{5.73} = 0.199$$

$$a = 2.174 - 0.35 \times 1.99 = -0.082$$

$$S^2_{Y.X} = \frac{0.629 - 0.199 \times 1.138}{18} = 0.022$$

$$R^2 = \frac{0.199 \times 1.138}{0.629} = 0.359$$

La prueba de ajuste es

FV	*gl*	*SC*	*CM*	F_o
Regresión	1	0.2265	0.2265	10.295**
Error	18	0.4025	0.0220	
Total	19	0.6290		

Por lo que la ecuación de la recta

$$Y = -0.082+0.199\ X$$

provee un buen ajuste.

Por tanto, si se tiene una muestra de leche de contenido proteico desconocido pero con una absorbancia de 0.370, el contenido proteico de dicha muestra se puede estimar con la ecuación anterior, despejando X

$$\hat{X}_i = \frac{(0.370+0.082)}{0.199} = 2.27$$

Siendo sus límites de confianza

$$D = 0.199^2 - 2.1^2 \times 0.062^2 = 0.023$$

$$H = \frac{2.1}{0.023}\sqrt{0.022\left[0.023\left(1+\frac{1}{20}\right)+\frac{(0.37-0.35)^2}{5.73}\right]} = 2.108$$

$$LC = 2.174+\frac{0.199(0.37-0.35)}{0.023}\times 2.108 = 2.347 \pm 2.108$$

$$L_i = 0.239$$

$$L^s = 4.45$$

Es decir, el contenido proteico de una muestra de leche con 0.370 de absorbancia es de 2.27, con un mínimo de 0.24 y un máximo de 4.45.

Pero en este caso, es razonable pensar que, teóricamente, si hay cero proteínas la absorción debe ser cero, por lo que se debería de realizar el ajuste por el origen. Comprobemos previamente si el valor de ordenada en el origen es cero

$$H_o : \alpha = 0$$
$$H_1 : \alpha \neq 0$$
$$S_a = \sqrt{0.022\left(\frac{1}{20}+\frac{2.174^2}{5.73}\right)} = 0.140$$
$$t_o = \frac{-0.082-0}{0.14} = -0.584 ns$$
$$t_{(18;\,0.05/2)} = 2.1009$$

Como *a* no es estadísticamente diferente de cero, se puede realizar el ajuste por el origen.

La pendiente del ajuste por el origen es

$$\Sigma_i X_i Y_i = 16.359$$
$$\Sigma_i X_i^2 = 100.24$$
$$\Sigma_i Y_i^2 = 3.08$$
$$b = \frac{16.359}{100.24} = 0.163$$

La descomposición de la suma de cuadrados sería

$$SC_{\text{regresión}} = 0.163 \times 1.6359 = 2.67$$
$$SC_{\text{error}} = 3.08 - 2.67 = 0.41$$
$$SC_{\text{total}} = 3.08$$
$$R^2 = 0.87$$

La prueba de ajuste es

FV	*gl*	*SC*	*CM*	F_o
Regresión	1	2.67	2.670	121.37***
Error	19	0.41	0.022	
Total	20	3.08		

Tanto el coeficiente de determinación como la prueba de ajuste da como resultado un mejor ajuste de estos datos a la recta que pasa por el origen, por lo que la ecuación de la recta sería

$$Y = 0.163\, X$$

Siendo la estima del contenido proteico de una muestra de absorbancia 0.37

$$\hat{X}_i = \frac{0.370}{0.163} = 2.27$$

Siendo sus límites de confianza

$$D = 0.163^2 - 2.09^2 \times 0.015^2 = 0.026$$

$$H = \frac{2.09}{0.026}\sqrt{0.022\left[0.026\left(1+\frac{1}{20}\right)+\frac{(0.37-0.35)^2}{5.73}\right]} = 1.972$$

$$LC = 2.174+\frac{0.163(0.37-0.35)}{0.026} \pm 1.972 = 2.299 \pm 1.972$$

$$L_i = 0.33$$

$$L^s = 4.27$$

Con el *SAS* habría que hacer un primer programa para constatar el buen ajuste por el centro de coordenadas y un segundo programa (o una segunda parte del mismo programa) para estimar el valor de *Y* dado el valor determinado de *X*, por ejemplo *X*=0.37.

Archivo del programa *SAS* (C12-2.SAS).-

```
options ls=75 ps=60;
data cali;
infile 'c12-2.dat';
input prot absor @@;
title .Ajuste a la recta con ordenada en el origen.;
proc reg;
  model absor = prot;
run;
title .Ajuste a la recta pasando por el origen.;
  model absor = prot / noint;
run;
```

Archivo de datos DAT (C12-2.DAT).-

```
   2.2364   0.2421   1.7331   0.4440
   2.7789   0.2590   2.4463   0.4375
   1.9522   0.2166   1.7045   0.4662
   2.6812   0.1704   2.4691   0.5275
   2.1099   0.1779   2.3606   0.5253
   2.1808   0.1786   2.3931   0.5094
   1.4619   0.1487   2.4587   0.5396
   1.2254   0.1306   2.5617   0.5744
   1.7761   0.1379   2.8031   0.5854
   1.0286   0.1345   3.1150   0.5963
```

Archivo de resultados (C12-2.LST).-

```
                .Ajuste a la recta con ordenada en el origen.

Model: MODEL1
Dependent Variable: ABSOR

                          Sum of          Mean
  Source           DF     Squares        Square      F Value       Prob>F
  Model             1     0.22603       0.22603       10.095       0.0052
  Error            18     0.40305       0.02239
  C Total          19     0.62908

      Root MSE        0.14964     R-square        0.3593
      Dep Mean        0.35010     Adj R-sq        0.3237
      C.V.           42.74198

                        Parameter Estimates

                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   INTERCEP   1     -0.081689     0.13995972        -0.584        0.5667
   PROT       1      0.198628     0.06251697         3.177        0.0052

                .Ajuste a la recta pasando por el origen
Model: MODEL2
NOTE: No intercept in model. R-square is redefined.
Dependent Variable: ABSOR

                        Analysis of Variance
                          Sum of          Mean
  Source           DF     Squares        Square      F Value       Prob>F
  Model             1     2.66973       2.66973      123.517       0.0001
  Error            19     0.41067       0.02161
  U Total          20     3.08041

      Root MSE        0.14702     R-square        0.8667
      Dep Mean        0.35010     Adj R-sq        0.8597
      C.V.           41.99382

                        Parameter Estimates
                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   PROT       1      0.163198     0.01468423        11.114        0.0001
```

Para estimar el valor de Y dado el valor de, por ejemplo, X=0.37, se tiene el siguiente programa SAS.

Archivo del programa *SAS* (C12-3.SAS).-

```
Title 'Absorbancia y límites de confianza para X=0.37';
options ls=75 ps=60;
data cali;
infile 'c12-2.dat';
input prot absor @@;
abs_o = 0.37;
sumn+1;
sumx +prot;
sumxx +prot*prot;
mprot = sumx/sumn;
sumy +absor;
sumyy +absor*absor;
mabsor = sumy/sumn;
sumxy +prot*absor;
sp = sumxy-(sumx*sumy/sumn);
scx = sumxx - (sumx*sumx/sumn);
scy = sumyy - (sumy*sumy/sumn);
b = sumxy/sumxx;
cme = (sumyy-(b*sumxy))/(sumn-1);
varb = cme/sumxx;
cmep = (scy-(b*sp))/(sumn-2);
prot_est = abs_o/b;
pt= tinv(0.975,sumn-2);
d = b*b - pt*pt*varb;
desme = abs_o - mabsor;
h = pt/d * sqrt(cme*(d*((1+1/sumn)+(desme*desme/scx))));
ls = mprot + b*desme/d + h;
li = mprot + b*desme/d - h;
proc print;
var li prot_est ls;
run;
```

Si cambia el problema solo hay que cambiar el valor de la absorbancia en la línea 6ª del programa (abs_o = ...;)

Archivo de resultados (C12-3.LST).-

Absorbancia y límites de confianza para X=0.37

OBS	LI	PROT_EST	LS
...	...	.....	....
13	-0.0688	2.50155	5.031
14	-0.0305	2.40177	4.791
15	0.0974	2.41733	4.732
16	0.1072	2.32685	4.547
17	0.1563	2.37380	4.624
18	0.1921	2.30501	4.460
19	0.2738	2.31002	4.403
20	0.3253	2.26719	4.275

Como siempre en estos casos, hay que mirar sólo la última línea.

Transformación *Probit* en los casos de predicción de *X* a partir de una *Y* no normal.-

El método de calcular X a partir de Y, también denominado método de *predicción inversa*, se aplica frecuentemente en el análisis estadístico de problemas sobre mortalidad por dosis en bioensayos. Este tipo de estudio implica una regresión de mortalidad de organismos sobre dosis de una sustancia tóxica. Así, la dosis X_1 causará un porcentaje de mortalidad de Y_1; la dosis $X_2 > X_1$, un porcentaje de mortalidad $Y_2 > Y_1$, y así sucesivamente. Cuando se alcanza una cierta dosis, la mortalidad es entonces del 100%; la muestra completa de organismos ha muerto. Una medida común de la potencia de la sustancia tóxica (o tolerancia de los organismos) es la dosis requerida para matar al 50% o al 95% de los organismos. Tales puntos se denominan dosis letal para el 50% de los organismos o para el 95%, simbolizado como DL_{50} (dosis letal 50) y DL_{95} (dosis letal 95), respectivamente. Este problema, como se ve, es claramente de predicción inversa: ¿Cuál es la dosis necesaria (X) para que se produzca la mortalidad (Y) del 50% o del 95% de los organismos?

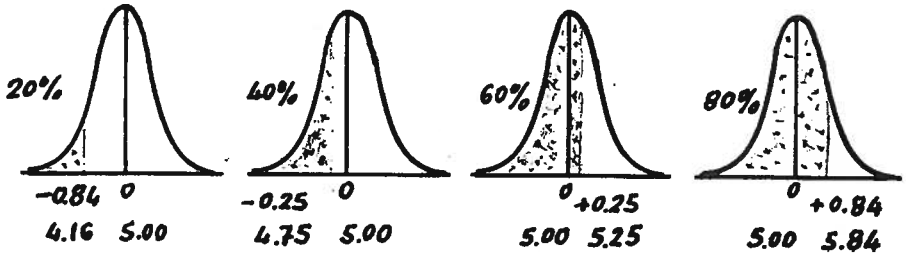
La prueba de hipótesis y los límites de confianza para dicha estimación se construyen de la misma manera indicada dos epígrafes más arriba. El único problema es que como variable dependiente se tiene una proporción, es decir, una variable no normal, lo que, además de no cumplir los supuestos paramétricos, puede dar como resultado que la línea resultante no sea recta. Es por todo ello por lo que los datos se transforman de manera que las mortalidades se transforman a una escala normal, denominada *probit* (valor Z que delimita el área dada por la proporción de organismos muertos) y las dosis a escala logarítmica, a fin de que la regresión de la mortalidad sobre las dosis sea lineal.

La transformación probit consiste en lo siguiente: la tabla de distribución normal (tabla Z) nos da el área bajo la curva normal para un determinado valor de Z de abscisas. Es decir, nos da la probabilidad o frecuencia relativa que delimitan unos determinados límites de Z. La denominada transformación probit no es nada más que la operación inversa, es decir, conocida el área, probabilidad o frecuencia relativa, saber a qué valor de Z le corresponde. Por tanto, estos valores probit se extraen de la tabla Z (Tabla 1). Pero como la media de los valores Z es 0, para no trabajar con números negativos, al valor Z obtenido se le sumará 5. Por lo que para un área determinada bajo la curva normal, el valor correspondiente que se toma es

$$\Pr obit\ Y = 5+Z$$

por lo tanto, el *probit* es una variable normal con media 5 y varianza uno.

En la siguiente figura tenemos ejemplos de esta transformación



Cuando el porcentaje bajo la curva normal es del 20%, el valor $Z = -0.84$ y el valor probit que se toma es 5 - 0.84 = 4.16.

A causa de la simetría de la distribución, las Z correspondientes al 20% y 80% tienen el mismo valor numérico pero con signo opuesto. Por lo tanto, sus correspondientes valores probit sumarán 10. Cuando dos porcentajes cualesquiera sumen 100% sus correspondientes valores probit sumarán 10.

La razón por la que se utiliza la transformación probit para los datos en los que intervienen conteos, especialmente en los datos de mortalidad, es muy sencilla. En experimentos de toxicología, por ejemplo, se cuenta el número de muertos y se calcula la proporción de estos en el grupo de animales utilizados. Se supone que la resistencia innata de los animales a la sustancia tóxica se distribuye normalmente. Cuando la proporción de muertes observada es del 40%, por ejemplo, significa que han muerto aquellos animales que tienen una resistencia inferior a un valor de Z de -0.25 (Y=4.75) y que los que tienen mayor resistencia han sobrevivido. Si una dosis mayor de sustancia tóxica mata el 60% de un grupo similar de animales, es decir, un grupo con la misma distribución de tolerancia, significa que mata a los que tienen una resistencia inferior al valor de 5.25 de probit. Para indicar la toxicidad de la sustancia, en lugar de utilizar directamente las proporciones, utilizamos el valor de probit.

Si Y indica el valor de probit y X la dosis (en la misma escala), la transformación será especialmente útil cuando se haga que Y sea linealmente dependiente de X.

Es bien conocida la dificultad de comparar porcentajes. Por ejemplo, una disminución de la mortalidad del 40 al 20% no significa que la disminución, numéricamente igual, del 20% al 1% sea biológicamente igual, y no digamos la disminución del 20% al 0%. La transformación probit soslaya la dificultad de trabajar con porcentajes.

Pero tenemos que salvar otra dificultad, como es que no todos los valores de probit tienen el mismo *peso*. Esto se debe al hecho de que las proporciones intermedias se determinan con mayor exactitud que las proporciones muy pequeñas o muy grandes por lo que tendremos que el *coeficiente de ponderación* de los diversos porcentajes es

$$W_i = \frac{h_i^2}{p_i \times q_i}$$

Siendo h_i la ordenada del valor Z_i del porcentaje *i-ésimo*. Esta ordenada o altura se calcula

$$h_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}$$

El peso o frecuencia de cada proporción o dato, sería igual a $F_i = W_i{}^* n_i$; siendo n_i el número de individuos totales en que se ha probado la *i-ésima* dosis.

Como se ha dicho anteriormente, debido a la simetría de la curva normal, el peso para p=0.20, por ejemplo, es el mismo que el p=0.8.

Una vez hallados los valores probit correspondientes a los porcentajes de mortalidad, que serán los valores de Y, y los logaritmos de las dosis, como valores de X, se realiza el análisis de regresión normal, tal como se ha realizado en los epígrafes anteriores, pero teniendo en cuenta el peso o frecuencia de cada par de valores, es decir, como si el *i-ésimo* par de valores se hubiera observado F_i veces. Entonces, la suma total de los pesos o frecuencias será el tamaño de muestra.

Por tanto, los sumatorios básicos que se necesitan son: $\Sigma_i X_i F_i$, $\Sigma_i X^2_i F_i$, $\Sigma_i Y_i F_i$, $\Sigma_i Y^2_i F_i$ y $\Sigma_i X_i Y_i F_i$. Siendo el cálculo de los estadísticos como sigue

$$\overline{X} = \frac{\Sigma_i X_i F_i}{\Sigma_i F_i}$$

$$\overline{Y} = \frac{\Sigma_i Y_i F_i}{\Sigma_i F_i}$$

$$SP_{(XY)} = \Sigma_i X_i Y_i F_i - \frac{\Sigma_i X_i F_i \Sigma_i Y_i F_i}{\Sigma_i F_i}$$

$$SC_{(X)} = \Sigma_i X_i^2 F_i - \frac{(\Sigma_i X_i F_i)^2}{\Sigma_i F_i}$$

$$SC_{(Y)} = \Sigma_i Y_i^2 F_i - \frac{(\Sigma_i Y_i F_i)^2}{\Sigma_i F_i}$$

$$b = \frac{SP}{SC_{(X)}}$$

$$a = \overline{X} - b\overline{Y}$$

La prueba de ajuste es la misma que la presentada en el Capítulo 11 pero teniendo en cuenta que, dada las transformaciones realizadas, las sumas de cuadrados son en realidad *ji-cuadrados* (χ^2)

FV	*gl*	χ^2
Regresión	1	$b\,SP_{(XY)}$
Error	*n*-2	$SC_{(Y)}-bSP_{(XY)}$
Total	*n*-1	$SC_{(Y)}$

siendo, en este caso, *n* igual al número de pares de valores, sin ponderación.

Una vez estimada la recta de regresión y comprobado que el ajuste es significativamente bueno, se puede estimar las dosis deseadas despejando la *X* de la recta y sustituyendo la *Y* por la mortalidad deseada, en valores probit.

Ejemplo.-

Se probó la toxicidad de la ouabaina en ranas (*Rana pipiens*), dando los siguientes resultados

Dosis	*n*	*Muertes*	*P*	*Z*	*Probit*	*W*	*F*
0.000200	20	5	0.25	-0.68	4.32	0.5346	10.69
0.000225	20	10	0.50	0.00	5.00	0.6366	12.73
0.000250	20	13	0.65	0.39	5.39	0.6009	12.02
0.000300	18	14	0.78	0.78	5.78	0.5047	9.08
0.000350	10	8	0.80	0.85	5.85	0.4830	4.83
0.000400	10	9	0.90	1.29	6.29	0.3349	3.35
							52.71

logdosi	*logd* × *F*	*prob* × *F*	$logd^{2}\times F$	$prob^{2}\times F$	*log*×*prob*×*F*
-3.6990	-39.547	46.186	146.282	119.525	-170.842
-3.6478	-46.445	63.662	169.425	318.310	-232.227
-3.6021	-43.287	64.774	155.924	349.131	-233.319
-3.5229	-32.007	52.515	112.758	303.535	-185.003
-3.4559	-16.692	28.255	57.687	165.295	-97.649
-3.3979	-11.379	21.065	38.668	132.500	-71.578
	-189.358	276.458	680.744	1468.296	-990.619

$$\overline{X} = \frac{-189.3583}{52.71} = -3.5927$$

$$\overline{Y} = \frac{276.4577}{52.71} = 5.2453$$

$$SP_{(XY)} = -990.6193 - \frac{-189.3529 \times 276.4577}{52.71} = 2.6270$$

$$SC_{(X)} = 680.7493 - \frac{(-189.3529)^2}{52.71} = 0.4223$$

$$SC_{(Y)} = 1468.2962 - \frac{(276.4572)^2}{52.71} = 18.1904$$

$$b = \frac{2.6270}{0.4223} = 6.2207$$

$$a = 5.2453 - 6.2270 \times -3.5927 = 27.5944$$

La prueba de ajuste es

FV	*gl*	χ^2
Regresión	1	16.3418***
Error	4	1.8486*ns*
Total	5	18.1904***

El error sale no significativo, por lo que el ajuste es perfecto.

Ahora se puede estimar, por ejemplo, la dosis letal cincuenta,

$$DL_{50} = \frac{Y_{50} - a}{b} = \frac{5 - 27.3944}{6.2207} \quad -3.6321$$

como este resultado esta en logaritmo, se halla el antilogaritmo, dando como resultado que DL_{50} = antlog(-3.6321) = 0.000233

El ***SAS*** tiene un procedimiento especial para estos tipos de problemas.

Archivo del programa *SAS* (C12-4.SAS).-

```
title 'Probit';
option ls=75 ps=60;
data probit;
infile 'c12-4.dat';
input dosis n muertos;
proc probit log10;
 model muertos/n = dosis / lackfit inversecl itprint;
run;
```

Archivo de datos (C12-4.DAT).-

```
0.000200 20  5
0.000225 20 10
0.000250 20 13
0.000300 18 14
0.000350 10  8
0.000400 10  9
```

Archivo de resultados (C12-4.LST).-

```
                          Probit Procedure

Iter Ridge   LogLikelihood     INTERCPT Log10(DOSIS)
   0     0 -67.92842369487            0            0
   1     0 -57.03649605075 18.845309153 5.1913118344
   2     0 -56.75619293861 22.421752959 6.1741976482
   3     0 -56.75554070832 22.603499289 6.2241101474
   4     0 -56.75554070398 22.603968254 6.2242389088

 Data Set          =WORK.PROBIT
 Dependent Variable=MUERTOS
 Dependent Variable=N
 Number of Observations=   6
 Number of Events      =      59    Number of Trials =       98

 Log Likelihood for NORMAL  -56.7555407

                    Last Evaluation of the Gradient

                          INTERCPT      Log10(DOSIS)
                      -2.117771E-9      7.2615096E-9

                    Last Evaluation of the Hessian
                                  INTERCPT      Log10(DOSIS)
              INTERCPT          53.703747       -193.146071
         Log10(DOSIS)        -193.146071        695.064093

                       Goodness-of-Fit Tests

         Statistic                  Value      DF   Prob>Chi-Sq
         ------------------        --------    --   -----------
         Pearson Chi-Square          1.8135     4        0.7700
         L.R.    Chi-Square          1.8148     4        0.7698

         Response Levels:   2  Number of Covariate Values:    6

NOTE: Since the chi-square is small (p > 0.1000), fiducial limits will be
      calculated using a t value of  1.96.

       Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

       INTERCPT   1 22.6039683 5.602868  16.27604  0.0001 Intercept
       Log10(DOS) 1 6.22423891   1.5574  15.97247  0.0001

             Probit Model in Terms of Tolerance Distribution

                                MU         SIGMA
                           -3.6316      0.160662
```

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	0.000558	-0.000353
SIGMA	-0.000353	0.001616

Probit Analysis on Log10(DOSIS)

Probability	Log10(DOSIS)	95 Percent Fiducial Limits	
		Lower	Upper
0.01	-4.00536	-4.40341	-3.86626
0.02	-3.96156	-4.31800	-3.83634
0.03	-3.93378	-4.26388	-3.81729
0.04	-3.91287	-4.22321	-3.80291
0.05	-3.89587	-4.19016	-3.79118
0.06	-3.88140	-4.16207	-3.78116
0.07	-3.86871	-4.13746	-3.77236
0.08	-3.85735	-4.11544	-3.76445
0.09	-3.84701	-4.09545	-3.75724
0.10	-3.83750	-4.07706	-3.75058
0.15	-3.79812	-4.00117	-3.72276
0.20	-3.76682	-3.94128	-3.70023
0.25	-3.73997	-3.89034	-3.68046
0.30	-3.71585	-3.84514	-3.66216
0.35	-3.69351	-3.80394	-3.64451
0.40	-3.67231	-3.76578	-3.62684
0.45	-3.65179	-3.73017	-3.60843
0.50	-3.63160	-3.69698	-3.58845
0.55	-3.61141	-3.66637	-3.56590
0.60	-3.59090	-3.63856	-3.53968
0.65	-3.56970	-3.61346	-3.50895
0.70	-3.54735	-3.59043	-3.47314
0.75	-3.52324	-3.56837	-3.43170
0.80	-3.49639	-3.54599	-3.38337
0.85	-3.46509	-3.52162	-3.32532
0.90	-3.42571	-3.49243	-3.25080
0.91	-3.41619	-3.48554	-3.23265
0.92	-3.40586	-3.47810	-3.21287
0.93	-3.39450	-3.46998	-3.19108
0.94	-3.38181	-3.46096	-3.16667
0.95	-3.36734	-3.45074	-3.13878
0.96	-3.35033	-3.43881	-3.10594
0.97	-3.32943	-3.42422	-3.06548
0.98	-3.30164	-3.40495	-3.01158
0.99	-3.25785	-3.37477	-2.92643

Probability	DOSIS	95 Percent Fiducial Limits	
		Lower	Upper
0.01	0.0000988	0.0000395	0.0001361
0.02	0.0001093	0.0000481	0.0001458
0.03	0.0001165	0.0000545	0.0001523
0.04	0.0001222	0.0000598	0.0001574
0.05	0.0001271	0.0000645	0.0001617
0.06	0.0001314	0.0000689	0.0001655
0.07	0.0001353	0.0000729	0.0001689
0.08	0.0001389	0.0000767	0.0001720
0.09	0.0001422	0.0000803	0.0001749
0.10	0.0001454	0.0000837	0.0001776
0.15	0.0001592	0.0000997	0.0001893
0.20	0.0001711	0.0001145	0.0001994
0.25	0.0001820	0.0001287	0.0002087

0.30	0.0001924	0.0001428	0.0002177
0.35	0.0002025	0.0001571	0.0002267
0.40	0.0002127	0.0001715	0.0002361
0.45	0.0002229	0.0001861	0.0002464
0.50	0.0002336	0.0002009	0.0002580
0.55	0.0002447	0.0002156	0.0002717
0.60	0.0002565	0.0002298	0.0002886
0.65	0.0002693	0.0002435	0.0003098
0.70	0.0002836	0.0002568	0.0003364
0.75	0.0002998	0.0002702	0.0003701
0.80	0.0003189	0.0002845	0.0004136
0.85	0.0003427	0.0003009	0.0004728
0.90	0.0003752	0.0003218	0.0005613
0.91	0.0003835	0.0003269	0.0005853
0.92	0.0003928	0.0003326	0.0006125
0.93	0.0004032	0.0003389	0.0006441
0.94	0.0004151	0.0003460	0.0006813
0.95	0.0004292	0.0003542	0.0007265
0.96	0.0004463	0.0003641	0.0007835
0.97	0.0004683	0.0003765	0.0008600
0.98	0.0004993	0.0003936	0.0009737
0.99	0.0005523	0.0004219	0.00118

En el caso de que no se quiera tantas salidas se puede realizar el siguiente programa ***SAS***.

Archivo del programa *SAS* (C12-5.SAS).-

```
title 'Probit';
options ls=75 ps=60;
data probit;
infile 'c12-4.dat';
input dosis n muertos ;
Z=probit(muertos/n);
pprobit=5+z;
logdosis=log10(dosis);
h=(exp(-(Z*Z/2)))/sqrt(2*3.14159);
w=n*(h*h)/(muertos/n*(1-(muertos/n)));
sumgl+1;
sumw+w;
sumlp+ w*logdosis;
sumpp+ w*pprobit;
sumllp+ w*logdosis*logdosis;
sumppp+ w*pprobit*pprobit;
sumlpp+ w*logdosis*pprobit;
sp=sumlpp-(sumlp*sumpp)/sumw;
scx=sumllp-(sumlp*sumlp)/sumw;
scy=sumppp-(sumpp*sumpp)/sumw;
b=sp/scx;
a=(sumpp/sumw)-b*sumlp/sumw;
chi_err=scy-b*sp;
gl=sumgl-2;
prji2=probchi(chi_err,gl);
pr_chi=1-prji2;
ldl50=((sumpp/sumw)-a)/b;
dl50=10**(ldl50);
proc print;
var a b dl50 chi_err gl pr_chi;
run;
```

Archivo de resultados (C12-5.LST).-

Probit

OBS	A	B	DL50	CHI_ERR	GL	PR_CHI
1	-3.4230	-2.0948	.00020000	-0.00000	-1	.
2	53.0996	13.1859	.00021318	0.00000	0	.
3	44.8162	10.9353	.00022501	0.10826	1	0.74213
4	34.2786	8.0561	.00023876	0.75819	2	0.68448
5	28.7805	6.5475	.00024793	1.66634	3	0.64444
6	27.3630	6.1574	.00025569	1.80284	4	0.77196

Los resultados son los de la última línea, en este ejemplo, la línea 6.

El modelo logístico o LOGIT.-

Las situaciones en las que se aplica el modelo *LOGIT* es muy semejante al *PROBIT*, si bien ambos modelos son diferentes. Como se ha visto en el epígrafe anterior, el modelo probit responde a una curva acumulativa normal, mientras que en el modelo logit, la curva acumulativa es la logística. La curva acumulativa logística es típica de los procesos de crecimiento con substrato limitado.

Este modelo se utiliza mucho cuando se tiene una variable dependiente dicotómica, es decir, una variable con dos posibles valores: 1 si se presenta el suceso favorable y 0 si se presenta el suceso desfavorable.

Si se denomina *p* como la probabilidad del suceso favorable, se tiene que

$$p_i = \frac{1}{1 + e^{-(a + b X_i)}}$$

siendo *e*=2.71828 la base de los logaritmos naturales.

Como

$$\mu_i = a + b X_i$$

se puede expresar la fórmula anterior como

$$p_i = \frac{1}{1 + e^{-\mu_i}}$$

Esta ecuación es la conocida como *función de distribución logística.*

Dado que μ_i está comprendida entre menos infinito y más infinito, P_i estará comprendida entre cero y uno y, por tanto, P_i esta relacionado no linealmente con μ_i. Pero la expresión anterior es intrínsecamente lineal, lo que se puede demostrar fácilmente.

Si P_i es la probabilidad de *éxito* y viene dada por la expresión anterior, entonces $(1-P_i)$, es la probabilidad de *fracaso*, siendo

$$1 - p_i = \frac{1}{1 + e^{\mu_i}}$$

Por tanto, se puede escribir

$$\frac{p_i}{1 - p_i} = \frac{1 + e^{\mu_i}}{1 + e^{-\mu_i}} = e^{\mu_i}$$

La expresión $P_i/(1-P_i)$ sencillamente es la razón entre la probabilidad de tener éxito y la probabilidad de no tenerlo. Si $P_i = 0.8$, esto significa que hay una razón de 4 a 1 de tener éxito.

Si ahora se toman logaritmos naturales se obtiene

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right) = \ln\left(e^{\mu_i}\right) = a + bX_i$$

Esto es, L, el logaritmo de la razón de las dos probabilidades es lineal con respecto a X. L se denomina *logit* y, por tanto, recibe el nombre de *modelo logístico*.

Las características del modelo logístico son:

1) Dado que P va de 0 a 1 (a medida que μ_i varía entre $-\infty$ y $+\infty$), el logit L esta entre $+\infty$ y $-\infty$. Es decir, aunque las probabilidades necesariamente se encuentran entre 0 y 1, los logit no tienen estos límites.
2) La interpretación del modelo logístico es la siguiente: b, la pendiente mide el cambio en L por unidad de cambio en X, es decir, nos muestra cómo varía la posibilidad del **ln** en favor de un suceso exitoso a medida que el valor de X cambia. La ordenada en el origen, a corresponde al valor de la probabilidad en **ln** de un suceso exitoso si X=0. Puede ocurrir que esto no tenga significado físico.
3) Dado un valor de X, y estimados a y b, la probabilidad del suceso favorable para ese valor de X se estimaría a partir de

$$p_i = \frac{1}{1 + e^{-(a + bX_i)}}$$

Por lo que el problema sería como estimar a y b. En el caso de tener los datos en valores individuales de X, sin que se puedan agrupar, la estima de los parámetros de regresión por máxima verosimilitud es muy complicada, por lo que se recomienda la utilización de paquetes estadísticos.

Archivo del programa *SAS* (C12-6.SAS).-

```
title 'Modelos logístico';
option ls=75 ps=60;
data logit;
infile 'c12-4.dat';
input dosis n muertos;
logdosis = log(dosis);
proc logistic;
 model muertos/n = logdosis;
run;
```

Archivo de resultados (C12-6.LST).-

```
                          Modelos logístico

                       The LOGISTIC Procedure
Data Set: WORK.LOGIT
Response Variable (Events): MUERTOS
Response Variable (Trials): N
Number of Observations: 6
Link Function: Logit

                          Response Profile
                     Ordered  Binary
                       Value  Outcome      Count
                           1  EVENT           59
                           2  NO EVENT        39

    Model Fitting Information and Testing Global Null Hypothesis BETA=0

                            Intercept
                 Intercept      and
 Criterion         Only      Covariates    Chi-Square for Covariates
 AIC              133.746      117.294          .
 SC               136.331      122.464          .
 -2 LOG L         131.746      113.294       18.452 with 1 DF (p=0.0001)
 Score               .            .          16.628 with 1 DF (p=0.0001)

               Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate       Ratio
INTERCPT 1    38.4382  10.1400    14.3697     0.0002          .          .
LOGDOSIS 1     4.5956   1.2208    14.1706     0.0002     0.570892   99.050

       Association of Predicted Probabilities and Observed Responses

                Concordant = 67.5%          Somers' D = 0.501
                Discordant = 17.4%          Gamma     = 0.590
                Tied       = 15.1%          Tau-a     = 0.242
                (2301 pairs)                c         = 0.750
```

Modelo II de regresión o cuando la *X* se mide con error.-

La utilización de la recta de regresión de *Y* en *X* (o de *X* en *Y*) no esta justificada nada más que cuando una de las dos variables (la dependiente) debe ser expresada en función de la otra variable (la independiente), con objeto de *previsión* o de *estimación* de una en función de la otra. Este es el caso, por ejemplo, de cuando las medidas de pesos o longitudes de un organismo (variable dependiente) se realizan en diferentes fechas sucesivas arbitrariamente elegidas (variable independiente), o cuando se estudia la producción de cierta especie (variable dependiente) en función de la cantidad de alimento ingerido (variable independiente). Como se ve en estos dos ejemplos, la variable independiente realmente no se mide, sino que su valor es elegido por el experimentador y ese es el que se aplica, en el ejemplo de las medidas de peso o de longitud, el experimentador decide, por ejemplo, realizar dichas medidas todas las semanas, por lo que los valores de la variable independientes serán, y se tendrán antes de realizar la experiencia, de 1 semana, 2 semanas, ...

Por contra, puede ocurrir que las dos variables estén medidas con error, sin decidir sobre el valor de una de ellas. Tal es el caso, por ejemplo, de cuando se estudia las variaciones simultáneas de dos características en un mismo organismo, como puede ser la longitud del cuerpo y la profundidad de las mamas en vacas lecheras, o el peso total del cuerpo y el peso del hígado en aves de carne.

En este epígrafe se va ha estudiará el caso, algo más complejo, en el que la *X* también se mide con error, en contraposición de lo estudiado hasta ahora que se basaba en el supuesto que la variable *X* se mide sin incurrir en error.

El modelo que se planteó al comienzo de Capítulo 11 fue

$$\mu_{Y.X_i} = \alpha + \beta X_i$$

Lo que se pretendía al utilizar este modelo es encontrar la relación entre los valores verdaderos de *X* (χ) y de *Y* (ψ), eliminando los errores de observación. Dicha relación puede expresarse

$$\psi = \alpha + \beta\chi$$

donde ψ (*psi*) es el verdadero valor de *Y*, y χ (*ji*) es el verdadero valor de *X*, y α y β son los valores paramétricos de *a* y *b*, respectivamente.

En el caso del Modelo I, se supone que la variable *X* se mide sin error, por lo que $\chi = X$ y $\sigma_x^2 = \sigma_X^2$. Y la variable *Y* se mide con error, es decir, $Y = \psi + \varepsilon$, siendo ε el error de medida de *Y*. La estima de β se realiza dividiendo la covarianza muestral por la varianza muestral de la variable independiente. Como la varianza de la variable independiente, σ_x^2 es igual a σ_χ^2, se espera que la covarianza muestral estime $\sigma_{\chi\psi}$, dado que ε, la porción del error de *Y*, es independiente de χ y, por lo tanto, no contribuirá a la covarianza entre *X* e *Y*. Y, por tanto, la estima de β será insesgada, de manera que

$$\beta = \frac{Cov_{XY}}{S_X^2} = \frac{\sigma_{\chi,\psi}}{\sigma_\chi^2}$$

En el Modelo II, ambas variables, X e Y están sujeta a error, por lo que

$$Y = \psi + \varepsilon$$
$$X = \chi + \delta$$

donde δ es el término de error de la variable X.

Por tanto, en el caso de que se mida la variable independiente con error, la covarianza de X e Y deberá ser un estimador imparcial de $\sigma_{\chi\psi}$, por lo que

$$\Sigma XY = \Sigma(\chi + \delta)(\psi + \varepsilon) = \Sigma\chi\psi + \Sigma\chi\varepsilon + \Sigma\psi\delta + \Sigma\delta\varepsilon$$

Esperándose que todos los sumatorios, del término de la derecha de la anterior expresión, sean igual a cero, menos el primero, puesto que tanto ε como δ se distribuyen con media cero y son independientes entre ellos. Sin embargo, ahora, la varianza muestral de X está estimando

$$\sigma_X^2 = \sigma_\chi^2 + \sigma_\delta^2$$

por lo que se espera que el coeficiente de regresión b calculado sea menor en valor absoluto que la pendiente verdadera de la autentica relación funcional.

Existe un caso especial de aparente medidas con error de la variable X que permite utilizar el Modelo I. Este es el caso en que la medida de X se realiza con error pero está controlada por el experimentador. Esto es como consecuencia de que no hay instrumentos de medición que sean perfectos. Esto es muy corriente en el trabajo experimental. Como se puede estar seguro, por ejemplo, de la dosis de una hormona administrada a las unidades experimentales; se puede cometer algún error en la dosis administrada o en la lectura de la concentración y sobre todo, lo que si puede ocurrir es que la dosis y concentración planteada, cuando interacciona con el organismo, será la dosis y concentración *efectiva* y no necesariamente la *planteada.* En casos como estos, también se puede plantear que

$$X = \chi + \delta$$

donde X es el valor decidido por el experimentador, χ es el valor real y δ es el error. Pero en este caso, X e δ no están correlacionados, pues no existe razón para suponer que la magnitud del valor planeado y su error de aplicación estén correlacionados, por lo que, en casos como estos, se puede utilizar la metodología anterior.

Para los demás casos, existen varias soluciones. Algunas de ellas parten del hecho de que se conocen los valores paramétricos de σ_ε^2 y de σ_δ^2, como esto no acostumbra a ocurrir en la investigación biológica, se verán otras dos soluciones más útiles, tal como es la de la regresión de *mínimos rectángulos* y la del *método de los tres grupos* de Bartlett.

Recta mínimos rectángulos.-

La recta de mínimos rectángulos es la que minimiza el valor absoluto de las sumas de productos de las desviaciones

$$/ X_i - \hat{X}_{(Y_i)} / \; ; / Y_i - \hat{Y}_{(X_i)} /$$

siendo $\hat{X}_{(Y_i)}$ el valor esperado de X sustituyendo en la recta para Y_i. Y $\hat{Y}_{(X_i)}$ el valor esperado de Y sustituyendo en la recta para X_i.

De manera que si se define la recta como

$$Y = a + c\,X$$

siendo Y, X y a lo mismo que en la recta del Modelo I, y siendo c el coeficiente de regresión mínimo rectángulo. La cantidad a minimizar es

$$\Sigma \left[/ X_i - \hat{X}_{(Y_i)} / \, / Y_i - \hat{Y}_{(X_i)} / \right] = \Sigma \left[\left| X_i - \frac{Y_i - a}{c} \right| \left| y_i - a - c\,X_i \right| \right] =$$

$$= \frac{1}{|c|} \Sigma \left(Y_i - a - c\,X_i \right)^2$$

Por desarrollo diferencial se llega al sistema de ecuaciones

$$\begin{cases} \Sigma (Y_i - a - c\,X_i) = 0 \\ \Sigma (Y_i - a - c\,X_i)\,(Y_i - a + c\,X_i) = 0 \end{cases}$$

Como con las rectas mínimo cuadrados, la primera ecuación muestra que

$$\overline{Y} = a + c\,\overline{X}$$

Lo que indica que esta recta, también, para con el punto $(\overline{X}, \overline{Y})$

Despejado a de esta última ecuación y sustituyéndola en la segunda ecuación del sistema, se obtiene

$$c = \pm \frac{S_y}{S_x}$$

el signo es el de la *Cov*.

Como se observa, exceptuando el signo, esta recta solo depende de los parámetros de las distribuciones de las dos variables.

Además, esta recta es intermedia entre las dos rectas mínimo cuadráticas, $b_{Y.X}$, $b_{X.Y}$.

Pruebas de hipótesis e intervalos de confianza con la recta mínimos rectángulos.-

Para dos variables independientes el coeficiente c se puede considerar como estima del parámetro γ, con el signo de la covarianza. Se puede demostrar que en una distribución normal bivariante, la variable

$$t_o = \frac{c^2 - \gamma_o^2}{\frac{2\, c\, \gamma_o \sqrt{1-R^2}}{\sqrt{n-2}}}$$

Esta t_o se distribuye como la t de *Student* con n-2 grados de libertad, siendo γ_o el valor teórico del coeficiente de regresión, c el coeficiente medido y R^2 el coeficiente de determinación

Observar que esta t_o, para el caso de H_o: γ=0, toma el valor de infinito, por lo que para probar esta hipótesis, lo mejor es calcular el intervalo de confianza.

Se puede, por tanto, con esta t_o, probar las hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_o: \gamma \leq \gamma_o$	$H_o: \gamma \geq \gamma_o$	$H_o: \gamma = \gamma_o$
$H_1: \gamma > \gamma_o$	$H_1: \gamma < \gamma_o$	$H_1: \gamma \neq \gamma_o$

Dado el valor paramétrico que se está midiendo, las pruebas de hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_o: \gamma \leq 1$	$H_o: \gamma \geq 1$	$H_o: \gamma = 1$
$H_1: \gamma > 1$	$H_1: \gamma < 1$	$H_1: \gamma \neq 1$

son pruebas de igualdad de desviaciones típicas o de varianzas de dos variables aleatorias eventualmente correlacionadas, equivalentes, limitándonos al caso de las dos colas a

$$H_o: \sigma_X = \sigma_Y$$
$$H_1: \sigma_X \neq \sigma_Y$$

En este caso particular, el valor de t_o es

$$t_o = \frac{c^2 - 1}{\frac{2\, c \sqrt{1-R^2}}{\sqrt{n-2}}}$$

Con respecto a los intervalos de confianza, si designamos como k la cantidad

$$k = \frac{1-R^2}{n-2} t_{(n-2;\alpha/2)}$$

los límites de confianza son

$$LC_{(Y)} \begin{cases} L_i = c\sqrt{1+2k-\sqrt{(1+2k)^2-1}} \\ L^s = c\sqrt{1+2k+\sqrt{(1+2k)^2-1}} \end{cases}$$

Ejemplo.-

Se tiene la longitud del cuerpo (X) y la profundidad de las ubres (Y), medidas en centímetros, de 22 vacas de aptitud lechera.

Se desea saber la función lineal que relaciones la profundidad e las ubres en función de la longitud del cuerpo. Los datos son

vaca	*X*	*Y*
1	168	71
2	169	68
3	150	65
4	148	67
5	154	67
6	145	66
7	165	69
8	163	69
9	148	68
10	161	69
11	151	70

vaca	*X*	*Y*
12	176	74
13	159	70
14	159	73
15	151	69
16	155	71
17	169	74
18	158	70
19	157	71
20	161	73
21	146	71
22	150	65

$$\sum X_i = 3463 \quad \overline{X} = 157.4091 \quad \sum X_i^2 = 546625$$
$$\sum Y_i = 1530 \quad \overline{Y} = 69.5454 \quad \sum Y_i^2 = 106550$$
$$\sum X_i Y_i = 241116$$
$$S_X = \sqrt{\frac{546625 - \frac{3463^2}{22}}{21}} = 8.5002$$
$$S_Y = \sqrt{\frac{106550 - \frac{1530^2}{22}}{21}} = 2.6318$$
$$c = \frac{2.6318}{8.5002} = 0.3096$$
$$a = 69.5454 - 0.3086 \times 157.4091 = 20.8088$$

Como la *SP* tiene signo positivo, la *c* también es positiva. L ecuación que relaciona ambas variables funcionalmente es, entonces

$$\hat{Y} = 20.8088 + 0.3096X$$

Si se quiere probar la significación de esta pendiente, se puede calcular el intervalo de confianza al 95% y ver si abarca el cero

$$R^2 = \frac{\frac{280.091^2}{1517.32}}{145.455} = 0.3555$$
$$t_{(20,\ 0.05/2)} = 2.0860$$
$$k = 2.0869\frac{1-0.3555}{22-2} = 0.06724$$
$$L_i = 0.3096\sqrt{1+2\times0.06724 - \sqrt{(1+2\times0.06724)^2 - 1}} = 0.2396$$
$$L^s = 0.3096\sqrt{1-2\times0.06724 + \sqrt{(1+2\times0.06724)^2 - 1}} = 0.4001$$

Como se ve, este intervalo al 95% de confianza, no abarca el cero, por lo que se puede concluir que la pendiente de esta recta es significativa.

Si se quiere probar la hipótesis de igualdad de varianzas o, lo que es lo mismo, H_o: γ=1, sería

$$H_0 : \gamma = 1$$

$$H_1 : \gamma \neq 1$$

$$R^2 = \frac{\frac{280.091^2}{1517.32}}{145.455} = 0.3555$$

$$t_o = \frac{0.3096^2 - 1}{\frac{2 \times 0.3096 \sqrt{1-0.3555}}{\sqrt{20}}} = -8.134^{***}$$

$$t_{(20,0.05/2)} = -2.0860$$

$$t_{(20,0.01/2)} = -2.8453$$

$$t_{(20,0.001/2)} = -3.85$$

Es altamente significativa y negativa, por lo que se concluye que γ es estadísticamente inferior a uno, o lo que es lo mismo, la varianza de Y es estadísticamente inferior a la varianza de X

Archivo del programa *SAS* (C12-7.SAS).-

```
title 'Regresión mínimo cuadrática y mínimo rectángulo';
option ls=75 ps=60;
data minrect;
infile 'c12-7.dat';
input X Y @@;
sumn + 1;
sumx + x;
sumy+ y;
sumxx + x*x;
sumyy + y*y;
sumxy + x*y;
title 'Regresión mínimo cuadrática de Y en X';
proc reg;
 model Y = X;
run;
title 'Regresión mínimo cuadrática de X en Y';
 model Y = X;
run;
title 'Regresión mínimo rectangulo de Y en X';
data minrect2;
set minrect (firstobs=22);
scx=sumxx-(sumx*sumx)/sumn;
scy=sumyy-(sumy*sumy)/sumn;
sp=sumxy-(sumx*sumy)/sumn;
c=sqrt(scy/scx);if sp<0 then c=c*-1;
a=sumy/sumn-c*sumx/sumn;
r2=sp*sp/scx/scy;
tc1=(c*c-1)/((2*c*sqrt(1-r2))/sqrt(sumn-2));
prtc1=probt(tc1,(sumn-2));
k=tinv(.975,(sumn-2))*(1-r2)/(sumn-2);
li=c*sqrt(1+2*k-sqrt((1+2*k)**2-1));
ls=c*sqrt(1+2*k+sqrt((1+2*k)**2-1));
proc print;
var a c tc1 prtc1 li ls;
run;
```

Archivo de datos (C12-7.DAT).-

```
168 71    176 74
169 68    159 70
150 65    159 73
148 67    151 69
154 67    155 71
145 66    169 74
165 69    158 70
163 69    157 71
148 68    161 73
161 69    146 71
151 70    150 65
```

Archivo de resultados (C12-7.LST).-

```
                Regresión mínimo cuadrática de Y en X
Model: MODEL1
Dependent Variable: Y

                          Analysis of Variance
                          Sum of          Mean
  Source           DF     Squares        Square      F Value       Prob>F
  Model             1    51.70367      51.70367       11.030       0.0034
  Error            20    93.75088       4.68754
  C Total          21   145.45455

      Root MSE       2.16507     R-square       0.3555
      Dep Mean      69.54545     Adj R-sq       0.3232
      C.V.           3.11318

                       Parameter Estimates
                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   INTERCEP   1     40.488362     8.76128371         4.621        0.0002
   X          1      0.184596     0.05558202         3.321        0.0034

                Regresión mínimo cuadrática de X en Y
Model: MODEL2
Dependent Variable: Y

                          Analysis of Variance

                          Sum of          Mean
  Source           DF     Squares        Square      F Value       Prob>F
  Model             1    51.70367      51.70367       11.030       0.0034
  Error            20    93.75088       4.68754
  C Total          21   145.45455

      Root MSE       2.16507     R-square       0.3555
      Dep Mean      69.54545     Adj R-sq       0.3232
      C.V.           3.11318

                       Parameter Estimates

                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   INTERCEP   1     40.488362     8.76128371         4.621        0.0002
   X          1      0.184596     0.05558202         3.321        0.0034
```

Regresión mínimo rectangulo de Y en X						
OBS	A	C	TC1	PRTC1	LI	LS
1	20.8089	0.30962	-8.13336	.000000045103	0.23958	0.40013

Recta de los tres grupos de *Bartlett*.-

El cálculo de la pendiente de esta recta se basa en lo siguiente

Se ordenan los datos con respecto a la variable *X*.

Se dividen los datos en tres grupos, aproximadamente del mismo tamaño, de manera que el grupo de las *X* pequeñas y el grupo de *X* mayores, sean del mismo tamaño, *k*.

Se calculan las medias y sumatorios par las dos variables en los tres grupos y en todos los datos.

El cálculo de la pendiente *b* es la diferencia de la media de *Y* del tercer grupo menos la media del primer grupo dividida por la misma diferencia para la variable *X*.

La ordenada en el origen se calcula como siempre.

Es decir, el valor de la regresión se calcula

$$b = \frac{\overline{Y}_3 - \overline{Y}_1}{\overline{X}_3 - \overline{X}_1}$$

Y el de la ordenada en el origen, como siempre

$$a = \overline{Y} - b\,\overline{X}$$

Para el intervalo de confianza se necesita calcular previamente el término

$$c = \frac{k\,(\overline{X}_3 - \overline{X}_1)^2\,(n-3)}{2\,t^2_{(n-3;\,\alpha/2)}}$$

siendo *k* el tamaño del grupo primero y tercero; y $t_{(n-3;\,\alpha/2)}$ el valor de las tablas.

Siendo, entonces, el intervalo

$$LC_{(b)} = \frac{b\,c - SP' \pm \sqrt{c(\,b^2\,SC'_{(X)} + SC'_{(Y)} - 2b\,SP'\,) - SC'_{(X)}\,SC'_{(Y)} - SP'^2}}{c - SC'_{(X)}}$$

donde

$$SC'_{(X)} = \Sigma X_i^2 - \left[\frac{(\Sigma X_{1i})^2}{k} + \frac{(\Sigma X_{2i})^2}{n-2k} + \frac{(\Sigma X_{3i})^2}{k} \right]$$

es decir, es como una *SC* normal, pero utilizando como término de corrección la suma de los términos de corrección de los tres grupos. Y lo mismo para $SC'_{(Y)}$ y *SP*..

Pongamos como ejemplo el mismo anterior.

Ejemplo.-

Se tiene la longitud del cuerpo (*X*) y la profundidad de las ubres (*Y*), medidas en centímetros, de 22 vacas de aptitud lechera.

Se desea saber la función lineal que relaciones la profundidad e las ubres en función de la longitud del cuerpo.

Los datos, ordenados por la variable *X*, y divididos en tres grupos, son

Grupo inferior		
vaca	*X*	*Y*
6	145	66
21	146	71
4	148	67
9	148	68
22	150	65
3	150	65
15	151	69

Grupo intermedio		
Vaca	*X*	*Y*
11	151	70
5	154	67
16	155	71
19	157	71
18	158	70
13	159	70
14	159	73
10	161	69

Grupo superior		
vaca	*X*	*Y*
20	161	73
8	163	69
7	165	69
1	168	71
2	169	68
17	169	74
12	176	74

$\Sigma X_{1i} = 1038$	$\Sigma X_{2i} = 1254$	$\Sigma X_{3i} = 1177$
$\Sigma Y_{1i} = 472$	$\Sigma Y_{2i} = 560$	$\Sigma Y_{3i} = 498$
$\overline{X}_1 = 148.286$	$\overline{X}_2 = 156.75$	$\overline{X}_3 = 167.286$
$\overline{Y}_1 = 67.4286$	$\overline{Y}_2 = 70.00$	$\overline{Y}_3 = 71.1429$
$\Sigma X_{1i}^2 = 153950$	$\Sigma X_{2i}^2 = 196638$	$\Sigma X_{3i}^2 = 196037$
$\Sigma Y_{1i}^2 = 31721$	$\Sigma Y_{2i}^2 = 39361$	$\Sigma Y_{3i}^2 = 35468$
$\Sigma X_{1i}^2 Y_{1i}^2 = 69835$	$\Sigma X_{2i}^2 Y_{2i}^2 = 87946$	$\Sigma X_{3i}^2 Y_{3i}^2 = 83335$

$$\overline{X} = 157.409$$

$$\overline{Y} = 69.5454$$

$$b = \frac{71.1429 - 67.2857}{167.2857 - 148.2857} = 0.2030$$

$$a = 69.5455 - 0.2030 \times 157.409 = 37.5914$$

Por lo que la recta que relacionan funcionalmente ambas variables es

$$\hat{Y} = 37.5914 + 0.203\ X$$

Como se puede apreciar, este método da valores algo diferentes al anterior.

Si se calculan los límites de confianza para comprobar entre que valores se encuentra el valor paramétrico en esta población, se necesita primero el término

$$c = \frac{7 \times (167.286 - 148.286)^2 (22 - 3)}{2 \times 2.0930^2} = 5480.1241$$

siendo k=7 y $t_{(n-3;\ \alpha/2)} = 2.093$.

Y las sumas de cuadrados primas

$$SC'_{(X)} = 546625 - \left[\frac{1038^2}{7} + \frac{1254^2}{8} + \frac{1171^2}{7}\right] = 248.36$$

$$SC'_{(Y)} = 106550 - \left[\frac{471^2}{7} + \frac{561^2}{8} + \frac{498^2}{7}\right] = 89.161$$

$$SP' = 241116 - \left[\frac{1038 \times 471}{7} + \frac{1254 \times 561}{8} + \frac{1171 \times 498}{7}\right] = 28.3929$$

Siendo, entonces, el intervalo al 95% de confianza

$$LC_{(b)} = \frac{0.203 \times 5480 - 28.3929 \pm \sqrt{\begin{array}{l}5480(0.203^2 \times 248.36 + 89.161 - \\ -2 \times 0.203 \times 28.3929) - \\ -248.36 \times 89.161 - 28.3929^2\end{array}}}{5480 - 248.36}$$

$$L_i = 0.0778$$

$$L^s = 0.3366$$

Como se ve no incluye el cero, luego la pendiente es significativa con un 95% de confianza.

Archivo del programa *SAS* (C12-8.SAS).-

```
title 'Regresión de los tres grupos de Bartlett';
options ls=75 ps=60;
data tresgrup;
infile 'c12-7.dat';
input X  Y @@ ;
n=22;
k=7;
*Los resultados de las regresiones minimocuadráticas pueden;
*consultarse en la salida del anterior problema;
*Si se desea realizar en este programa quítese los asteriscos;
*title 'Regresión mínimo cuadrática de Y en X';
*proc reg;
* model Y = X;
*run;
*title 'Regresión mínimo cuadrática de X en Y';
* model Y = X;
*run;
proc sort;by X y;
run;
data tres3;
set tresgrup (firstobs=16);
sumx3 + x;
sumy3 + y;
sumxx3 + x*x;
sumyy3 + y*y;
sumxy3 + x*y;
xm3=sumx3/k;
ym3=sumy3/k;
run;
data tres2;
set tresgrup (obs=15 firstobs=8);
sumx2 + x;
sumy2 + y;
sumxx2 + x*x;
sumyy2 + y*y;
sumxy2 + x*y;
xm2=sumx2/(n-(2*k));
ym2=sumy2/(n-(2*k));
run;
data tres1;
set tresgrup (obs=7);
sumx1 + X;
sumy1 + Y;
sumxx1 + x*x;
sumyy1 + y*y;
sumxy1 + x*y;
xm1=sumx1/k;
ym1=sumy1/k;
run;
data tres4;
merge tres1 tres2 tres3;
by n;
scx=sumxx1+sumxx2+sumxx3-
     (sumx1*sumx1/k+sumx2*sumx2/(n-(2*k))+sumx3*sumx3/k);
scy=sumyy1+sumyy2+sumyy3-
     (sumy1*sumy1/k+sumy2*sumy2/(n-(2*k))+sumy3*sumy3/k);
sp=sumxy1+sumxy2+sumxy3-
    (sumx1*sumy1/k+sumx2*sumy2/(n-(2*k))+sumx3*sumy3/k);
```

```
c=(k*(xm3-xm1)**2*(n-3))/(2*tinv(.975,(n-3))**2);
b=(ym3-ym1)/(xm3-xm1);
ym=(sumy1+sumy2+sumy3)/22;
xm=(sumx1+sumx2+sumx3)/22;
a=ym-b*xm;
li=(b*c-sp-
   sqrt(c*(b*b*scx+scy-2*b*sp)-scx*scy-sp*sp))/(c-scx);
ls=(b*c-sp+
   sqrt(c*(b*b*scx+scy-2*b*sp)-scx*scy-sp*sp))/(c-scx);
proc print;
title 'Regresión de los tres grupos de Bartlett';
var a b c li ls;
run;
```

Archivo de resultados (C12-8.LST).-

Regresión de los tres grupos de Bartlett

OBS	A	B	C	LI	LS
1	0.4119	0.43750	79.31	.	.
2	12.5730	0.15152	337.37	.	.
3	19.4855	0.14000	774.49	.	.
4	25.7792	0.14286	1518.01	.	.
5	31.5679	0.14607	2453.90	.	.
6	30.5168	0.20370	3613.47	.	.
7	35.9395	0.20301	5480.00	.	.
8	37.5902	0.20301	5480.00	0.077779	0.33666

Como siempre, en casos como este, hay que mirar la última línea.

Bibliografía

Dagnelie, P. 1970. THÉORIE ET MÉTHODES STATISTIQUES. Ed J. Duculot, S.A. Gembloux.

Freund, R.J., and Littel, R.C. 1991. *SAS* SYSTEM FOR REGRESION. SAS Institute Inc., Cary, NC, USA.

*Infante Gil, S. y Zárate De Lara, G.P.*1984. METODOS ESTADISTICOS. Ed. TRILLAS. México.

Lite, TM, y Jackson Hills, F. 1987. METODOS ESTADISTICOS PARA LA INVESTIGACION EN LA AGRICULTURA. Ed TRILLAS. México.

Ostle, B. 1965. ESTADISTICA APLICADA. Ed. Limusa-Wiley. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

Littel, R.C., Freund, R.J. and Spector, P.C. 1991. *SAS* FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. 1990. *SAS* PROCEDURE GUIDE. Cary, NC, USA.

CAPÍTULO 13

Regresión Múltiple

CAPÍTULO 13

Regresión Múltiple

Regresión múltiple.-

La regresión de Y en una sola variable independiente es muchas veces inadecuada. A menudo, puede haber disponibles dos o más X para proporcionar más información acerca de Y, por medio de la regresión múltiple en las X. Entre los usos principales de la regresión múltiple están los siguientes

1 La elaboración de una ecuación en la que varias X permiten una mejor predicción de los valores de Y.

2 Cuando se tiene muchas X se puede buscar el subconjunto de X que da la mejor ecuación lineal. Por ejemplo, en la predicción del tiempo se puede usar hasta 50 variables X; una ecuación con 50 variables X puede resultar difícil de manejar, además de no ser preciso el utilizarlas todas si muchas de estas variables no contribuyen significativamente a un mejor pronóstico, por lo que se puede buscar una ecuación con solo las tres o cuatro variables más predictivas.

3) En otros estudios semejantes al anterior, el objetivo no es el pronóstico, sino que se tiene un número de variables X y se quiere saber cuáles están relacionadas con la Y, y, de ser posible, ordenar las variables en categorías según su importancia en influenciar a Y.

Dos variables independientes.-

Cuando se tenía una sola variable X, los valores de X e Y se podían representar en un plano junto con la recta de regresión. Pero si Y es la variable dependiente de X_1 y de X_2, para su representación se necesita tres dimensiones y, en lugar de línea de regresión, se tiene un espacio de regresión. Ahora se estudiará un solo tipo de espacio como son las *superficies planas*, es decir, la regresión lineal en X_1 y X_2 o *plano de regresión.* Como se verá más adelante, la generalización a cualquier *espacio de regresión* es automática.

El plano de regresión poblacional se expresa

$$\mu_{Y.X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2$$

β_1 mide el cambio medio o esperado de Y cuando X_1 aumenta una unidad, permaneciendo constante X_2. Por esta razón a β_1 se le denomina *coeficiente de regresión parcial.*

Dado una X_1 y una X_2, los valores individuales de Y varían en torno al plano de regresión ajustándose a una distribución normal, con media cero y varianza σ^2, dado este supuesto el modelo es

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

siendo

$$\varepsilon \sim N(0, \sigma^2)$$

Tomada una muestra de n valores de (Y, X_1, X_2) la ecuación de regresión es

$$Y = a + b_1 X_1 + b_2 X_2$$

siendo a, b_1 y b_2 estimas mínimo cuadráticas, es decir, estimas que minimizan el valor de

$$\Sigma_i (Y_i - \hat{Y})^2$$

El valor de a viene dado por la ecuación

$$a = \overline{Y} + b_1 \overline{X}_1 + b_2 \overline{X}_2$$

Sustituyendo a en la ecuación de regresión se tiene

$$Y = \overline{Y} + b_1 (X_1 - \overline{X}_1) + b_2 (X_2 - \overline{X}_2)$$

Las b satisfacen las ecuaciones normales

$$b_1 SC_{(X_1)} + b_2 SP_{(X_1,X_2)} = SP_{(X_1,Y)}$$
$$b_1 SP_{(X_1,X_2)} + b_2 SC_{(X_2)} = SP_{(X_2,Y)}$$

Resolviendo este sistema de ecuaciones se obtienen las soluciones

$$b_1 = \frac{SC_{(X_2)}SP_{(X_1,Y)} - SP_{(X_1,X_2)}SP_{(X_2,Y)}}{SC_{(X_1)}SC_{(X_2)} - SP^2_{(X_1,X_2)}}$$

$$b_2 = \frac{SC_{(X_1)}SP_{(X_2,Y)} - SP_{(X_1,X_2)}SP_{(X_1,Y)}}{SC_{(X_1)}SC_{(X_2)} - SP^2_{(X_1,X_2)}}$$

Ejemplo.-

Con objeto de constatar la fuente de la que obtiene el fósforo la planta de maíz se estudiaron 18 plantaciones en suelos diferentes en los que se determinó químicamente la concentración (partes por millón) de *fósforo inorgánico* (X_1) y de *fósforo orgánico* (X_2) a 20°C. Así mismo se determinó el contenido de *fósforo en el maíz* cultivado (Y)

Suelo	X_1	X_2	Y
1	0.4	53	64
3	0.4	23	60
5	3.1	19	71
7	0.6	34	61
9	4.7	24	54
11	1.7	65	77
13	9.4	44	81
15	10.1	31	93
17	11.6	29	93
2	12.6	58	51
4	10.9	37	76
6	23.1	46	96
8	23.1	50	77
10	21.6	44	93
12	23.1	56	95
14	1.9	36	54
16	26.8	58	168
18	29.9	51	99
Σ	215.0	758	1463
$\overline{X}$	11.94	42.11	81.28

$\Sigma X_1^2 = 4321.02$ $\quad \Sigma X_1 X_2 = 10139.50$ $\quad \Sigma X_1 Y = 20706.20$

$TC = \frac{215^2}{18} =$ $\quad TC = \frac{215 \times 758}{18} =$ $\quad TC = \frac{215 \times 1463}{18} =$

$= 2568.06$ $\quad = 9053.89$ $\quad = 17474.72$

$SC_{(X_1)} = 1752.96$ $\quad SP_{(X_1,X_2)} = 1085.61$ $\quad SP_{(X_1,Y)} = 3231.48$

$$\Sigma X_2^2 = 35076.00 \qquad \Sigma X_2 Y = 63825.56 \qquad \Sigma Y^2 = 131299.00$$

$$TC = \frac{758^2}{18} = 31920.22 \qquad TC = \frac{758 \times 1463}{18} = 61608.56 \qquad TC = \frac{1463^2}{18} = 118909.39$$

$$SC_{(X_2)} = 3155.78 \qquad SP_{(X_2,Y)} = 2216.44 \qquad SC_{(Y)} = 12389.61$$

Sustituyendo en las soluciones del sistema de ecuaciones normales se tiene

$$b_1 = \frac{3155.78 \times 3231.48 - 1085.61 \times 2216.44}{1752.96 \times 3155.78 - 1085.61^2} = 1.7898$$

$$b_2 = \frac{1752.96 \times 2216.44 - 1085.61 \times 3231.48}{1.75296 \times 3255.78 - 1085.61^2} = 0.0866$$

$$a = 81.28 - 1.7898 \times 11.94 - 0.0866 \times 42.11 = 56.26$$

Por tanto, la ecuación de regresión múltiple es

$$Y = 56.26 + 1.7898 X_1 + 0.0866 X_2$$

El significado de esta ecuación es que para cada parte por millón de fósforo inorgánico adicionada en el suelo al comienzo de la época de crecimiento, el fósforo del maíz aumentó en 1.7898 *ppm*, contra 0.0866 *ppm* para cada *ppm* de fósforo orgánico adicionada. Por lo que la conclusión es que el fósforo inorgánico en el suelo fue la principal fuente de fósforo asimilable por la planta.

Prueba de ajuste en la regresión múltiple.-

En el modelo de regresión múltiple, como en la regresión simple, las desviaciones de las *Y* del plano de regresión poblacional tienen media cero y varianza σ^2. Una estima imparcial de σ^2 es

$$S^2_{Y.X_1,X_2} = \frac{\Sigma_i (Y_i - \hat{Y})^2}{n-k} = \frac{SC_{(Error)}}{n-k} = CM_{(Error)}$$

siendo

n el tamaño de muestra
k el número de parámetros estimados.
Los demás términos se verán dentro de un momento.

En el ejemplo anterior n=18, con tres parámetros estimados, α, β_1 y β_2, por lo que n-k=15.

La suma de cuadrados de las desviaciones

$$\Sigma_i(Y_i - \hat{Y})^2$$

se puede deducir de la siguiente manera.

Partiendo de la ecuación de regresión

$$Y = \overline{Y} + b_1(X_1 - \overline{X}_1) + b_2(X_2 - \overline{X}_2)$$

como las medias muestrales de $(X_1 - \overline{X}_1)$ y de $(X_2 - \overline{X}_2)$ son ambas cero, la media muestral de los valores ajustados, $\hat{Y}$, será $\overline{Y}$, por lo que se tiene

$$Y - \overline{Y} = (\hat{Y} - \overline{Y}) + (Y - \hat{Y})$$

Sumando para todos los valores de Y_i y elevando al cuadrado, se tiene

$$\Sigma_i (Y_i - \overline{Y})^2 = \Sigma_i (\hat{Y} - \overline{Y})^2 + \Sigma_i (Y_i - \hat{Y})^2$$

Este resultado dice que la *suma de cuadrados total* de Y se subdivide en dos componentes, la debida a la suma de cuadrados de las desviaciones de los valores ajustados de Y con respecto a su media (*suma de cuadrados de la regresión*) y la debida a las desviaciones de los valores ajustados (*suma de cuadrados del error* o *residuo*).

La suma de cuadrados

$$\Sigma_i(\hat{Y} - \overline{Y})^2$$

se designa como la *suma de cuadrados debida a la regresión.*

$$SC_{(\text{Regresión})} = \Sigma_i (\hat{Y} - \overline{Y})^2 = b_1 SP_{(X_1, Y)} + b_2 SP_{(X_2, Y)}$$

De esta manera, la suma de cuadrados de las desviaciones de la regresión, *error* o *residuo* puede obtenerse restando de la suma de cuadrados total, la suma de cuadrados debida a la regresión.

$$\Sigma_i (Y - \hat{Y})^2 = \Sigma_i (Y - \overline{Y})^2 - \Sigma_i (\hat{Y} - \overline{Y})^2$$

Coeficiente de determinación y coeficiente de alineación o factor de mejoramiento.-

Igual que en el ANOVA, la razón de la suma de cuadrados debida al modelo por la suma de cuadrados total, es el denominado ***coeficiente de determinación*** que se simboliza como r^2 o más comúnmente como R^2

$$R^2 = \frac{\frac{SP^2}{SC_{(X)}}}{SC_{(Y)}}$$

Por tanto, este coeficiente indica la proporción de la suma de cuadrados total de Y que es atribuible al ajuste del modelo, es decir, a la regresión. Es una medida de bondad de ajuste al modelo.

El campo de variación de este coeficiente esta entre el 0 y el +1. No puede ser negativo, independientemente de que lo sea b, pues todas las sumas de cuadrados son positivas y el numerador es una expresión elevada al cuadrado. Y no puede ser mayor de uno puesto que la suma de cuadrados explicada de cualquier variable tiene que ser menor que su suma de cuadrados total, o, en caso extremo, si explica toda la variación de una variable, puede ser tan grande como la suma de cuadrados total, pero no mayor.

Este coeficiente de determinación puede indicar ajustes casi perfectos (valores cercanos al uno) si se ha introducido en el modelo muchas variables aunque estas variables sean superfluas y realmente no mejoren el ajuste; para evitar este problema, se puede ajustar el coeficiente de determinación para el número (m) de coeficientes de regresión estimados, siendo este ajuste

$$R^2_{adj} = 1 - \left(\frac{(1 - r^2)(n-1)}{n-m-1} \right)$$

Este R^2 ajustado tiende a estabilizarse en un cierto valor cuando se introducen variables adecuadas al ajuste. Aunque para un solo coeficiente de regresión, ambos valores son próximos.

Hay que hacer notar que el coeficiente de determinación ajustado puede tener valores negativos.

Puesto que el valor del coeficiente de determinación oscila entre 0 y +1, se le puede restar a 1 con el fin de poder calcular la fracción de la suma de cuadrados explicada y no explicada, es decir, la fracción de la suma de cuadrados total de Y explicada por la variabilidad de X (debida a la regresión) es

$$SC_{(\text{regresión})} = \frac{SP^2}{SC_{(Y)}} = r^2 SC_{(Y)}$$

Y la suma de cuadrados no explicada por X es

$$SC_{(error)} = SC_{(Y)} - \frac{SP^2}{SC_{(Y)}} = (1 - r^2)\, SC_{(Y)}$$

La cantidad $1\text{-}r^2$ se denomina *coeficiente de indeterminación* o de no determinación y expresa la proporción de la varianza de una variable que no ha sido explicada por la otra variable. La raíz cuadrada de este coeficiente de indeterminación

$$\sqrt{1 - r^2}$$

se denomina *coeficiente de alineación* o *factor de mejoramiento* y mide la falta de asociación entre las variables X e Y.

Tanto el coeficiente de *determinación* como de *indeterminación* como el coeficiente de *alineación* son cantidades o proporciones indicativas de lo que denomina su nombre, pero no son estadísticos que permitan hacer inferencias.

Sin embargo, una estima insesgada de la varianza no explicada por la regresión con la que si se puede realizar inferencias es el *CM* del residuo o error, que se simboliza como $S^2_{Y.X}$, es decir, la fracción de la suma de cuadrados total de Y no explicada por la variabilidad de X dividida por sus grados de libertad (n-2). A su raíz cuadrada se le denomina *error típico o desviación típica de Y para X fijo* o bien se le denomina *desviación típica o error típico de Y manteniendo constante X.*

Ejemplo.-

Siguiendo con el mismo ejemplo del fósforo,

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_1 \neq 0 \text{ ó } \beta_2 \neq 0 \text{ ó } \beta_3 \neq 0$$

$$SC_{(regresión)} = 1.7898 \times 3231.48 + 0.0866 \times 2216.44 = 5975.6$$
$$SC_{(error)} = 12389.6 - 5975.6 = 6414.0$$

Se puede, ahora, contrastar las hipótesis

FV	*gl*	*SC*	*CM*	*F*
Regresión	2	5975.6	2987.8	6.99**
Error	15	6414.0	427.6	
Total	17	12389.6		

El valor de F, significativo al 0.01 indica que ambos coeficientes de regresión o uno de ellos es diferente de cero.

El coeficiente de determinación es

$$R^2 = \frac{5975.6}{12389.6} = 0.4823$$

es decir, del 48%.

Y el coeficiente de determinación ajustado es

$$R^2_{adj} = 1-\left(\frac{(1-0.4823)(18-1)}{18-2-1}\right) = 0.4133$$

Se puede hacer la prueba de significación de las *b* individuales, ampliando el anterior análisis, de la siguiente manera.

Si se hubiese ajustado la regresión de *Y* sobre, solamente, X_1, el coeficiente de regresión sería

$$b_{Y.X_1} = \frac{SP_{(Y,X_1)}}{SC_{(X_1)}} = \frac{3231.48}{1752.96} = 1.8434$$

La suma de cuadrados debida a esta regresión es

$$SC_{(\text{Regresión})} = \frac{SP^2_{(Y,X_1)}}{SC_{(X_1)}} = \frac{3231.48^2}{1752.96} = 5957.0$$

Por tanto ha habido una reducción de la suma de cuadrados con respecto a cuando se incluían las dos *X* de, 5975.6-5957.0=18.6, que con *gl*=1 mide la contribución de X_2 a la regresión dado que X_1 ya estaba incluida.

De la misma manera se puede probar la H_o: $\beta_2=0$

La tabla de los resultados de estas pruebas es

FV	*gl*	*SC*	*CM*	*F*
X_1 y X_2	2	5975.6		
X_1 *solo*	1	5957.0		
X_2	1	18.6	18.6	0.04*ns*
X_1 y X_2	2	5975.6		
X_2 *solo*	1	1556.7		
X_1	1	4418.9	4418.9	10.30**
Error	15	6414.0	427.6	
Total	17	12389.6		

Por lo que se concluye que X_2 no influye significativamente en la variación de Y, siendo la mayoría de la variación de Y explicada por la variación de X_1 y por otros factores que no han sido estudiado (residuo).

Pruebas de hipótesis e Intervalos de confianza para los coeficientes de regresión.-

Si se desean las t de *Student* para el cálculo de los límites de confianza no se tiene más que hallar las raíces cuadradas de las F_o anteriores.

O bien, si no se ha hecho la prueba de ajuste por preferir el uso de los límites de confianza o bien porque se prefiere hacer una prueba de hipótesis $\beta=\beta_o$, siendo β_o un número diferente a cero, se pueden hallar directamente las t de esta manera.

Los errores típicos de b_1 y de b_2 son

$$S_{b_1} = \sqrt{S^2_{(Y.X_1,X_2)}\left[\frac{SC_{(X_2)}}{SC_{(X_1)}\ SC_{(X_2)} - SP^2_{(X_1,X_2)}}\right]}$$

$$S_{b_2} = \sqrt{S^2_{(Y.X_1,X_2)}\left[\frac{SC_{(X_1)}}{SC_{(X_1)}SC_{(X_2)} - SP^2_{(X_1,X_2)}}\right]}$$

Se puede probar que la cantidad

$$\frac{b_1+\beta_1}{S_{b_1}} \sim t_{(n-k;\alpha)}$$

Es decir, se distribuye como t de *Student* con $gl=(n-k)$.

Por lo tanto se pueden probar las hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_0:\beta_1 \le \beta_{o_1}$	$H_0:\beta_1 \ge \beta_{o_1}$	$H_0:\beta_1 = \beta_{o_1}$
$H_1:\beta_1 > \beta_{o_1}$	$H_1:\beta_1 < \beta_{o_1}$	$H_1:\beta_1 \ne \beta_{o_1}$

$$t_o = \frac{b_1-\beta_{o_1}}{S_{b_1}}$$

pudiendo ser β_{o1} cualquier número, incluido el cero.

Esta t_o se distribuye como la t de *Student* por lo que se puede contrastar con la $t_{(n-k;\,\alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-k;\,\alpha)}$ para las hipótesis de una cola.

Lo mismo se haría con b_2.

Y los límites de confianza para cualquier β_i se encuentran en la forma acostumbrada

$$LC_{(\beta_1)} = b_1 \pm S_{b_1} t_{(n-k;\alpha/2)}$$
$$LC_{(\beta_2)} = b_2 \pm S_{b_2} t_{(n-k;\alpha/2)}$$

Ejemplo.-

Siguiendo con el ejemplo anterior, puede ocurrir que se tenga una hipótesis previa a la realización del experimento del valor esperado para b_1, por ejemplo. se espera que $\beta_1=1$, en este caso la prueba sería

$$H_o : \beta_1 = 1$$
$$H_1 : \beta_1 \neq 1$$
$$S_{b_1} = \sqrt{427.6 \left[\frac{3155.78}{1752.96 \times 3155.78 - 1085.61^2}\right]} = 0.557$$
$$t_o = \frac{1.7898 - 1}{0.557} = 1.418 ns$$
$$t_{(15;\,0.05/2)} = 2.1315$$

Se confirma la hipótesis nula, por lo que se puede concluir que el valor de b_1 no es estadísticamente diferente de uno.

Si la hipótesis que se quiere probar es la de significación de la pendiente, es decir, la de $\beta_1=0$, la prueba sería

$$H_o : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$
$$S_{b_1} = \sqrt{427.6 \left[\frac{3155.78}{1752.96 \times 3155.78 - 1085.61^2}\right]} = 0.557$$
$$t_o = \frac{1.7898}{0.557} = 3.21^{**}$$
$$t_{(15;\,0.05/2)} = 2.1315$$
$$t_{(15;\,0.01/2)} = 2.9467$$

Se rechaza la hipótesis nula, por lo que se puede concluir que la pendiente de este coeficiente es significativa, o lo que es lo mismo, que los valores de X_1 influyen en los valores de Y.

Nótese que el valor de t_o para contrastar la hipótesis de β=0 es la raíz cuadrada del valor de F_o de la prueba de ajuste.

Y para b_2

$$H_o : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$
$$S_{b_2} = \sqrt{427.6 \left[\frac{1752.96}{1752.96 \times 3155.78 - 1085.61^2} \right]} = 0.416$$
$$t_o = \frac{0.0866}{0.416} = 0.21 ns$$
$$t_{(15;\,0.05/2)} = 2.1315$$

Se confirma la hipótesis nula, por lo que se puede concluir que el valor de b_2 no es estadísticamente diferente de cero, o lo que es lo mismo, que los valores de X_2 no influyen significativamente en la variación de Y.

Nótese que el valor de t_o para contrastar la hipótesis de β=0 es la raíz cuadrada del valor de F_o de la prueba de ajuste

Los intervalos de confianza, al 95%, son, para la b_1

$$LC_{(\beta_1)} = 1.7898 \pm 0.557 \times 2.1315 = 1.7898 \pm 1.1872$$
$$L_i = 0.6025$$
$$L^S = 2.9770$$

Y para la b_2

$$LC_{(\beta_2)} = 0.0866 \pm 0.416 \times 2.1315 = 0.0866 \pm 0.8867$$
$$L_i = -0.8001$$
$$L^S = 0.9733$$

Pruebas de hipótesis e intervalos de confianza para la ordenada en el origen.-

Estos son casos especiales de valores estimados (ver más adelante) de $\hat{Y}$ para X_1=0, X_2 = 0, si se quiere la ordenada en el origen. Y para $X_1 = \overline{X}_1$, $X_2 = \overline{X}_2$ si se quiere la media de la variable dependiente.

El *error típico de* α es

$$S_{\hat{Y}.0,0} = \sqrt{S^2_{(Y.X_1,X_2)}\left(\frac{1}{n}+\frac{(-\overline{X}_1)^2 SC_{(X_2)}+(-\overline{X}_2)^2 SC_{(X_1)}}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right) - \left(\frac{2\,SP_{(X_1X_2)}(-\overline{X}_1)(-\overline{X}_2)}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)}$$

Se pueden realizar, también, pruebas de hipótesis para contrastar la hipótesis sobre α por medio de la siguiente prueba *t*.

Cola derecha	Cola izquierda	Dos colas
$H_0 : \alpha \leq \alpha_0$	$H_0 : \alpha \geq \alpha_0$	$H_0 : \alpha = \alpha_0$
$H_1 : \alpha > \alpha_0$	$H_1 : \alpha < \alpha_0$	$H_1 : \alpha \neq \alpha_0$

$$t_o = \frac{a-\alpha_0}{S_{\hat{Y}.0,0}}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n\text{-}k;\,\alpha/2)}$ para las hipótesis de dos cola y con $t_{(n\text{-}k;\,\alpha)}$ para las hipótesis de una cola.

Ejemplo.-

Estímese el error típico de la ordenada en el origen del mismo ejemplo y compruébese si esta ordenada en el origen es estadísticamente igual a cero

El error típico es

$$S_{(\hat{Y}.0,0)} = \sqrt{427.6\left(\frac{1}{18}+\frac{(-11.94)^2\times 3155.78+(-42.11)^2\times 1752.96}{1752.96\times 3155.78-1085.61^2}\right) - \left(\frac{2\times(-11.94)\times(-42.11)\times 1085.61}{1752.96\times 3155.78-1085.61^2}\right)} =$$

$$= 16.3106$$

Y la prueba para α=0 es

$$H_0 : \alpha = \alpha_0$$

$$H_1 : \alpha \neq \alpha_0$$

$$t_o = \frac{56.26}{16.3106} = 3.4493\,**$$

$$t_{(15;\,0.05/2)} = 2.1315$$

$$t_{(15;\,0.01/2)} = 2.9467$$

Luego se concluye que este plano de regresión no pasa por el origen de coordenadas.

Todos estos resultados se obtienen con el siguiente programa

Archivo de programa SAS (C13-1.SAS).-

```
title 'Regresión Múltiple';
options ls=75 ps=60;
data regresi;
infile 'c13-1.dat';
input suelo X1 X2 Y ;
proc reg;
  model Y = X1 X2 ;
run;
title 'Prueba para b1=1 haciendo X2 igual a cero';
   test X1=1;
run;
title 'Prueba para b1=0 haciendo X2 igual a cero';
   test X1=0;
run;
title 'Prueba para b2=0 haciendo X1 igual a cero';
   test X2=0;
run;
```

Archivo de datos (C13-1.DAT).-

```
 1   0.4     53     64
 2  12.6     58     51
 3   0.4     23     60
 4  10.9     37     76
 5   3.1     19     71
 6  23.1     46     96
 7   0.6     34     61
 8  23.1     50     77
 9   4.7     24     54
10  21.6     44     93
11   1.7     65     77
12  23.1     56     95
13   9.4     44     81
14   1.9     36     54
15  10.1     31     93
16  26.8     58    168
17  11.6     29     93
18  29.9     51     99
19   6.0     30      .
20  35.0     70      .
21  11.9     42      .
```

Las tres últimas líneas se utilizarán en los siguientes ejemplos.

Archivo de resultados (C13-1.LST).-

```
                          Regresión Múltiple

Model: MODEL1
Dependent Variable: Y

                        Analysis of Variance

                             Sum of          Mean
Source                DF    Squares        Square      F Value       Prob>F
Model                  2  5975.66853   2987.83427        6.988       0.0072
Error                 15  6413.94258    427.59617
C Total               17 12389.61111

    Root MSE       20.67840     R-square       0.4823
    Dep Mean       81.27778     Adj R-sq       0.4133
    C.V.           25.44164

                        Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|
INTERCEP   1       56.251024   16.31073734         3.449        0.0036
X1         1        1.789774    0.55674341         3.215        0.0058
X2         1        0.086649    0.41494299         0.209        0.8374

Dependent Variable: Y
Test:          Numerator:    860.4575  DF:    1   F value:    2.0123
               Denominator:  427.5962  DF:   15   Prob>F:     0.1765

Dependent Variable: Y
Test:          Numerator:   4418.9601  DF:    1   F value:   10.3344
               Denominator:  427.5962  DF:   15   Prob>F:     0.0058

Dependent Variable: Y
Test:          Numerator:     18.6460  DF:    1   F value:    0.0436
               Denominator:  427.5962  DF:   15   Prob>F:     0.8374
```

Las salidas de este programa (**C13-1.LST**) son:

.Análisis de varianza que prueba la significación del modelo completo, explicado en la página 484 y desarrollado en la página 487.

.La desviación típica del error (**Root MSE**), que no es sino la raíz cuadrada del cuadrado medio del error.

.La media de la variable dependiente (**Dep Mean**).

.El coeficiente de variación (**C.V.**) que es la fracción de las dos cantidades anteriores puesta en porcentajes; nos da la magnitud relativa de la desviación típica del error con respecto a la media, por lo que cuanto menor sea el valor del coeficiente de variación mejor será el ajuste del modelo.

.El coeficiente de determinación (**R-Square**) y el coeficiente de determinación ajustado (**Adj R-Sq**) explicados en la página 486 y desarrollados en las páginas 488.

. Las estimas de los tres parámetros de este modelo (**Parameter Estimate**) que son, α (**INTERCEP**), β_1 (**X1**) y β_2 (**X2**), explicados a partir de la página 482.

.El error típico de estos parámetros (**Standard Error**); la prueba ***t*** de *Student* para el valor de estos parámetros igual a cero (**T for H0: Parameter=0**) y el nivel de

significación de esta prueba (**Prob > |T|**), explicados y desarrollados entre la página 489 y la página 492.

.Las tres últimas pruebas especificadas en el programa (**test**) son pruebas ***F*** explicadas y desarrolladas en la página 488. Como se ve, las dos últimas pruebas son redundantes con las ***t*** de *Student* del apartado anterior, puesto que esta *F* no son sino el cuadrado de las *t*, y su significación es exactamente la misma.

Pruebas de hipótesis e intervalos de confianza para una estimación por interpolación.-

El *error típico de un valor estimado* $\hat{Y}$, para unos valores dados de X_{1o} y X_{2o}, dentro del rango de las *X* muestreadas en el experimento, es

$$S_{\hat{Y}.X_{1o},X_{2o}} = \sqrt{S^2_{(Y.X_1,X_2)}\left(\frac{1}{n}+\frac{(X_{1o}-\overline{X}_1)^2 SC_{(X_2)}+(X_{2o}-\overline{X}_2)^2 SC_{(X_1)}}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)-\left(\frac{2\ SP_{(X_1X_2)}(X_{1o}-\overline{X}_1)(X_{2o}-\overline{X}_2)}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)}$$

Como se ve, la varianza del error se ve incrementada en un factor que es proporcional a la distancia que exista entre las X_o y las respectivas media $\overline{X}$, por lo que cuanto mayor sea la distancia entre X_o y su media mayor será el error típico de la estima de $\hat{Y}$ y mayor, por tanto, el intervalo de confianza, que sería

$$LC_{(\hat{Y}.X_{1o},X_{2o})} = \hat{Y} \pm S_{(\hat{y}.X_{1o}X_{2o})}\, t_{(n-k,\,\alpha/2)}$$

Si se estiman los límites de confianza de $\hat{Y}$ para todos los puntos X_o del experimento, y se representan estos puntos en una gráfica tridimensional donde se tiene el plano de regresión, se obtendrá dos superficies curvas cóncavas a ambos lados del plano. Es decir, a medida que nos alejamos de la media, menos fiable son las estimas de *Y* como consecuencia de la incertidumbre de las verdaderas pendientes de β_1 y β_2.

Se pueden realizar, también, pruebas de hipótesis para contrastar la hipótesis nula de que $\hat{Y}$ es una estima de $\mu_{Y.Xo}$ por medio de la siguiente prueba *t*.

Cola derecha	Cola izquierda	Dos colas
$H_o: \hat{Y} \le \mu_{Y.X_1X_2}$	$H_o: \hat{Y} \ge \mu_{Y.X_1X_2}$	$H_o: \hat{Y} = \mu_{Y.X_1X_2}$
$H_1: \hat{Y} > \mu_{Y.X_1X_2}$	$H_1: \hat{Y} < \mu_{Y.X_1X_2}$	$H_1: \hat{Y} \ne \mu_{Y.X_1X_2}$

$$t_o = \frac{\hat{Y}-\mu_{Y.X_1X_2}}{S_{\hat{Y}.X_{1o}X_{2o}}}$$

Esta t_o se distribuye como la t de *Student* por lo que se puede contrastar con la $t_{(n-k;\,\alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-k;\,\alpha)}$ para las hipótesis de una cola.

Ejemplo.-

Siguiendo con el ejemplo anterior, se tiene que el intervalo de confianza al 95% del valor estimado $\hat{Y}$ para X_{1o}=0.04 y X_{2o}=54 (la primera observación o suelo) es

$$\hat{Y} = 56.26+1.7898\times 0.4+0.0866x53 = 61.56$$

$$S_{(\hat{Y}.X_{1o}X_{2o})} = \sqrt{427.6\left(\frac{1}{18}+\frac{(0.4-11.94)^2\times 3155.78+(53-42.11)^2\times 1752.96}{1752.96\times 3155.78-1085.61^2}\right)-\left(\frac{2\times(0.4-11.94)\times(53-42.11)\times 1085.61}{1752.96\times 3155.78-1085.61^2}\right)} =$$

$$= 10.595$$

$$LC_{(\hat{Y}.0.04,54)} = 61.56 \pm 10.595\times 2.1315 = 61.56 \pm 22.5832$$

$$L_i = 38.9767$$

$$L^s = 84.1432$$

También se puede realizar la estima de cual sería el valor de $\hat{Y}$ para unos valores de X no medidos, como pueden ser, por ejemplo, para X_1=6.0 y X_2=30; en el ejemplo *SAS* se verá este resultado.

Pruebas de hipótesis e intervalos de confianza de la predicción, por extrapolación, de un valor *Y*.-

Entre los usos de la regresión está la predicción de valores de Y para valores de X fuera del campo de variación medido. Estos pueden ser valores futuros o bien pueden ser valores de X que son posibles observar pero es imposible o poco práctico medir el correspondiente valor de Y.

Lo que se ha hecho en el epígrafes anterior era estimar un valor de Y, esto es, $\hat{Y}$, entre dos valores de X observados, es decir, se ha estimado $\hat{Y}$ por *interpolación*. Mientras que ahora de trata de predecir un valor futuro cuya X aún no existe o, si existe, es inasequible como consecuencia del valor de Y.

Es importante que no se confundan los dos tipos de predicción. El resultado de esta predicción esta sujeto al supuesto de que el nuevo miembro proviene de la misma población como los datos originales. A no ser que la predicción satisfagan esta condición, habrá de considerar el error típico como aproximado. Será muy bajo si el paso del tiempo o de ambiente a cambiado el valor de β.

Al predecir un valor o al estimar una media $\mu_{Y.X}$ para un X fuera del intervalo observado, esto es, al *extrapolar*, se supone que la relación se mantiene lineal y este supuesto puede no ser acertado, especialmente si se usa la recta como una aproximación y se extrapolan puntos muy alejados del rango muestreado. Para extrapolar con cierta garantía hay que revisar la recta de regresión a medida que se acumulan experiencias.

La predicción se hace igual que la estima de $\hat{Y}$, es decir, sustituyendo en la ecuación de regresión pero el error típico de la predicción es, ahora

$$S_{\hat{Y}.X_{1i},X_{2i}} = \sqrt{S^2_{(Y.X_1,X_2)}\left(1+\frac{1}{n}+\frac{(X_{1o}-\overline{X}_1)^2 SC_{(X_2)}+(X_{2o}-\overline{X}_2)^2 SC_{(X_1)}}{SC_{(X_1)}\ SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)-\left(\frac{2SP_{(X_1X_2)}(X_{1o}-\overline{X}_1)(X_{2o}-\overline{X}_2)}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)}$$

Por lo que se puede realizar pruebas de hipótesis para contrastar la hipótesis nula de que $\hat{Y}$ es una estima de $\mu_{Y.X1,X2}$ por medio de la prueba *t* siguiente.

Cola derecha	Cola izquierda	Dos colas
$H_o: \hat{Y} \le \mu_{Y.X_1X_2}$	$H_o: \hat{Y} \ge \mu_{Y.X_1X_2}$	$H_o: \hat{Y} = \mu_{Y.X_1X_2}$
$H_1: \hat{Y} > \mu_{Y.X_1X_2}$	$H_1: \hat{Y} < \mu_{Y.X_1X_2}$	$H_1: \hat{Y} \ne \mu_{Y.X_1X_2}$

$$t_o = \frac{\hat{Y}-\mu_{Y.X_1X_2}}{S_{\hat{Y}.X_{1i}X_{2i}}}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n-k;\ \alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-k;\ \alpha)}$ para las hipótesis de una cola.

El intervalo de confianza sería

$$LC_{(\hat{Y}.X_{1i},X_{2i})} = \hat{Y} \pm S_{(\hat{y}.X_{1i}X_{2i})}\, t_{(n-k,\ \alpha/2)}$$

Ejemplo.-

Siguiendo con el ejemplo anterior, supóngase que se quiere predecir, por extrapolación, el valor de *Y* del primer suelo. Se tiene que el intervalo de confianza al 95% del valor estimado $\hat{Y}$ para X_{1o}=0.04 y X_{2o}=54 (la primera observación o suelo) es

$$\hat{Y} = 56.26+1.7898\times 0.4+0.0866\times 53 = 61.56$$

$$S_{(\hat{Y}.X_{1i}X_{2i})} = \sqrt{427.6\left(1+\frac{1}{18}+\frac{(0.4-11.94)^2\times 3155.78+(53-42.11)^2\times 1752.96}{1752.96x3155.78-1085.61^2}\right)-\left(\frac{2\times(0.4-11.94)\times(53-42.11)\times 1085.61}{1752.96\times 3155.78-1085.61^2}\right)} =$$

$$= 23.235$$

$$LC_{(\hat{Y}.0.04,54)} = 61.56 \pm 23.235\times 2.1315 = 61.56 \pm 49.5249$$

$$L_i = 12.0351$$

$$L^s = 111.0849$$

Como se ha dicho anteriormente, esta es una prueba para la predicción de valores que están fuera del rango de variación de las X por lo que un ejemplo sería el predecir el valor de Y para los valores de X_1=30 y X_2=70; el resultado se verá con el ejemplo *SAS*.

Pruebas de hipótesis e intervalos de confianza para la media de la variable dependiente.-

Un caso especial de valores estimados es el de la estima de $\overline{Y}$. Dado que el plano de regresión pasa por el punto cuyas coordenadas son las medias, no hay más que estimar Y por interpolación para los lo valores medios de las X.

Recuérdese que el *error típico de la media muestral observada,* $\overline{Y}$, se calcula mediante la expresión

$$S_{(\overline{Y})} = \sqrt{\frac{S^2_{(Y)}}{n}}$$

$$gl = n-1$$

Pero ahora se puede explicar parte de la variación de Y en función de la variación de las X, quedando como varianza no explicada $S^2_{Y.X1,X2}$. Se puede, por tanto, utilizar el error típico

$$S_{(\overline{Y})} = S_{\hat{Y}.\overline{X}_{1o},\overline{X}_{2o}} = \sqrt{S^2_{(Y.X_1,X_2)}\left(\frac{1}{n}+\frac{(X_{1o}-\overline{X}_1)^2 SC_{(X_2)}+(X_{2o}-\overline{X}_2)^2 SC_{(X_1)}}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)-\left(\frac{2\ SP_{(X_1X_2)}(X_{1o}-\overline{X}_1)(X_{2o}-\overline{X}_2)}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)} =$$

$$= \sqrt{\frac{S^2_{(Y.X_1,X_2)}}{n}}$$

$$gl = n - n_{Xs}$$

para el calculo del intervalo de confianza de $\overline{Y}$, de la siguiente manera

$$LC_{(\overline{Y})} = \overline{Y} \pm S_{(\overline{Y})} t_{(n-2,\,\alpha/2)}$$

Ejemplo.-

Siguiendo con los mismos datos de niveles de fósforo, los límites de confianza al 95% de $\overline{Y}$, son

$$\hat{Y} = 56.26 + 1.7898 \times 11.9 + 0.0866 \times 42.1 = 81.20$$

$$S_{(\overline{Y})} = \sqrt{\frac{427.6}{18}} = 4.874$$

$$LC_{(\overline{Y})} = 81.2 \pm 4.874 \times 2.1199 = 81.2 \pm 10.33$$

$$L_i = 70.86$$

$$L^s = 91.53$$

Este intervalo es notablemente mas estrecho que el intervalo de confianza para la misma media de la variable *Y*, sin considerar la influencia de las variables *X*, a pesar de que aquí se tiene dos grados de libertad menos. Recuérdese que este intervalo hubiera sido

$$S_{(\overline{Y})} = \sqrt{\frac{728.80}{18}} = 40.49$$

$$LC_{(\overline{Y})} = 81.28 \pm 40.49 \times 2.1098 = 81.28 \pm 85.43$$

$$L_i = 4.15$$

$$L^s = 116.71$$

Se pueden realizar también pruebas de hipótesis para contrastar la hipótesis nula de que $\overline{Y}$ es una estima de la media poblacional μ_o, por medio de la siguiente prueba *t*.

Cola derecha	Cola izquierda	Dos colas
$H_o : \mu_Y \leq \mu_o$	$H_o : \mu_Y \geq \mu_o$	$H_o : \mu_Y = \mu_o$
$H_1 : \mu_Y > \mu_o$	$H_1 : \mu_Y < \mu_o$	$H_1 : \mu_Y \neq \mu_o$

$$t_o = \frac{\overline{Y} - \mu_o}{S_{(\overline{Y})}}$$

Esta t_o se distribuye como la *t* de *Student* por lo que se puede contrastar con la $t_{(n-3;\, \alpha/2)}$ para las hipótesis de dos cola y con $t_{(n-3;\, \alpha)}$ para las hipótesis de una cola.

Archivo de programa SAS (C13-2.SAS).-

```
title 'Regresión Múltiple: Estimas y predicciones';
options ls=75 ps=60;
data regresi;
infile 'c13-1.dat';
input suelo X1 X2 Y ;
proc reg;
  model Y = X1 X2 / clm cli;
run;
```

La estima de los valores extrapolados se realizan con la opción **CLM** y la estima de los valores extrapolados se realizan con la opción **CLI**.

Para obtener la estima o la predicción de un valor de *Y* diferente a las *X* medidas no hay más que poner los valores de estas *X* en el archivo de datos, poniendo un punto en el sitio que le correspondería al valor de *Y* (ver C13-1.DAT).

Archivo de resultados (C13-2.LST).-

Regresión Múltiple: Estimas y predicciones

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual
1	64.0000	61.5593	10.597	38.9732	84.1455	12.0342	111.1	2.4407
2	51.0000	83.8278	8.071	66.6257	101.0	36.5149	131.1	-32.8278
3	60.0000	58.9599	8.994	39.7887	78.1311	10.8960	107.0	1.0401
4	76.0000	78.9656	5.240	67.7977	90.1334	33.4978	124.4	-2.9656
5	71.0000	63.4457	9.817	42.5211	84.3702	14.6559	112.2	7.5543
6	96.0000	101.6	7.462	85.6758	117.5	54.7238	148.4	-5.5807
7	61.0000	60.2710	7.440	44.4134	76.1285	13.4301	107.1	0.7290
8	77.0000	101.9	7.367	86.2243	117.6	55.1385	148.7	-24.9273
9	54.0000	66.7425	8.278	49.0993	84.3858	19.2674	114.2	-12.7425
10	93.0000	98.7227	7.027	83.7451	113.7	52.1724	145.3	-5.7227
11	77.0000	64.9258	14.018	35.0478	94.8038	11.6783	118.2	12.0742
12	95.0000	102.4	7.906	85.5965	119.3	55.2609	149.6	-7.4472

13	81.0000	76.8875	5.235	65.7301	88.0448	31.4222	122.4	4.1125
14	54.0000	62.7710	6.955	47.9474	77.5945	16.2700	109.3	-8.7710
15	93.0000	77.0139	6.457	63.2506	90.7771	30.8400	123.2	15.9861
16	168.0	109.2	9.235	89.5581	128.9	60.9717	157.5	58.7574
17	93.0000	79.5252	7.241	64.0922	94.9582	32.8264	126.2	13.4748
18	99.0000	114.2	10.161	92.5258	135.8	65.0753	163.3	-15.1844
19	.	69.5891	6.679	55.3531	83.8252	23.2721	115.9	.
20	.	125.0	13.613	95.9432	154.0	72.1903	177.7	.
21	.	81.1886	4.874	70.7997	91.5775	35.9058	126.5	.

Sum of Residuals	0
Sum of Squared Residuals	6413.9426
Predicted Resid SS (Press)	9776.5478

Las salidas de este programa (**C13-2.LST**) son:

Los tres primeros bloques son la salida por defecto y ya se vio en el programa anterior, por lo que se han eliminados de la esta tabla.

El siguiente bloque de salidas lo constituye siete columnas que son:

Valor de la variable dependiente (**Dep Var Y**) para las 18 unidades experimentales medidas. En las observaciones 19, 20 y 21 no hay valor de la variable dependiente pues son valores que se han introducido en el archivo de datos con objeto de que nos estime y prevea, respectivamente, el valor de *Y*.

Valores estimados de *Y* (**Predict Value**) por la regresión.

Error típico de esta estima (**Std Err Predict**).

Límite inferior (**Lower95% Mean**) y superior (**Upper95% Mean**) con el nivel de significación del 0.05 (95% de confianza) de la estima por interpolación

Y límite inferior (**Lower95% Predict**) y superior (**Upper95% Predict**) con el nivel de significación del 0.05 (95% de confianza) de la estima por extrapolación.

Obsérvese que la fila correspondiente a la observación 19 es para la estima del valor de *Y* si los valores de *X* son 6.0 y 30, respectivamente. Mientras que la fila de la observación 20 es para la predicción del valor de *Y* para los valores de *X* de 35.0 y 70, respectivamente. Y la observación 21 es para la estima de la *Y* media.

Análisis de los residuo ($e_{Y.X}$).-

Como se ha visto anteriormente, los valores determinados por la ecuación de regresión, son estimaciones de parámetros poblacionales, es decir, de $\mu_{Y.X} = a+b_1X_{1i}+b_2X_{2i} = \hat{Y}$. Las diferencias entre éstos valores estimados y los valores observados son estimaciones de la variación de *Y* no explicada por la variación de X_1 y de X_2, esto es lo que se ha denominado *residuo*.

Los residuos ($e_{Y.X}$) pueden ser particularmente útiles cuando se representan respecto de *X*. Si tienden a ser del mismo signo en ambos extremos de la gráfica y de signos opuestos en el medio, entonces queda comprobado que la respuesta no es lineal. Si sus magnitudes cambian de manera regular, por ejemplo, aumentando con *X*, entonces hay evidencia de heterogeneidad de la varianza. Los valores extremos o alejados pueden detectarse de la manera que se describe más adelante.

Una dificultad con los residuos es que no todos se estiman con la misma precisión. Sin embargo, se puede estimar errores típicos de los residuos y dividiendo cada residuo por su error típico se obtienen residuos típicos o residuos *Student*. Estos

residuos *Student* se ajustan a la distribución *t* con *n*-2 grados de libertad, por lo que podría utilizarse como prueba de hipótesis. Sin embargo, esto no es posible porque la elección del residuo no es aleatorio, como es el supuesto de la prueba *t*, pero si se puede utilizar para detectar residuos excesivamente grandes (desviaciones sospechosamente grandes), pues residuos stundetizados mayores de 2.5 son relativamente raros y habría que investigar esta desviación inusualmente grande.

El error típico del residuo es,

$$S_{e_{X_1,X_2}} = \sqrt{S^2_{(Y.X_1,X_2)}\left(1-\frac{1}{n}+\frac{(X_{1o}-\overline{X}_1)^2 SC_{(X_2)}+(X_{2o}-\overline{X}_2)^2 SC_{(X_1)}}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)-\left(\frac{2SP_{(X_1X_2)}(X_{1o}-\overline{X}_1)(X_{2o}-\overline{X}_2)}{SC_{(X_1)}SC_{(X_2)}-SP^2_{(X_1,X_2)}}\right)}$$

Archivo de programa SAS (C13-3.SAS).-

```
Title 'Regresión Múltiple: Residuos';
Options ls=75 ps=40;
Data regresi;
Infile 'c13-1.dat';
Input suelo X1 X2 Y @@;
Proc reg;
  Model Y = X1 X2 / R;
  Plot residual.*X1='1';
  Plot residual.*X2='2';
  Plot residual.*obs.='*';
run;
```

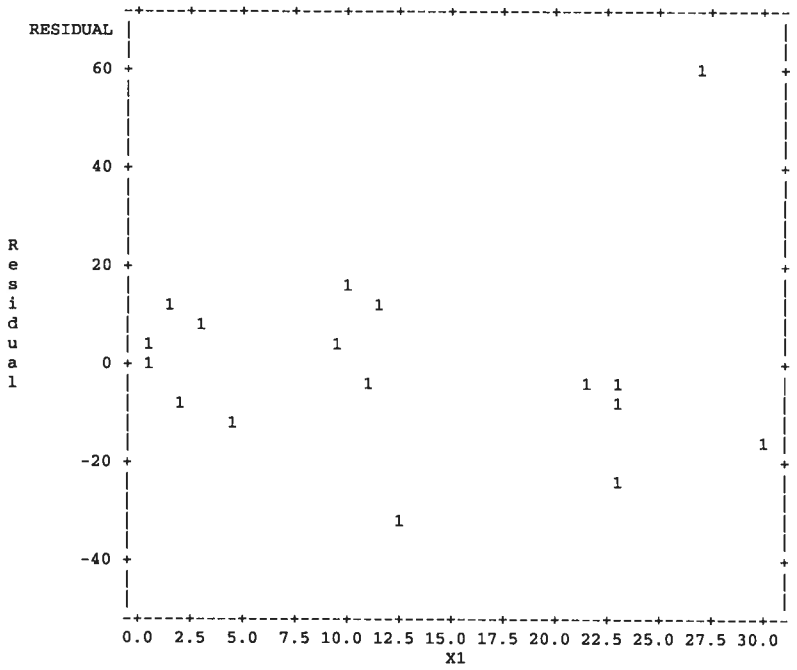
La estima de los residuos realiza con la opción **R**. Se han añadido tres estamentos **PLOT** para que nos representen, respectivamente, los residuos de cada valor de las dos *X* y de cada observación para las dos *X* en conjunto.

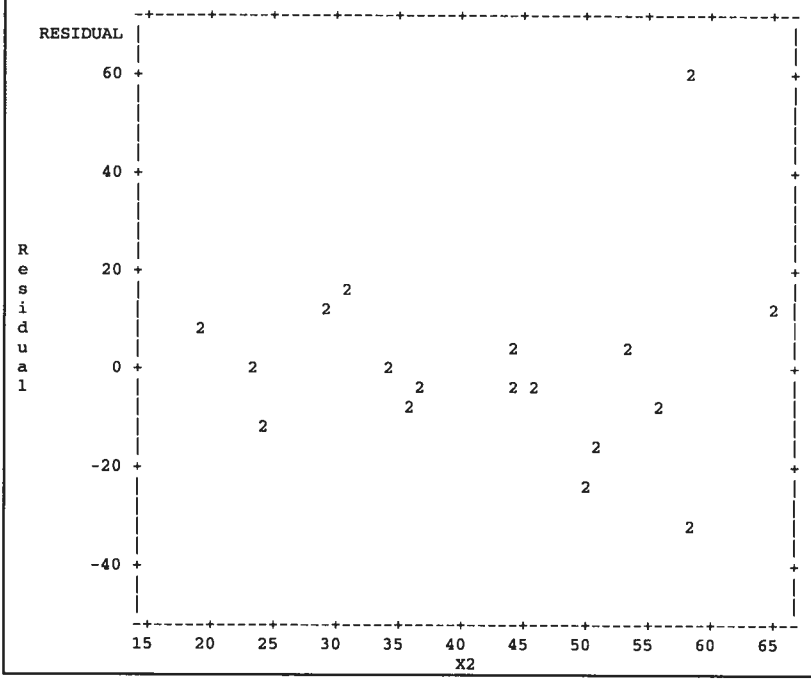
Archivo de resultados (C13-3.LST).-

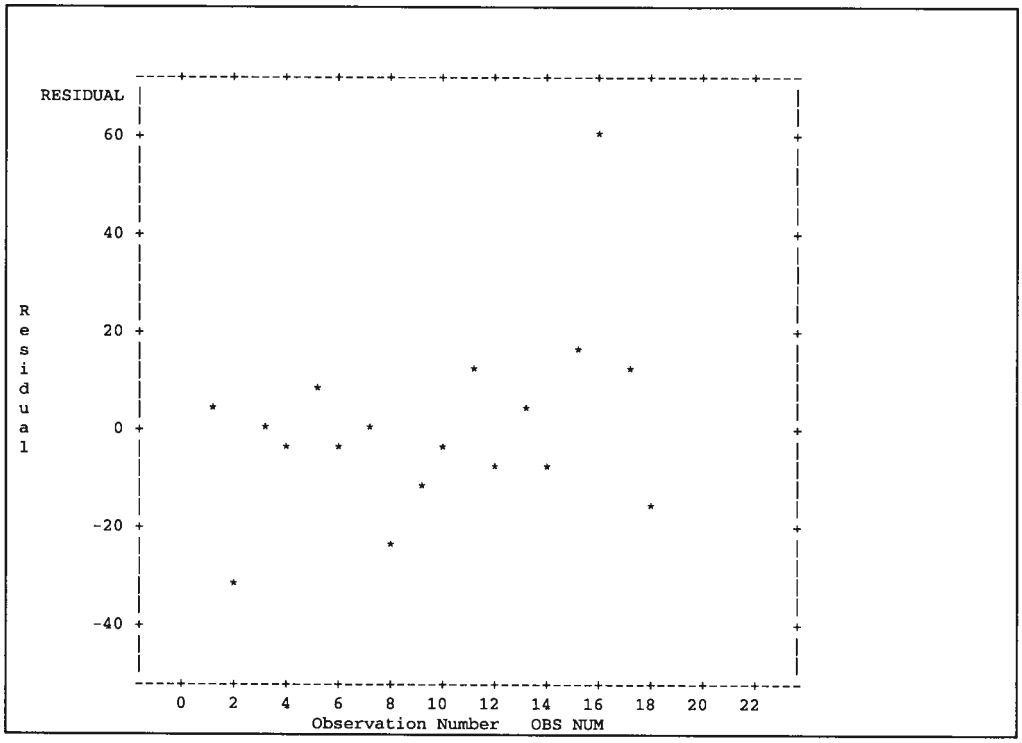
Regresión Múltiple: Residuos

Obs	Dep Var Y	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	64.0000	61.5593	10.597	2.4407	17.757	0.137	| | |	0.002
2	51.0000	83.8278	8.071	-32.8278	19.038	-1.724	***| | |	0.178
3	60.0000	58.9599	8.994	1.0401	18.620	0.056	| | |	0.000
4	76.0000	78.9656	5.240	-2.9656	20.004	-0.148	| | |	0.001
5	71.0000	63.4457	9.817	7.5543	18.199	0.415	| | |	0.017
6	96.0000	101.6	7.462	-5.5807	19.285	-0.289	| | |	0.004
7	61.0000	60.2710	7.440	0.7290	19.294	0.038	| | |	0.000
8	77.0000	101.9	7.367	-24.9273	19.321	-1.290	*| | |	0.081
9	54.0000	66.7425	8.278	-12.7425	18.949	-0.672	*| | |	0.029
10	93.0000	98.7227	7.027	-5.7227	19.448	-0.294	| | |	0.004
11	77.0000	64.9258	14.018	12.0742	15.202	0.794	| |* |	0.179
12	95.0000	102.4	7.906	-7.4472	19.107	-0.390	| | |	0.009
13	81.0000	76.8875	5.235	4.1125	20.005	0.206	| | |	0.001
14	54.0000	62.7710	6.955	-8.7710	19.474	-0.450	| | |	0.009
15	93.0000	77.0139	6.457	15.9861	19.644	0.814	| |* |	0.024
16	168.0	109.2	9.235	58.7574	18.502	3.176	| |******	0.838
17	93.0000	79.5252	7.241	13.4748	19.369	0.696	| |* |	0.023
18	99.0000	114.2	10.161	-15.1844	18.009	-0.843	| *| |	0.075
19	.	69.5891	6.679	.	.	.		.

```
 20          .    125.0  13.613          .      .          .                          .
 21          .  81.1886   4.874          .      .          .                          .
Sum of Residuals                          0
Sum of Squared Residuals           6413.9426
Predicted Resid SS (Press)         9776.5478
```







Las salidas de este programa (**C13-3.LST**) son:

Las primeras salidas ya se conocen de los anteriores ejemplos, por lo que la hemos excluido. Y las salidas que son por columnas, las tres primeras columnas son conocidas, las columnas nuevas son

La columna encabezada por **Residual** son los residuos. Los errores típicos de los residuos son los encabezados por **Std Err Residual**. Y la división de los residuos por su error típico da el residuo Student (**Student Residual**).

La siguiente columna es una representación gráfica de la magnitud del residuo Student, como se ve, ninguna supera el dos, y la mayoría vale cero, lo que significa un buen ajuste.

La siguiente columna, la *D* de *Cook* se estudiará en el siguiente apartado

Tal como era de esperar por el criterio minimocuadrático seguido para la estima de la recta, las sumas de los residuos vale cero, y la suma de los cuadrados de los residuos (que es la más pequeña posible) es igual a la de la $SC_{(error)}$ de la prueba de bondad de ajuste, con una ligera variación debido a errores de redondeo.

Como se observa en la columna correspondiente, los errores típicos de los residuos oscilan alrededor de los mismos valores, no hay ningún error típico excesivamente grande. Y en la columna de los residuos Student se comprueba que ningún valor sobrepasa en valor de 25, por lo que se puede concluir que no hay ningún dato excesivamente alejado de su valor esperado.

Como se ve en las salidas de los ***PLOT***, aproximadamente la mitad de los residuos están por encima del cero y la otra mitad están por debajo del cero, no

observándose ninguna tendencia entre unos y otros. Esto ya se sabía viendo la columna encabezada por **Residuals**. La ventaja de la gráfica es que si hay una desviación significativa se puede visualizar la tendencia de esta desviación, como se puede comprobar cuando se hace ajustes a *Curvas polinómicas*.

Análisis de las *influencias*.-

Los valores extremos pueden, a veces, pasar desapercibidos al análisis de los residuos y sus valores Student del epígrafe anterior, porque las estimaciones minimocuadráticas de la recta de regresión tienden a situarse en valores intermedios de las observaciones extremas. Por tanto, las estimas de los residuos de dichas observaciones pueden no ser especialmente grandes, interfiriendo en la búsqueda de valores extremos. Esta dificultad puede ser soslayada si se calculan las estimas y estadísticos de la observación cuestionada por medio de una recta de regresión estimada con todos los pares de valores menos con el punto cuestionado. Esto no requiere la repetición de todos los cálculos, pues basta con restarle a los sumatorios básicos (ΣX, ΣX^2, ΣY, ΣY^2 y (ΣXY) el valor correspondiente de la observación eliminada. En todo caso, los paquetes estadísticos proveen de estos estadísticos.

Dichos estadísticos se anotan con el subíndice *-i* refiriéndose este subíndice al subíndice de la observación omitida. Por ejemplo, $S^2_{b\text{-}2}$, es la varianza de la regresión estimada sin el segundo valor de las parejas de valores que se tienen. Estos estadísticos indican la potencial ***influencia*** de una observación concreta.

Residuos *stundetizados* (Rstudent).-

El primer estadístico es una versión de los *residuos stundetizados*, en el que los residuos se dividen, como en el caso anterior, por los errores típicos de los residuos, pero utilizando para el cálculo de dichos errores típicos, la varianza del error calculada con todos los pares de valores menos con el par en cuestión, es decir, $S^2_{Y.X\text{-}i}$.

Este residuo stundetizado se distribuye con la misma *t* y con los mismos *gl* que el residuo Student, pero es más sensible que este, siendo el criterio de rechazo el mismo, es decir, rechazar por extremos las observaciones que de un residuo stundetizado mayor de 25.

Diferencias de ajustes (Dffits).-

Un segundo estadístico es el que se puede denominar *diferencias de ajustes*, consiste en la diferencia entre dos valores estimados de la misma *Y*, la primera estimación es la realizada con la ecuación de la recta calculada con todos los pares de valores, es decir, $\hat{Y}_i$, y la segunda es la estimación es realizada con la ecuación de la recta calculada con todos los pares de valores menos con el par en cuestión, es decir, $\hat{Y}_{i\text{-}i}$. Esta diferencia es dividida por el error típico de un valor estimado, pero utilizando la varianza del error calculada con todos los pares de valores menos el par en cuestión, esto es, con la simbolizada como $S^2_{Y.X\text{-}i}$.

Este estadístico es un buen indicador de la *influencia*. Se sugiere, que un buen criterio para detectar observaciones influyentes, es el de las observaciones que superen, en valor absoluto, el valor

$$2\sqrt{\frac{m+1}{n}},$$

siendo m el número de variables independientes y n el número de pares de valores.

D de *Cook* (Cook.s D).-

Un tercer estadístico es la *D* de *Cook*, que es semejante al anterior, pero en éste, a diferencia del anterior, se utiliza la varianza del error de todas las observaciones, es decir, el error típico de un valor estimado; y se eleva al cuadrado y divide por dos, para resaltar más los valores extremos.

Medida tipificada de lo extrema que es una observación (Hat Diag H).-

Otro estadístico sería h_i, que indica la influencia de cada observación. Este es una medida tipificada de lo extrema que es una observación, en el espacio de las X, con respecto al centro.

La suma de todos los h_i vale m+1, siendo m el número de variables independientes, por tanto, el valor esperado de h_i es

$$\frac{m+1}{n}$$

si una observación supera dos veces este valor, se puede considerar que tiene una gran influencia.

Proporción de las varianzas generalizadas (Cov Ratio).-

Otro estadístico es la *proporción de las varianzas generalizadas*. La varianza generalizada de una muestra de pares de valores, es la razón entre la varianza del error y la suma de cuadrados de la variable independiente por el número de observaciones, esto es

Este estadístico es el resultado de dividir la varianza generalizada sin la *i-ésima* observación y la varianza generalizada con todas las observaciones, es decir, Si el valor de este estadístico es superior a uno indica que la inclusión de dicha observación tiene como resultado incrementar la precisión, mientras que un valor inferior a uno tiene como resultado una disminución de la precisión. Se sugiere que valores que excedan a la unidad en

$$\frac{3(m+1)}{n}$$

pueden ser considerados como excesivos.

Diferencias tipificadas de las *b* y de la *a* (... Dfbetas, INTERCEP Dfbetas).-

Otros estadísticos son la diferencias tipificada de las regresiones y de la ordenada en el origen calculadas con todos y sin el *i-ésimo* valor. Se pueden rechazar como extremos las observaciones que den un valor de *dfbetas* superior a 2.

También puede ser de utilidad comparar la suma de cuadrados residual (suma de los cuadrados de todos los residuos) con la suma de cuadrados residual estimada, esto es, la suma de los cuadrados de los residuos obtenidos restándole al valor observado el valor estimando con arreglo a la ecuación calculada con todos los demás valores. La suma de cuadrados residual estimada se espera sea mayor que la original, y será mucho mayor cuanto más valores extremos haya.

Archivo de programa SAS (C13-4.SAS).-

Las estimas de las influencias se realiza con la opción **INFLUENCE**.

```
title 'Regresión Múltiple: Influencias';
options ls=75 ps=40;
data regresi;
infile 'c13-1.dat';
input suelo X1 X2 Y @@;
proc reg;
  model Y = X1 X2 / Influence;
run;
```

Archivo de resultados (C13-4.LST).-

```
                    Regresión Múltiple: Influencias

                          Hat Diag      Cov                INTERCEP     X1        X2
Obs  Residual  Rstudent        H       Ratio    Dffits      Dfbetas    Dfbetas   Dfbetas
  1    2.4407    0.1329    0.2626      1.6617    0.0793     -0.0231   -0.0637    0.0560
  2  -32.8278   -1.8604    0.1523      0.7479   -0.7886      0.4235    0.2617   -0.6278
  3    1.0401    0.0540    0.1892      1.5160    0.0261      0.0229   -0.0080   -0.0144
  4   -2.9656   -0.1433    0.0642      1.3086   -0.0375     -0.0235   -0.0028    0.0133
  5    7.5543    0.4033    0.2254      1.5337    0.2176      0.2105   -0.0110   -0.1622
  6   -5.5807   -0.2803    0.1302      1.3905   -0.1085     -0.0083   -0.0795    0.0182
  7    0.7290    0.0365    0.1294      1.4124    0.0141      0.0073   -0.0090   -0.0009
  8  -24.9273   -1.3219    0.1269      0.9899   -0.5040      0.0613   -0.3215   -0.0278
  9  -12.7425   -0.6597    0.1602      1.3361   -0.2882     -0.2696    0.0197    0.1968
 10   -5.7227   -0.2851    0.1155      1.3666   -0.1030     -0.0180   -0.0735    0.0249
 11   12.0742    0.7840    0.4595      2.0005    0.7229     -0.3830   -0.5202    0.6256
 12   -7.4472   -0.3785    0.1462      1.3972   -0.1566      0.0613   -0.0703   -0.0574
 13    4.1125    0.1989    0.0641      1.3031    0.0520      0.0064   -0.0177    0.0143
 14   -8.7710   -0.4381    0.1131      1.3313   -0.1565     -0.0722    0.0995   -0.0010
 15   15.9861    0.8041    0.0975      1.1902    0.2643      0.2226    0.0451   -0.1693
 16   58.7574    5.3603    0.1995      0.0540    2.6757     -1.0570    1.5146    0.8041
 17   13.4748    0.6832    0.1226      1.2705    0.2554      0.2202    0.0818   -0.1888
```

18	-15.1844	-0.8346	0.2415	1.4017	-0.4709	0.0433	-0.3844	0.0429
19	.	.	0.1043	.	.	.	.	.
20	.	.	0.4334	.	.	.	.	.
21	.	.	0.0556	.	.	.	.	.

Sum of Residuals	0
Sum of Squared Residuals	6413.9426
Predicted Resid SS (Press)	9776.5478

En el archivo de resultados, se han eliminado las salidas repetidas de anteriores programas, presentando solamente las *influencias*, en las que se observa.

- La columna encabezada por **Residual** son los mismos residuos explicados en el anterior ejemplo

- La columna encabezada por **Rstudent** son los *residuos stundetizados*, como se observa, ningún valor supera el 2.5 por lo que no hay ninguna observación que influya excesivamente en la regresión.

- La siguiente columna, la encabezada por **Hat Diag H**, es la h_i, como se ha dicho, la suma de los 18 h_i vale 3 (dos variables independientes más uno), por lo que el valor esperado de h_i es 3/18=0.1666, y dos veces este valor esperado da el valor crítico, esto es 0.3333. Como se observa, ninguna h_i supera este valor crítico, por lo que no hay ninguna observación que influya excesivamente en la regresión.

- La siguiente columna, la encabezada por **Cov Ratio** es la *proporción de las varianzas generalizadas*. Como se ve, todas (menos la 2, la 8 y la 16) superan la unidad por lo que todas estas observaciones incrementan la precisión del análisis de regresión, y de las tres observaciones que disminuyen la precisión del análisis (por ser menores que la unidas) solamente la observación 16 esta disminuyendo claramente la precisión.. Para saber si la influencia de una observación es excesiva, el valor crítico es la unidad más 3(2+1)/18=0.5, esto es, 1.5, y como se ve, solamente el 1 y el 11 superan este valor.

- La siguiente columna, la encabezada por **Dffits** es la *diferencias de ajustes*, el valor crítico para este estadístico es

$$2\sqrt{\frac{m+1}{n}} = 2\sqrt{\frac{2+1}{18}} = 0.8165$$

y como se ve este valore es ampliamente superado por la observación 16 y ligeramente superado por la observación 5.

- La tres últimas columnas son las de las *diferencias tipificadas* para la ordenada en el origen y las dos regresiones, respectivamente. Como se ve ninguno en ninguna observación se supera, en valor absoluto, el 2, por lo que se concluye que, según este criterio no hay ninguna observación excesivamente influyente.

- Falta la *D de Cook*, que el SAS la saca con la opción **R**, por lo tanto está en la salida del ejemplo C13-3. Como se ha dicho anteriormente, este estadístico es semejante a *Dffits* y tiene el mismo valor crítico, por lo tanto,

mirando la salida C13-3.lst, se observa que solo la observación 16 supera ligeramente este valor crítico.

Valores de regresión y valores ajustados.-

Los valores de Y determinados por la ecuación de regresión son ***valores de regresión***, y por lo tanto son estimaciones de parámetros poblacionales, es decir, de $\mu_{Y.X} = a + b_1X_1 + b_2X_2$. Las diferencias entre éstos y los valores observados, esto es, los residuos tal como se han calculado anteriormente, son estimaciones de la variación de Y no explicada por la variación de las X. En la siguiente tabla se presentan los residuos observados, $e_{Y.X}$, para el ejemplo del FÓSFORO (se obtienen de la salida del programa C13-3)

Y	$\hat{Y}$	$e_{Y.X}$	$\overline{Y} + e_{Y.X}$
64.0	61.55	2.44	83.72
51.0	83.82	-32.82	48.46
60.0	58.95	1.04	82.32
76.0	78.96	-2.96	78.32
71.0	63.44	7.55	88.83
96.0	101.60	-5.58	75.70
61.0	60.27	0.72	82.00
77.0	101.90	-24.92	56.36
54.0	66.74	-12.74	68.54
93.0	98.70	-5.72	75.56
77.0	64.92	12.07	93.35
95.0	102.40	-7.44	73.84
81.0	76.88	4.11	85.39
54.0	62.77	-8.77	72.51
93.0	77.01	15.98	97.26
168.0	109.20	58.75	140.03
93.0	79.52	13.47	94.75
99.0	114.20	-15.18	66.10

Teniendo en cuenta lo que se dijo en el epígrafe *Fuentes de variación de la regresión*, si a la media de la variable dependiente se le suma los residuos se tendrá los valores ajustados (última columna de tabla anterior) de dicha variable, esto es, los valores que tendría la variable dependiente si se le ha eliminado la contribución debida a la regresión. Es como si el contenido de fósforo del maíz cultivado en cada suelo se moviera paralelamente al plano de regresión muestral para cada valor de las X y se midiese entonces como un nuevo valor ajustado de Y, es decir, en este ejemplo son los contenido en fósforo de la planta ajustados, que son los esperados si todos los suelos tuvieran la misma cantidad de fósforo orgánico y de fósforo inorgánico. Estos se obtienen sumándole los residuos a la media $\overline{Y}$=81.28

$$Y_{adj} = \overline{Y} + e_{Y.X} = Y - b_1(X_1 - \overline{X_1}) - b_2(X_2 - \overline{X_2})$$

Las comparaciones entre medias ajustadas son muy útiles. Para entender mejor las siguiente explicaciones pongamos un ejemplo semejante al del fósforo,

Ejemplo.-

Se tienen los promedio del aumentos diarios de peso (Y, en libras) de 40 cerdos sometidos a cuatro tratamientos. Probablemente este ritmo de incremento de peso es ***predecible*** por la edad (X_1) y el peso (X_2) con los que comenzaron el experimento cada cerdo, por lo que el aumento diario de peso es función lineal de la edad y peso de inicio. Los datos son

Trat	X_1	X_2	Y
1	78	61	1.40
1	90	59	1.69
1	94	76	1.62
1	71	50	1.47
1	99	61	1.26
1	80	54	1.28
1	83	57	1.34
1	75	45	1.55
1	62	41	1.57
1	67	40	1.26
2	78	74	1.61
2	99	75	1.21
2	80	64	1.12
2	75	48	1.35
2	94	62	1.29
2	91	42	1.24
2	75	52	1.29
2	63	43	1.43
2	62	50	1.29
2	67	40	1.26
3	78	80	1.67
3	83	61	1.41
3	79	62	1.73
3	70	47	1.23
3	85	59	1.29
3	83	42	1.22
3	71	47	1.39
3	66	42	1.39
3	67	40	1.56
3	67	40	1.66
4	77	62	1.40
4	71	55	1.47
4	78	62	1.37
4	70	43	1.15
4	95	57	1.22
4	96	51	1.28
4	71	41	1.31
4	63	40	1.27
4	62	45	1.22
4	67	39	1.36
Σ	3082	2109	55.83
$\overline{X}$	77.05	52.72	1.40

Si esta función lineal es cierta, los efectos de los tratamientos y del error pueden mejorarse ajustando los valores de los incrementos de peso. Esto es, puede ser interesante saber cual es el incremento de peso diario debido a los *tratamientos* quitándole la influencia de la edad y del peso inicial, como si los cuarenta cerdos hubieran comenzado la experiencia con la misma edad y el mismo peso. Ese aumento diario de peso estimado o ajustado es la media (1.40) más los residuos. Ahora se puede hacer un análisis de varianza de estos valores ajustados, lo que nos dará si existe diferencia para el incremento de peso diario entre los cuatro tratamientos como si todos los cerdos fueran de la misma edad y peso.

Archivo de programa SAS (C13-5.SAS).-

Tanto el *peso inicial* como el *ritmo* del aumento de peso diario se han transformado a gramos

```
title 'Valores ajustados';
Options ls=75 ps=60;
Data covar;
Infile 'c13-5.dat';
Input trata edadini pesoini ritmo @@;
Pesoini=pesoini*453.59;
Ritmo=ritmo*453.59;
Proc reg ;
  Model ritmo = edadini pesoini;
   output out=residuos R=r_ritmo;
run;
proc anova;
  class trata;
  model ritmo = trata;
run;
  proc anova ;
  class trata;
  model r_ritmo = trata;
run;
proc glm ;
  class trata;
  model ritmo = edadini pesoini trata / ss3 solution;
run;
```

Archivo de datos (C13-5.DAT).-

```
1 78 61 1.40  2 78 74 1.61  3 78 80 1.67  4 77 62 1.40
1 90 59 1.69  2 99 75 1.21  3 83 61 1.41  4 71 55 1.47
1 94 76 1.62  2 80 64 1.12  3 79 62 1.73  4 78 62 1.37
1 71 50 1.47  2 75 48 1.35  3 70 47 1.23  4 70 43 1.15
1 99 61 1.26  2 94 62 1.29  3 85 59 1.29  4 95 57 1.22
1 80 54 1.28  2 91 42 1.24  3 83 42 1.22  4 96 51 1.28
1 83 57 1.34  2 75 52 1.29  3 71 47 1.39  4 71 41 1.31
1 75 45 1.45  2 63 43 1.43  3 66 42 1.39  4 63 40 1.27
1 72 41 1.47  2 62 50 1.29  3 77 40 1.46  4 62 45 1.22
1 77 40 1.26  2 67 40 1.26  3 77 40 1.46  4 67 39 1.36
```

Archivo de resultados (C13-5.LST).-

```
                          Valores ajustado
Model: MODEL1
Dependent Variable: RITMO
                        Analysis of Variance
                          Sum of          Mean
  Source            DF      Squares        Square      F Value       Prob>F
  Model              2  32009.69233  16004.84617        4.234       0.0221
  Error             37 139864.12255   3780.11142
  C Total           39 171873.81488

       Root MSE       61.48261     R-square       0.1862
       Dep Mean      619.49054     Adj R-sq       0.1423
       C.V.            9.92471
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	603.567852	74.96536971	8.051	0.0001
EDADINI	1	-1.830588	1.13781839	-1.609	0.1161
PESOINI	1	0.006640	0.00228183	2.910	0.0061

Analysis of Variance Procedure
Dependent Variable: RITMO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	28423.0038	9474.3346	2.38	0.0860
Error	36	143450.8111	3984.7448		
Corrected Total	39	171873.8149			

R-Square	C.V.	Root MSE	RITMO Mean
0.165371	10.18980	63.1248	619.491

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	3	28423.0038	9474.3346	2.38	0.0860

Analysis of Variance Procedure
Dependent Variable: R_RITMO Residual

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	31041.7474	10347.2491	3.42	0.0273
Error	36	108822.3751	3022.8438		
Corrected Total	39	139864.1225			

R-Square	C.V.	Root MSE	R_RITMO Mean
0.221942	9999.99	54.9804	0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRATA	3	31041.7474	10347.2491	3.42	0.0273

General Linear Models Procedure
Dependent Variable: RITMO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	63618.9313	12723.7863	4.00	0.0059
Error	34	108254.8836	3183.9672		
Corrected Total	39	171873.8149			

R-Square	C.V.	Root MSE	RITMO Mean
0.370149	9.108558	56.4267	619.491

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EDADINI	1	14443.5904	14443.5904	4.54	0.0405
PESOINI	1	34634.6176	34634.6176	10.88	0.0023
TRATA	3	31609.2389	10536.4130	3.31	0.0316

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		605.8852354 B	8.67	0.0001	69.91002301
EDADINI		-2.2733515	-2.13	0.0405	1.06736584
PESOINI		0.0069725	3.30	0.0023	0.00211406
TRATA	1	54.1663600 B	2.08	0.0448	25.99695505
	2	-7.8508090 B	-0.31	0.7613	25.63523146
	3	50.8435478 B	2.01	0.0527	25.32621330
	4	0.0000000 B	.	.	.

```
NOTE: The X'X matrix has been found to be singular and a generalized
      inverse was used to solve the normal equations.  Estimates followed
      by the letter 'B' are biased, and are not unique estimators of the
      parameters.
```

Las salidas de este programa (**C13-5.LST**) son:

La salida correspondiente al primer procedimiento se podría haber evitado poniendo la opción **NOPRINT** en el **proc reg**, pero es interesante contrastar los valores de los dos coeficientes de regresión con los que se obtendrán en el último procedimiento. Como se observa, no es significativa la regresión del *ritmo* de crecimiento con la edad inicial y si es significativa la regresión con el peso inicial.

La salida del segundo procedimiento es el análisis de los efectos de los tratamientos sobre el ritmo de crecimiento sin considerar que cada cerdo comienza con un peso y una edad diferente. Este análisis da no significativo al nivel 0.05, por lo que habría que concluir que los tratamientos no son efectivos sobre el ritmo de crecimiento.

La salida del tercer procedimiento es el análisis de los efectos de los tratamientos sobre la variable **R_RITMO**. Los valores de esta variable son los residuos del ritmo de crecimiento en base a la recta (plano) de regresión con la edad y el peso inicial, obtenidos en el primer procedimiento. Por lo tanto, este equivale al análisis de varianza de los tratamiento sobre los ritmos de crecimiento ajustados para la edad y el peso inicial. Este análisis si da significativo al 0.05, esto es, los tratamientos si son efectivos sobre el ritmo de crecimiento, si todos los cerdos comienzan la experiencia con la misma edad y el mismo peso. Lo que ha ocurrido en el procedimiento anterior es que la variabilidad, del ritmo de crecimiento, debida a los diferentes pesos iniciales y diferentes edades iniciales, solapaba la variabilidad debida a los tratamientos, no pudiendo ser detectada ésta por el ANOVA anterior.

Obsérvese que la suma de cuadrados total (del ritmo de crecimiento) es de 171873.8149 (primero, segundo y cuarto procedimiento), mientras que en el análisis de los datos ajustados (el tercer procedimiento) ha bajado a 139864.1225, pues se le ha quitado la variabilidad debida a la regresión, esta es, 32009.6923 (primer procedimiento).

No es preciso hacer los tres primeros procedimientos, se han hecho con fines didácticos, con hacer el cuarto procedimiento es suficiente y más exacto, como se verá más adelante.

El cuarto procedimiento es el del ***Análisis de Covarianza*** (ver Capítulo 17), este es un análisis de varianza de los tratamientos sobre los ritmos ajustados para la edad y el peso inicial, esto es, como si todos los cerdos tuvieran la misma edad y peso inicial, y también es un análisis de regresión con los ritmos ajustados para los tratamientos, esto es, como si todos los cerdos estuvieran sometidos al mismo tratamiento (misma dieta).

En la salida de este procedimiento se observa, primeramente, que los tratamientos son significativos sobre los ritmos ajustados, cuando el análisis de varianza dio no significativo. Y en segundo lugar se observa que las dos regresiones, con los ritmos ajustados para los tratamientos, son significativas, cuando el análisis de

regresión dio no significativa la regresión con la edad inicial. Los valores de los coeficiente de regresión se tienen en la salida de la opción SOLUTION, esto son, -2.2733* con la edad al inicio y 0.0069** con el peso al inicio.

Obsérvese que la suma de cuadrados debida a los tratamientos en este análisis (31609.2389) no vale exactamente lo mismo que la suma de cuadrados de los tratamientos con los ritmos ajustados (31041.7474; tercer procedimiento) porque aquí se ha ajustado los datos para la regresión hallada con los valores corregidos para los tratamientos (las regresiones de la salida de SOLUTION) y allí se ajustaron para la regresión de los datos originales, sin corregir.

Prueba de homogeneidad de dos o más líneas de regresión.-

A menudo el experimentador obtiene dos o más líneas de regresión a partir de datos análogos y desea saber si las relaciones funcionales descritas por las ecuaciones de regresión son las mismas o diferentes. Por ejemplo, se puede haber establecido la regresión de la concentración sanguínea de colesterol sobre la edad en una muestra de individuos y se puede desear, ahora, comparar esta ecuación de regresión con la de otra u otras muestras sometidas a una dieta diferente. Por lo tanto, lo que se desea es contrastar la ***homogeneidad* de los *b***, es decir, determinar si las pendientes halladas pueden considerarse o no estimaciones de un β común. El diseño básico de tal tipo de prueba es el del análisis de varianza. Existirán *t* muestras representando los grupos de tratamiento y el control, si lo hay.

Existe, sin embargo, un nuevo aspecto de bastante importancia: en los análisis de varianza realizados hasta el momento lo han sido para una o varias variables, consideradas individualmente o en conjunto. Por ejemplo, ahora además del nivel de colesterol (*Y*) se tiene la edad del individuo (*X*). De manera que son posibles dos análisis separados de varianza univariante, uno para cada variable, un análisis multivariante de la varianza para las dos variables, y también un análisis conjunto de la ***covarianza de X e Y***. Es decir, tenemos un análisis de la covarianza, que aunque se estudiará más detenidamente en el Capítulo 17, se verá su utilidad en la comparación o prueba de homogeneidad de varios coeficientes de regresión.

Considérese el modelo de una variable independiente en un diseño de una vía, de manera que

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$$

donde el subíndice, *i*, indica el grupo o tratamiento. La hipótesis que se quiere contrastar es

$$H_0 : \beta_i = \beta_{i'}$$

para todo $i \neq i'$.

Una formulación alternativa del modelo es

$$Y_{ij} = \bar{\alpha} + \alpha_i + \bar{\beta} X_{ij} + \beta_i X_{ij} + \varepsilon_{ij}$$

donde $\overline{\alpha}$ es la media de las ordenadas en el origen, $\overline{\beta}$ es la media de las pendientes, y α_i y β_i son la ordenada y la pendiente en cada tratamiento. La hipótesis que se quiere contrastar es

$$H_0: \beta_i = 0$$

para $i = 1, 2, ..., t$.

Nótese que las diferencias entre las ordenadas en el origen son irrelevantes para ambas hipótesis.

La diferencia de la regresión entre los grupos de los tratamientos de hecho refleja la interacción entre los tratamientos y la variable independiente o covariable. Por tanto, se puede probar esta hipótesis como la hipótesis de la interacción de la siguiente manera.

Ejemplo.-

Siguiendo con el ejemplo anterior, se nos puede plantear la cuestión de si el ritmo de incremento de peso diario como función del peso inicial y de la edad inicial es mayor en un tratamiento que en otro. Dicho de otra manera, podría ocurrir que la recta (o plano) de regresión en un/os tratamiento fuera más empinada que en otro/s tratamiento.

Archivo de programa SAS (C13-6.SAS).-

```
Title 'Prueba de homogeneidad de varias regresiones';
options ls=75 ps=60;
data homoreg;
infile 'c13-5.dat';
input trata edadini pesoini ritmo @@;
pesoini=pesoini*453.59;
ritmo=ritmo*453.59;
proc sort;
by trata;
run;
proc reg;
model ritmo = edadini pesoini;
by trata;
run;
proc glm;
class trata;
model ritmo =  edadini pesoini trata edadini*pesoini*trata ;
run;
*Hubiera sido igual expresar el modelo de la siguiente forma;
*model ritmo =  edadini pesoini trata edadini*pesoini(trata) ;
```

Archivo de resultados (C13-6.LST).-

```
---------------------------------- TRATA=1 ----------------------------------

Model: MODEL1
Dependent Variable: RITMO

                          Analysis of Variance
                             Sum of          Mean
  Source           DF       Squares        Square      F Value       Prob>F
  Model             2    7475.78809    3737.89404        0.795       0.4884
  Error             7   32899.39251    4699.91322
  C Total           9   40375.18060

      Root MSE       68.55591     R-square       0.1852
      Dep Mean      645.91216     Adj R-sq      -0.0477
      C.V.           10.61381

                       Parameter Estimates
                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   INTERCEP   1    633.545098  205.09497999         3.089        0.0176
   EDADINI    1     -2.386330    3.66118252        -0.652        0.5353
   PESOINI    1      0.008422    0.00701439         1.201        0.2689
---------------------------------- TRATA=2 ----------------------------------

Model: MODEL1
Dependent Variable: RITMO

                          Analysis of Variance
                             Sum of          Mean
  Source           DF       Squares        Square      F Value       Prob>F
  Model             2    7202.22480    3601.11240        0.975       0.4232
  Error             7   25858.76058    3694.10865
  C Total           9   33060.98538

      Root MSE       60.77918     R-square       0.2178
      Dep Mean      593.74931     Adj R-sq      -0.0056
      C.V.           10.23651

                       Parameter Estimates
                    Parameter       Standard    T for H0:
   Variable  DF      Estimate          Error   Parameter=0    Prob > |T|
   INTERCEP   1    684.060759  126.69170519         5.399        0.0010
   EDADINI    1     -2.552321    1.89734265        -1.345        0.2205
   PESOINI    1      0.004401    0.00415017         1.060        0.3242

---------------------------------- TRATA=3 ----------------------------------

Model: MODEL1
Dependent Variable: RITMO

                          Analysis of Variance
                             Sum of          Mean
  Source           DF       Squares        Square      F Value       Prob>F
  Model             2   19449.41584    9724.70792        2.043       0.2000
  Error             7   33313.60426    4759.08632
  C Total           9   52763.02010
```

```
      Root MSE        68.98613     R-square        0.3686
      Dep Mean       646.36575     Adj R-sq        0.1882
      C.V.            10.67292

                          Parameter Estimates
                    Parameter       Standard     T for H0:
   Variable  DF      Estimate          Error    Parameter=0     Prob > |T|
   INTERCEP   1    637.038073  286.11600476          2.227         0.0613
   EDADINI    1     -2.450850    3.99038024         -0.614         0.5585
   PESOINI    1      0.008386    0.00415917          2.016         0.0836

--------------------------------- TRATA=4 ---------------------------------

Model: MODEL1
Dependent Variable: RITMO

                          Analysis of Variance
                             Sum of          Mean
  Source           DF       Squares        Square      F Value       Prob>F
  Model             2   5382.23747   2691.11873        1.587       0.2701
  Error             7  11869.38755   1695.62679
  C Total           9  17251.62502

      Root MSE        41.17799     R-square        0.3120
      Dep Mean       591.93495     Adj R-sq        0.1154
      C.V.            6.95651

                          Parameter Estimates
                    Parameter       Standard     T for H0:
   Variable  DF      Estimate          Error    Parameter=0     Prob > |T|
   INTERCEP   1    551.798441   92.91144670          5.939         0.0006
   EDADINI    1     -1.603653    1.38441184         -1.158         0.2847
   PESOINI    1      0.007144    0.00403621          1.770         0.1200

                    General Linear Models Procedure

Dependent Variable: RITMO
                                  Sum of          Mean
Source                     DF     Squares        Square  F Value     Pr > F
Model                       9  68417.1649   7601.9072      2.20     0.0506
Error                      30 103456.6500   3448.5550
Corrected Total            39 171873.8149

                  R-Square         C.V.      Root MSE          RITMO Mean
                  0.398066     9.479467       58.7244              619.491

Source                     DF       Type I SS  Mean Square  F Value     Pr > F
EDADINI                     1          0.1856       0.1856     0.00     0.9942
PESOINI                     1   32009.5067   32009.5067      9.28     0.0048
TRATA                       3   31609.2389   10536.4130      3.06     0.0435
EDADIN*PESOINI*TRATA        4    4798.2336    1199.5584      0.35     0.8434

Source                     DF      Type III SS  Mean Square  F Value     Pr > F
EDADINI                     1          0.19615      0.19615     0.00     0.9940
PESOINI                     1   1606.84393   1606.84393      0.47     0.5001
TRATA                       3    877.56543    292.52181      0.08     0.9678
EDADIN*PESOINI*TRATA        4    4798.23359   1199.55840     0.35     0.8434
```

Las salidas de este programa (**C13-6.LST**) son:

El primer procedimiento del archivo programa (**SORT**) no tiene salida, su función es ordenar los datos por tratamientos para que el siguiente procedimiento pueda hacer las regresiones por tratamientos.

El segundo procedimiento (**REG**) ha calculado las cuatro rectas de regresión, correspondientes a los cuatro tratamientos (**BY TRATA**). Se observa que son no significativas las pendientes de los cuatro planos de regresión, si bien los valores de ***F*** del modelo de las dietas tres y cuatro son mayor que uno y los de las otras dietas están próxima a la unidad, lo que podría indicar que si el tamaño de muestra dentro de cada dieta fuera mayor podrían ser significativas las pendientes.

El tercer procedimiento es el que hacer la prueba de homogeneidad de las rectas de regresión. En este caso hay que mirar las suma de cuadrados del **Type I,** que nos da la siguiente información:

(a) **EDADINI**: esta suma de cuadrados es debida a la regresión simple del ritmo de aumento de peso sobre la edad inicial, ignorando los tratamientos. Se observa que es muy pequeña y no significativa.

(b) **PESOINI**: esta suma de cuadrados es debida a la regresión simple del ritmo de aumento de peso sobre el peso inicial más la regresión simple de la anterior, ignorando los tratamientos. La suma de esta SC más la anterior da la suma de cuadrados del modelo de la regresión del primer procedimiento de ejemplo C13-5. Esta regresión es significativa.

(c) **TRATA**: esta suma de cuadrados es la debida a las cuatro ordenadas en el origen, o lo que es lo mismo, es la debida a los efectos tratamientos sobre el *ritmo* ajustados para el peso inicial y la edad inicial, y esta es la misma que la obtenida en el último procedimiento del ejemplo C13-5. Los tratamientos de los ritmos ajustados son significativos

(d) **EDADINI*PESOINI*TRATA** es la suma de cuadrados debida a los cuatro planos de regresión. Los cuatro planos de regresión son iguales por lo tanto se pueden considerar como estimas de unos mismo coeficiente de regresión paramétrico, esto es, de una misma población.

Importancia relativa de diferentes variables *X*.-

En un análisis de regresión múltiple puede surgir la pregunta de cuál o cuales variables *X* son más importantes para determinar el valor de *Y*. Esta misma pregunta se presentará en el análisis discriminante (ver epígrafe *Selección de variables en el Análisis Discriminante*, del Capítulo 19).

Se han probado varios enfoques para responder preguntas de este tipo. Considérese primero que el objetivo del estudio fuera el de predecir *Y* o *explicar* la variación de *Y* en función de las *X*. Este problema sería muy fácil de resolver si las *X* fueran independientes (no estuvieran correlacionadas). Es decir, que si tenemos el modelo

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

se tendría, si las X son independientes, que

$$\sigma_Y^2 = \beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2+...+\beta_k^2\sigma_k^2$$

donde σ_i^2 designa la varianza de X_i. Por lo que la fracción

$$\beta_i^2\frac{\sigma_i^2}{\sigma_Y^2}$$

mide la fracción de la varianza de Y atribuible a su regresión lineal con X_i. Esta fracción puede considerarse razonablemente como una medida de la importancia relativa de X_i.

Si se toma una muestra de esta población, las cantidades

$$b_i^2\ \frac{SC_{(X_i)}}{SC_{(Y)}}$$

son estimas de las fracciones anteriormente citadas.

Las raíces cuadradas de estas estimas son

$$b_i\ \sqrt{\frac{SC_{(X_i)}}{SC_{(Y)}}}$$

A esta expresión se le denomina *coeficientes de regresión parcial típico*, y se ha utilizado como medida de la importancia relativa, ignorando el signo. A la expresión

$$\sqrt{\frac{SC_{(X_i)}}{SC_{(Y)}}}$$

se le considera como un coeficiente de *corrección por escala.*

Pero en la práctica, las X_i están correlacionadas, lo que hace más difícil la respuesta a la pregunta de la importancia relativa de cada X. Una solución de compromiso puede ser estudiar la contribución de cada X a la Y por separado y juntas y tomar una decisión en base a la que salga mejor predictora en ambos casos.

Selección del modelo adecuado.-

Un problema que se plantea cuando se realiza una regresión múltiple es que, tal vez, la mayoría de las variables X elegidas pueden contribuir muy poco o nada a la precisión del pronóstico. Por ejemplo, se puede comenzar con 11 variables X y seleccionar tres de ellas que proveen un mejor pronóstico.

El problema en este caso es saber cuantas variable se necesitan y cuales son.

Un enfoque lógico para resolver este problemas sería el de elaborar la regresión de *Y* con todos los posibles subgrupos de las *k* variables *X*, es decir, realizar la regresión simple con todas la variables *X*, después todas las regresiones dobles con todas las combinaciones de dos de las variables *X*, después todas la regresiones triples, *etc.* El conjunto que tenga un menor cuadrado medio del *error* sería el escogido para hacer la predicción. Si hubiera dos subconjuntos con el mismo CM_{error}, se elegiría el de menos cantidad de variables *X*. Como es fácil ver, este método falla por la cantidad de cálculos que son necesarios. Si se tienen *k* variable independientes, *X*, habría que calcular 2^k-1 regresiones, es decir, si se tienen 11 variables *X* habría que calcular 2047 regresiones. Aún con un ordenador esto es muy complicado.

Hay varias maneras de resolver esta cuestión, se van ha estudiar aquí el denominado *método ascendente* (*forward*) de introducción de variables; el *método descendente* (*backward*) de eliminación de variables; el *método paso a paso* (*stepwise*); el *método CP* (*CP*) de Mallows de introducción de variables; el *método de mejora del máximo* R^2 y el *método de* R^2 *ajustado* de introducción de variables.

En el método descendente primero se calcula la regresión de *Y* en todas la *k* variables *X* y se elimina la variable que tenga una menor contribución a la reducción de la suma de cuadrados total de *Y*. Se ajusta los datos para dicha variable y se vuelve a calcular las regresiones simples para todas las demás variables y se elimina aquella que tenga una menor contribución a la suma de cuadrados total de *Y*. Y así sucesivamente. El límite esta en el nivel de significación del 0.1, es decir, solo se elimina variable mientras que su nivel de significación de la regresión sea mayor de 0.1. Las que den con un nivel de significación inferior a 0.1 se dejan en el modelo.

En el método ascendente, se comienza calculando todas la regresiones simples, las variable que de una mayor reducción de la suma de cuadrados de *Y* es la seleccionada. Después se calculan todas las regresiones bivariantes en la que aparezca la *X* seleccionada anteriormente; la pareja que dé una mayor reducción de la SC_y es la seleccionada. Después se calculan todas la regresiones triples en las que aparezca la pareja de *X* anteriormente seleccionada, y así sucesivamente.

El método de paso a paso es un perfeccionamiento del método ascendente consistente en que en cada paso se considera la inclusión o exclusión de las variables que se habían introducido en pasos anteriores. Una variable que fuera la mejor en un paso anterior queda definitivamente incluida en el modelo por el método ascendente, sin embargo puede ocurrir que esta variable sea superflua en una fase posterior debido a la relación existente entre dicha variable y variables que se han introducido posteriormente en el modelo, esto lo evalúa el método paso a paso por medio del estadístico *F*.

El método *CP* de *Mallows* consiste en buscar el modelo que más simple que tenga un cuadrado medio del error más pequeño. Esta medida denominada *Cp* tiene en cuenta la suma de las desviaciones al cuadrado respecto al modelo completo más el cuadrado de los residuos, en el conjunto de las *p* variables seleccionadas

$$Cp = \frac{SC_{E_p}}{S^2_{Y.X}} - N+2p$$

siendo SC_{Ep} la suma de cuadrados del error para el modelo con los p parámetros, incluyendo la ordenada en el origen, y $S^2_{Y.X}$ es el cuadrado medio del error para el modelo completo.

El método de la inclusión por el R^2_{adj} es como el anterior pero utilizando como criterio el coeficiente de determinación ajustado.

Y el método del máximo R^2 busca el modelo con una variable que tiene un mayor coeficiente de determinación, después el modelo con dos variables que tiene un mayor R^2, después el modelo con tres variables, y así sucesivamente.

Estos métodos no eligen las mismas variables ni ninguno elige las variables que elegiría el método lógico impracticable. Estas diferencias no son muy preocupantes pues la interrelación entre las X hace que diferentes subgrupos puedan proporcionar pronósticos semejantes.

Ejemplo.-

Pongamos el mismo ejemplo, reducido en el número de individuos, del manual del *SAS* para este tema.

Se esta estudiando la capacidad de consumir oxigeno en personas adultas. Como las medidas directas de este carácter es complicado y caro se pretende realizar un ajuste por medio de los resultados de algunos ejercicios físicos simples. El objetivo es desarrollar una ecuación que prediga la capacidad de consumo de oxigeno basada en ejercicios físicos para evitar la medida directa de consumo de oxigeno que son caras y lentas. Se usan los métodos de selección del modelo anteriormente descritos.

Las variable fueron medidas en 10 personas adultas. Estas fueron: ***edad*** de la persona en años; ***peso*** de la persona en *Kg*; tasa de inspiración de ***oxig*****eno** (*ml* por *Kg* de peso por minuto); ***tiempo*** en minutos en correr 2414 *m*; media de pulsaciones de la persona en reposo (***pulrep***); media de pulsaciones corriendo, a la vez que se mide el consumo de oxigeno (***pulcor***); y máxima de pulsaciones registradas corriendo (***pulmax***).

Archivo del programa *SAS* (C13-7.SAS).-

```
title 'Selección del modelo más adecuado';
option ls=80 ps=60;
data aerobic;
infile 'c13-7.dat';
input edad peso oxigeno tiempo pulrep pulcor pulmax;
proc reg;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=forward;
run;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=backward;
run;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=stepwise;
run;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=cp;
run;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=maxr;
run;
model oxigeno = edad peso tiempo pulrep pulcor pulmax / selection=adjrsq;
run;
```

Archivo de datos (C13-7.DAT).-

```
44  89.47  44.609  11.37  62 178  182
40  75.07  45.313  10.07  62 185  185
44  85.84  54.297   8.65  45 156  168
42  68.15  59.571   8.17  40 166  172
38  89.02  49.874   9.22  55 178  180
47  77.45  44.811  11.63  58 176  176
40  75.98  45.681  11.95  70 176  170
43  81.19  49.091  10.85  64 162  170
44  81.42  39.442  13.08  63 174  176
45  87.66  37.388  14.03  56 186  192
```

Archivo de resultados (C13-7.LST).-

```
          Forward Selection Procedure for Dependent Variable OXIGENO

Step 1   Variable TIEMPO Entered    R-square = 0.87387787   C(p) =  3.28968235

                 DF         Sum of Squares      Mean Square          F   Prob>F
 Regression       1           339.51558180     339.51558180      55.43   0.0001
 Error            8            49.00047230       6.12505904
 Total            9           388.51605410

                Parameter        Standard          Type II
 Variable        Estimate           Error   Sum of Squares          F   Prob>F
 INTERCEP      82.27974896      4.80178194    1798.42084181     293.62   0.0001
 TIEMPO        -3.23537415      0.43456005     339.51558180      55.43   0.0001

Bounds on condition number:            1,            1
--------------------------------------------------------------------------------

Step 2   Variable PULCOR Entered    R-square = 0.92183235   C(p) =  1.75753550

                 DF         Sum of Squares      Mean Square          F   Prob>F
 Regression       2           358.14666705     179.07333352      41.28   0.0001
 Error            7            30.36938705       4.33848386
 Total            9           388.51605410
```

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	107.85569192	12.98668253	299.24553356	68.97	0.0001
TIEMPO	-2.74000804	0.43692324	170.62056032	39.33	0.0004
PULCOR	-0.17833290	0.08605604	18.63108525	4.29	0.0770

Bounds on condition number: 1.427194, 5.708776

Step 3 Variable PESO Entered R-square = 0.94478260 C(p) = 2.06710635

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	367.06320734	122.35440245	34.22	0.0004
Error	6	21.45284676	3.57547446		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	118.76907226	13.66571699	270.06951440	75.53	0.0001
PESO	-0.14802253	0.09373385	8.91654029	2.49	0.1654
TIEMPO	-2.57601285	0.41001522	141.13339230	39.47	0.0008
PULCOR	-0.18232212	0.07816389	19.45360530	5.44	0.0584

Bounds on condition number: 1.525025, 12.11944

Step 4 Variable PULREP Entered R-square = 0.95266370 C(p) = 3.48661377

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	370.12514308	92.53128577	25.16	0.0016
Error	5	18.39091102	3.67818220		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	119.44654892	13.88048025	272.37764900	74.05	0.0003
PESO	-0.15120974	0.09513476	9.29210963	2.53	0.1728
TIEMPO	-2.34557106	0.48655179	85.48152677	23.24	0.0048
PULREP	-0.08459987	0.09272317	3.06193574	0.83	0.4034
PULCOR	-0.17119200	0.08021164	16.75426587	4.56	0.0859

Bounds on condition number: 2.087544, 25.47404

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable OXIGENO

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	TIEMPO	1	0.8739	0.8739	3.2897	55.4306	0.0001
2	PULCOR	2	0.0480	0.9218	1.7575	4.2944	0.0770
3	PESO	3	0.0230	0.9448	2.0671	2.4938	0.1654
4	PULREP	4	0.0079	0.9527	3.4866	0.8325	0.4034

Backward Elimination Procedure for Dependent Variable OXIGENO

Step 0 All Variables Entered R-square = 0.95927026 C(p) = 7.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	372.69189441	62.11531574	11.78	0.0342
Error	3	15.82415969	5.27471990		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	137.12659119	32.14003439	96.01743237	18.20	0.0236
EDAD	-0.20252295	0.44901634	1.07305898	0.20	0.6826

PESO	-0.11362829	0.14893752	3.07018213	0.58	0.5010
TIEMPO	-1.95112749	0.89211597	25.23062885	4.78	0.1166
PULREP	-0.16805599	0.16553483	5.43661543	1.03	0.3848
PULCOR	-0.09383318	0.27332288	0.62167052	0.12	0.7540
PULMAX	-0.14127517	0.32411723	1.00213634	0.19	0.6924

Bounds on condition number: 11.84161, 213.0076

Step 1 Variable PULCOR Removed R-square = 0.95767014 C(p) = 5.11785849

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	372.07022390	74.41404478	18.10	0.0075
Error	4	16.44583020	4.11145755		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	135.92194162	28.20594443	95.47587741	23.22	0.0085
EDAD	-0.14374608	0.36647548	0.63255459	0.15	0.7149
PESO	-0.08276987	0.10484593	2.56234264	0.62	0.4740
TIEMPO	-2.00579151	0.77497797	27.54156896	6.70	0.0608
PULREP	-0.19473939	0.12903425	9.36469039	2.28	0.2057
PULMAX	-0.24278331	0.11721255	17.63945761	4.29	0.1071

Bounds on condition number: 4.737984, 64.22518

Step 2 Variable EDAD Removed R-square = 0.95604201 C(p) = 3.23778042

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	371.43766931	92.85941733	27.19	0.0014
Error	5	17.07838479	3.41567696		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	127.22907403	15.90277711	218.62689105	64.01	0.0005
PESO	-0.08597413	0.09527298	2.78146021	0.81	0.4082
TIEMPO	-2.22563892	0.48781606	71.10072181	20.82	0.0060
PULREP	-0.16253522	0.09073036	10.96141095	3.21	0.1332
PULMAX	-0.22381159	0.09731515	18.06679210	5.29	0.0698

Bounds on condition number: 2.259676, 26.74121

Step 3 Variable PESO Removed R-square = 0.94888282 C(p) = 1.76509946

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	368.65620909	122.88540303	37.13	0.0003
Error	6	19.85984501	3.30997417		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	125.02419024	15.46889692	216.21907955	65.32	0.0002
TIEMPO	-2.26285254	0.47848969	74.02730700	22.36	0.0032
PULREP	-0.16450581	0.08928957	11.23532447	3.39	0.1150
PULMAX	-0.24781360	0.09214991	23.93784435	7.23	0.0361

Bounds on condition number: 2.243527, 16.15205

Step 4 Variable PULREP Removed R-square = 0.91996426 C(p) = 1.89513189

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	357.42088462	178.71044231	40.23	0.0001
Error	7	31.09516948	4.44216707		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	114.88752962	16.74844626	209.02181866	47.05	0.0002
TIEMPO	-2.83934022	0.41936677	203.62997957	45.84	0.0003
PULMAX	-0.20849996	0.10385145	17.90530282	4.03	0.0847

Bounds on condition number: 1.284115, 5.136459

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable OXIGENO

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	PULCOR	5	0.0016	0.9577	5.1179	0.1179	0.7540
2	EDAD	4	0.0016	0.9560	3.2378	0.1539	0.7149
3	PESO	3	0.0072	0.9489	1.7651	0.8143	0.4082
4	PULREP	2	0.0289	0.9200	1.8951	3.3944	0.1150

Stepwise Procedure for Dependent Variable OXIGENO

Step 1 Variable TIEMPO Entered R-square = 0.87387787 C(p) = 3.28968235

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	339.51558180	339.51558180	55.43	0.0001
Error	8	49.00047230	6.12505904		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	82.27974896	4.80178194	1798.42084181	293.62	0.0001
TIEMPO	-3.23537415	0.43456005	339.51558180	55.43	0.0001

Bounds on condition number: 1, 1

Step 2 Variable PULCOR Entered R-square = 0.92183235 C(p) = 1.75753550

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	358.14666705	179.07333352	41.28	0.0001
Error	7	30.36938705	4.33848386		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	107.85569192	12.98668253	299.24553356	68.97	0.0001
TIEMPO	-2.74000804	0.43692324	170.62056032	39.33	0.0004
PULCOR	-0.17833290	0.08605604	18.63108525	4.29	0.0770

Bounds on condition number: 1.427194, 5.708776

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable OXIGENO

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	TIEMPO		1	0.8739	0.8739	3.2897	55.4306	0.0001
2	PULCOR		2	0.0480	0.9218	1.7575	4.2944	0.0770

N = 10 Regression Models for Dependent Variable: OXIGENO

C(p)	R-square	In	Variables in Model
1.75754	0.92183235	2	TIEMPO PULCOR
1.76510	0.94888282	3	TIEMPO PULREP PULMAX
1.89513	0.91996426	2	TIEMPO PULMAX
2.06711	0.94478260	3	PESO TIEMPO PULCOR
3.23778	0.95604201	4	PESO TIEMPO PULREP PULMAX
3.24824	0.92874678	3	TIEMPO PULREP PULCOR
3.28968	0.87387787	1	TIEMPO
3.31588	0.92782848	3	PESO TIEMPO PULMAX
3.46502	0.92580369	3	TIEMPO PULCOR PULMAX
3.48661	0.95266370	4	PESO TIEMPO PULREP PULCOR
3.51661	0.92510326	3	EDAD TIEMPO PULMAX
3.60364	0.95107494	4	EDAD TIEMPO PULREP PULMAX
3.68714	0.94994130	4	TIEMPO PULREP PULCOR PULMAX
3.75519	0.89471104	2	PESO TIEMPO
3.75686	0.92184146	3	EDAD TIEMPO PULCOR
4.04329	0.94510594	4	PESO TIEMPO PULCOR PULMAX
4.06678	0.94478701	4	EDAD PESO TIEMPO PULCOR
4.30332	0.88726929	2	TIEMPO PULREP
4.34511	0.88670190	2	EDAD TIEMPO
4.66295	0.90953996	3	PESO TIEMPO PULREP
4.84105	0.90712195	3	EDAD PESO TIEMPO
4.89325	0.93356640	4	EDAD PESO TIEMPO PULMAX
5.04046	0.93156780	4	EDAD TIEMPO PULREP PULCOR
5.11786	0.95767014	5	EDAD PESO TIEMPO PULREP PULMAX
5.18999	0.95669086	5	EDAD PESO TIEMPO PULREP PULCOR
5.20343	0.95650831	5	PESO TIEMPO PULREP PULCOR PULMAX
5.38810	0.92684796	4	EDAD TIEMPO PULCOR PULMAX
5.58206	0.95136793	5	EDAD TIEMPO PULREP PULCOR PULMAX
6.03069	0.94527697	5	EDAD PESO TIEMPO PULCOR PULMAX
6.05014	0.89070666	3	EDAD TIEMPO PULREP
6.46201	0.91226801	4	EDAD PESO TIEMPO PULREP
6.83098	0.88010543	3	EDAD PULREP PULMAX
7.00000	0.95927026	6	EDAD PESO TIEMPO PULREP PULCOR PULMAX
8.13861	0.88950554	4	EDAD PESO PULREP PULCOR
8.33929	0.88678100	4	EDAD PESO PULREP PULMAX
8.79880	0.88054235	4	EDAD PULREP PULCOR PULMAX
9.56959	0.84292455	3	EDAD PULREP PULCOR
9.78331	0.89432924	5	EDAD PESO PULREP PULCOR PULMAX
13.79946	0.75834422	2	PULREP PULMAX
14.71731	0.77303613	3	PESO PULREP PULMAX
14.93947	0.77001992	3	PULREP PULCOR PULMAX
16.56942	0.77504389	4	PESO PULREP PULCOR PULMAX
16.79567	0.77197221	4	EDAD PESO PULCOR PULMAX
17.69250	0.73264313	3	PESO PULREP PULCOR
18.77757	0.71791168	3	EDAD PESO PULCOR
21.17120	0.65826115	2	PULREP PULCOR
21.52084	0.68066744	3	EDAD PESO PULREP
21.88612	0.64855502	2	EDAD PULCOR
23.48489	0.65400232	3	EDAD PULCOR PULMAX
24.30693	0.61568861	2	EDAD PULREP
26.82367	0.58151990	2	PESO PULCOR
27.97639	0.59302315	3	PESO PULCOR PULMAX
28.29544	0.56153833	2	PESO PULREP
32.10438	0.48267274	1	PULCOR
33.00700	0.47041832	1	PULREP
33.79484	0.48687533	2	PULCOR PULMAX
34.92212	0.47157064	2	EDAD PULMAX
35.12273	0.49600019	3	EDAD PESO PULMAX
38.02544	0.42943816	2	PESO PULMAX
38.50002	0.39584183	1	PULMAX
52.86815	0.22792487	2	EDAD PESO
55.86098	0.16013935	1	PESO
60.90080	0.09171580	1	EDAD

Maximum R-square Improvement for Dependent Variable OXIGENO

Step 1 Variable TIEMPO Entered R-square = 0.87387787 C(p) = 3.28968235

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	339.51558180	339.51558180	55.43	0.0001
Error	8	49.00047230	6.12505904		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	82.27974896	4.80178194	1798.42084181	293.62	0.0001
TIEMPO	-3.23537415	0.43456005	339.51558180	55.43	0.0001

Bounds on condition number: 1, 1

The above model is the best 1-variable model found.

Step 2 Variable PULCOR Entered R-square = 0.92183235 C(p) = 1.75753550

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	358.14666705	179.07333352	41.28	0.0001
Error	7	30.36938705	4.33848386		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	107.85569192	12.98668253	299.24553356	68.97	0.0001
TIEMPO	-2.74000804	0.43692324	170.62056032	39.33	0.0004
PULCOR	-0.17833290	0.08605604	18.63108525	4.29	0.0770

Bounds on condition number: 1.427194, 5.708776

The above model is the best 2-variable model found.

Step 3 Variable PESO Entered R-square = 0.94478260 C(p) = 2.06710635

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	367.06320734	122.35440245	34.22	0.0004
Error	6	21.45284676	3.57547446		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	118.76907226	13.66571699	270.06951440	75.53	0.0001
PESO	-0.14802253	0.09373385	8.91654029	2.49	0.1654
TIEMPO	-2.57601285	0.41001522	141.13339230	39.47	0.0008
PULCOR	-0.18232212	0.07816389	19.45360530	5.44	0.0584

Bounds on condition number: 1.525025, 12.11944

The above model is the best 3-variable model found.

Step 4 Variable PULREP Entered R-square = 0.95266370 C(p) = 3.48661377

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	370.12514308	92.53128577	25.16	0.0016
Error	5	18.39091102	3.67818220		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F

INTERCEP	119.44654892	13.88048025	272.37764900	74.05	0.0003
PESO	-0.15120974	0.09513476	9.29210963	2.53	0.1728
TIEMPO	-2.34557106	0.48655179	85.48152677	23.24	0.0048
PULREP	-0.08459987	0.09272317	3.06193574	0.83	0.4034
PULCOR	-0.17119200	0.08021164	16.75426587	4.56	0.0859

Bounds on condition number: 2.087544, 25.47404

Step 5 Variable PULCOR Removed R-square = 0.95604201 C(p) = 3.23778042
Variable PULMAX Entered

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	371.43766931	92.85941733	27.19	0.0014
Error	5	17.07838479	3.41567696		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	127.22907403	15.90277711	218.62689105	64.01	0.0005
PESO	-0.08597413	0.09527298	2.78146021	0.81	0.4082
TIEMPO	-2.22563892	0.48781606	71.10072181	20.82	0.0060
PULREP	-0.16253522	0.09073036	10.96141095	3.21	0.1332
PULMAX	-0.22381159	0.09731515	18.06679210	5.29	0.0698

Bounds on condition number: 2.259676, 26.74121

The above model is the best 4-variable model found.

Step 6 Variable EDAD Entered R-square = 0.95767014 C(p) = 5.11785849

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	372.07022390	74.41404478	18.10	0.0075
Error	4	16.44583020	4.11145755		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	135.92194162	28.20594443	95.47587741	23.22	0.0085
EDAD	-0.14374608	0.36647548	0.63255459	0.15	0.7149
PESO	-0.08276987	0.10484593	2.56234264	0.62	0.4740
TIEMPO	-2.00579151	0.77497797	27.54156896	6.70	0.0608
PULREP	-0.19473939	0.12903425	9.36469039	2.28	0.2057
PULMAX	-0.24278331	0.11721255	17.63945761	4.29	0.1071

Bounds on condition number: 4.737984, 64.22518

The above model is the best 5-variable model found.

Step 7 Variable PULCOR Entered R-square = 0.95927026 C(p) = 7.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	372.69189441	62.11531574	11.78	0.0342
Error	3	15.82415969	5.27471990		
Total	9	388.51605410			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	137.12659119	32.14003439	96.01743237	18.20	0.0236
EDAD	-0.20252295	0.44901634	1.07305898	0.20	0.6826
PESO	-0.11362829	0.14893752	3.07018213	0.58	0.5010
TIEMPO	-1.95112749	0.89211597	25.23062885	4.78	0.1166

PULREP	-0.16805599	0.16553483	5.43661543	1.03	0.3848
PULCOR	-0.09383318	0.27332288	0.62167052	0.12	0.7540
PULMAX	-0.14127517	0.32411723	1.00213634	0.19	0.6924

Bounds on condition number: 11.84161, 213.0076

The above model is the best 6-variable model found.

No further improvement in R-square is possible.

N = 10 Regression Models for Dependent Variable: OXIGENO

Adjusted R-square	R-square	In	Variables in Model
0.92332423	0.94888282	3	TIEMPO PULREP PULMAX
0.92087562	0.95604201	4	PESO TIEMPO PULREP PULMAX
0.91717390	0.94478260	3	PESO TIEMPO PULCOR
0.91479467	0.95266370	4	PESO TIEMPO PULREP PULCOR
0.91193488	0.95107494	4	EDAD TIEMPO PULREP PULMAX
0.90989435	0.94994130	4	TIEMPO PULREP PULCOR PULMAX
0.90475781	0.95767014	5	EDAD PESO TIEMPO PULREP PULMAX
0.90255444	0.95669086	5	EDAD PESO TIEMPO PULREP PULCOR
0.90214370	0.95650831	5	PESO TIEMPO PULREP PULCOR PULMAX
0.90119069	0.94510594	4	PESO TIEMPO PULCOR PULMAX
0.90061661	0.94478701	4	EDAD PESO TIEMPO PULCOR
0.89949874	0.92183235	2	TIEMPO PULCOR
0.89709691	0.91996426	2	TIEMPO PULMAX
0.89312017	0.92874678	3	TIEMPO PULREP PULCOR
0.89174271	0.92782848	3	PESO TIEMPO PULMAX
0.89057783	0.95136793	5	EDAD TIEMPO PULREP PULCOR PULMAX
0.88870554	0.92580369	3	TIEMPO PULCOR PULMAX
0.88765490	0.92510326	3	EDAD TIEMPO PULMAX
0.88276219	0.92184146	3	EDAD TIEMPO PULCOR
0.88041952	0.93356640	4	EDAD PESO TIEMPO PULMAX
0.87781077	0.95927026	6	EDAD PESO TIEMPO PULREP PULCOR PULMAX
0.87687319	0.94527697	5	EDAD PESO TIEMPO PULCOR PULMAX
0.87682204	0.93156780	4	EDAD TIEMPO PULREP PULCOR
0.86832633	0.92684796	4	EDAD TIEMPO PULCOR PULMAX
0.86462848	0.89471104	2	PESO TIEMPO
0.86430994	0.90953996	3	PESO TIEMPO PULREP
0.86068292	0.90712195	3	EDAD PESO TIEMPO
0.85811260	0.87387787	1	TIEMPO
0.85506052	0.88726929	2	TIEMPO PULREP
0.85433101	0.88670190	2	EDAD TIEMPO
0.84208241	0.91226801	4	EDAD PESO TIEMPO PULREP
0.83606000	0.89070666	3	EDAD TIEMPO PULREP
0.82015814	0.88010543	3	EDAD PULREP PULMAX
0.80110997	0.88950554	4	EDAD PESO PULREP PULCOR
0.79620580	0.88678100	4	EDAD PESO PULREP PULMAX
0.78497623	0.88054235	4	EDAD PULREP PULCOR PULMAX
0.76438683	0.84292455	3	EDAD PULREP PULCOR
0.76224078	0.89432924	5	EDAD PESO PULREP PULCOR PULMAX
0.68929971	0.75834422	2	PULREP PULMAX
0.65955420	0.77303613	3	PESO PULREP PULMAX
0.65502987	0.77001992	3	PULREP PULCOR PULMAX
0.59896469	0.73264313	3	PESO PULREP PULCOR
0.59507901	0.77504389	4	PESO PULREP PULCOR PULMAX
0.58954998	0.77197221	4	EDAD PESO PULCOR PULMAX
0.57686752	0.71791168	3	EDAD PESO PULCOR
0.56062148	0.65826115	2	PULREP PULCOR
0.54814217	0.64855502	2	EDAD PULCOR
0.52100116	0.68066744	3	EDAD PESO PULREP
0.50588536	0.61568861	2	EDAD PULREP
0.48100349	0.65400232	3	EDAD PULCOR PULMAX
0.46195415	0.58151990	2	PESO PULCOR
0.43626357	0.56153833	2	PESO PULREP

```
0.41800683   0.48267274    1   PULCOR
0.40422061   0.47041832    1   PULREP
0.38953473   0.59302315    3   PESO PULCOR PULMAX
0.34026828   0.48687533    2   PULCOR PULMAX
0.32059082   0.47157064    2   EDAD PULMAX
0.32032206   0.39584183    1   PULMAX
0.26642049   0.42943816    2   PESO PULMAX
0.24400029   0.49600019    3   EDAD PESO PULMAX
0.05515676   0.16013935    1   PESO
0.00733198   0.22792487    2   EDAD PESO
-.02181972   0.09171580    1   EDAD
-----------------------------------
```

Como se ve, el método ascendente (*forward*) en el primer paso (*step 1*) ha elegido el *tiempo* como la variable que propicia una mayor reducción de la suma de cuadrados total de *oxigeno*. Después, en el segundo paso, de todas las combinaciones de *tiempo* con las demás variables independientes, es la combinación *tiempo, pulcor* la que propicia una mayor reducción de la suma de cuadrados total. En el paso tercero, el grupo de tres variables independientes de entre las combinaciones que lleve *tiempo, pulcor* el el trío *tiempo, pulcor, peso* el que propicia una mayor reducción de la suma de cuadrados total. En el paso cuarto, el grupo de cuatro variables (que incluya *tiempo, pulcor, peso*) que propicia una mayor reducción de la suma de cuadrados total es *tiempo, pulcor, peso, pulrep.*

Ya no elige mas variables, pues las demás (*pulmax, edad*) tienen un nivel de significación del 0.5, es decir, azar total, por lo que se supone que, al menos en esta muestra, las demás variables no indican nada con respecto a la capacidad de consumo de oxigeno.

Por contra, el método descendente (*backward*) primeramente (paso cero) ha calculado todas las regresiones simples y la que tiene una menor contribución a la reducción (*Type II Sum of Squares*) es la eliminada, esta es *pulcor*. En el primer paso se hace lo mismo con todas las demás después de ajustados los datos a *pulcor* y dá que la que tiene una menor reducción es *edad*, por lo que es eliminada. En el paso segundo se ajusta para *edad* (ya están ajustados para *pulcor*) y se repite el proceso, la variable que tiene una menor reducción de *peso*, por lo que es eliminada. En el paso tercero es la variable *pulrep* la que provee una menor reducción, por lo que es eliminada. Se ajusta los datos también para esta variable y se calcula la regresión simple para las demás, no dando ninguna un nivel de significación superior a 0.11, por lo que estas variables restantes son las que tendrán el modelo. Es decir, el modelo lo formarán las variables *tiempo y pulmax* además de la ordenada en el origen.

Con el método descendente queda un modelo más simple que con el método ascendente y, además, con una variable, *pulmax*, que no incluía el método ascendente.

Con el método paso a paso introduce en el modelo *tiempo* y *pulcor*, no ha sacado ninguna variable después de meterla.

Con el método de *Cp* se concluye también que el modelo con *tiempo* y *pulcor* es el más óptimo.

Con respecto al método de R^2, no encuentra variables que se pueda eliminar. Y el método de R^2_{adj} indica que el modelo con las variables *tiempo, pulrep y pulmax* es el que tiene un mayor coeficiente de determinación ajustado, mientras que el coeficiente de determinación normal, el más grande, lógicamente, es el del modelo con todas las variables, si bien el modelo con las tres variables anteriormente citadas tiene un coeficiente de determinación ligeramente menor (una centésima) que el modelo completo.

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

Freund, R.J., and Littel, R.C. 1991. *SAS* SYSTEM FOR REGRESION. SAS Institute Inc., Cary, NC, USA.

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

Lite, TM, y Jackson Hills, F. 1987. METODOS ESTADISTICOS PARA LA INVESTIGACION EN LA AGRICULTURA. Ed TRILLAS. México.

Littel, R.C., Freund, R.J. and Spector, P.C. 1991. *SAS* FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. METODOS ESTADISTICOS. Ed. C.E.C.S.A. México.

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

SAS Institute Inc. 1990. *SAS* PROCEDURE GUIDE. Cary, NC, USA.

CAPÍTULO 14

Regresión Curvilínea

CAPÍTULO 14

Regresión Curvilínea

Ajustes de curvas.-

El tipo más común y sencillo de ajuste de curvas es el de la línea recta. Sin embargo, cuando se representan pares de observaciones, éstas suelen quedar sobre una línea curva; y hay observaciones para las que la teoría exige el ajuste a una curva de forma específica.

Regresión no lineal.-

Una relación entre dos variables puede ser aproximadamente lineal cuando se estudia en un intervalo limitado, pero puede ser marcadamente curvilínea si se amplía el intervalo. Por ejemplo, la relación entre la maduración y el rendimiento para envasado en guisantes, usualmente es una recta en el intervalo de grado de maduración aceptable para la industria conservera. Pero al aumentar el grado de maduración el rendimiento para envasado disminuye, es decir, se hace curvilínea. Análogamente, la tasa de aumento en rendimiento tiende a mejorar en la etapas de inmadurez. Así, pues, para describir la relación en todo el intervalo es inadecuada la ecuación de una recta.

Además, si se usa una recta para describir un proceso curvilíneo se está sobrecargando sobre el error la componente de la regresión curvilínea que no se está teniendo en cuenta. Así, si unas observaciones se describen apropiadamente con la ecuación

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

y se usa como modelo:

$$Y = \alpha + \beta_1 X + \varepsilon$$

entonces se asigna a la medida del error la parte de la variación correspondiente a $\beta_2 X^2$, por lo que se está sobreestimando la medida del error.

La selección de la forma de la curva o de la ecuación de regresión que mejor describa una relación curvilínea no es fácil, pues es prácticamente infinito el número de ecuaciones que se pueden tomar como buenas para minimizar la $SC_{(residuo)}$. Por lo que es deseable tener una teoría previa de cuál es la curva a la que se ajusta las observaciones.

Las relaciones curvilíneas se pueden clasificar en dos tipos: *lineales* y *no lineales*. Los modelos lineales son aquellos para los cuales se dispone de Las técnicas de regresión. Los modelos que no son lineales pueden subdividirse a su vez en los que se pueden linealizar por medio de una transformación y los que no se puede linealizar.

La transformación tiene por objeto proporcionar un procedimiento más fácil de ajuste y procedimientos válidos de estimación y prueba. Por ejemplo, se puede convenir en que la ecuación que mejor describe unos datos determinados es

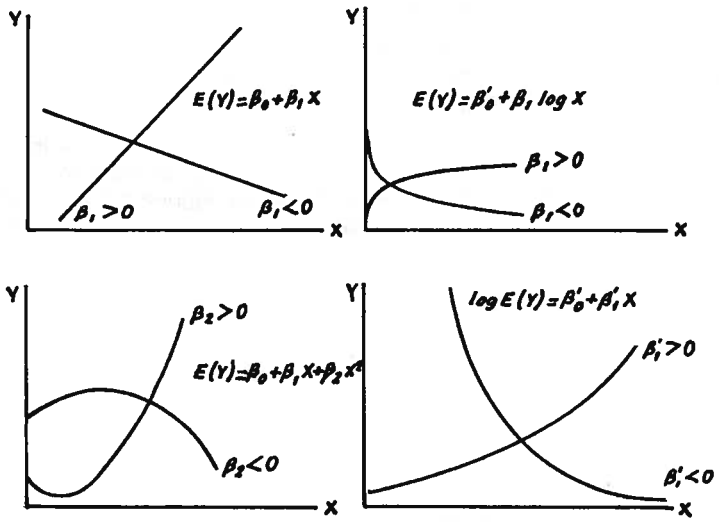
$$Y = \alpha\, X^{\beta}$$

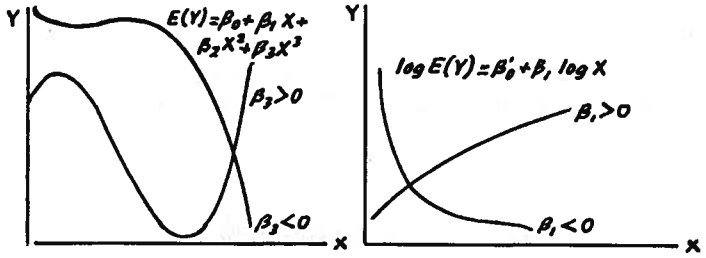
Esta convicción procede de los conocimientos previos del problema o de una teoría previa ya contrastada. Entonces, en lugar de trabajar con esta ecuación, se puede trabajar con esta otra

$$\log Y = \log \alpha + \beta \log X$$

que es una ecuación lineal si el par de observaciones que se consideran son **log*Y*** y **log*X***. En este caso son perfectamente aplicables los procedimientos estudiados para la regresión lineal simple.

Por tanto, hay dos tipos generales de curvas: las polinomiales (no linealizables) y las logarítmicas (linealizables).





Las polinomiales pueden ser

$$\begin{array}{ll} \textit{Lineal} & Y = \alpha + \beta_1 X \\ \textit{Cuadrática} & Y = \alpha + \beta_1 X + \beta_2 X^2 \\ \textit{Cubica} & Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \end{array}$$

.....

Las logarítmicas

$$\begin{array}{ll} \exp^Y = \alpha X^{\beta} & Y = \log\alpha + \beta \log X \\ Y = \alpha \beta^X & \log Y = \log\alpha + \log\beta X \\ Y = \alpha X^{\beta} & \log Y = \log\alpha + \beta \log X \end{array}$$

Para las ecuaciones exponenciales, *Exp* puede ser cualquier constante, sin que esta afecte a la forma de la curva.

Los polinomios pueden tener picos y depresiones cuyo número, como máximo, es uno menos que el exponente más alto.

Curvas logarítmicas.-

Para determinar si una curva logarítmica puede describir nuestros datos, suele ser suficiente con representar los datos en papel logarítmico o semilogarítmico. Una vez tomada la decisión respecto al tipo de curva, se transforman los valores observados de *X* o de *Y* o de ambas, a logaritmos antes de realizar los cálculos. Los datos transformados se tratarán por los métodos ya estudiados. Los supuestos se aplican a los datos transformados en lugar de a los originales.

Curvas de crecimiento exponencial.-

Una característica de algunos fenómenos sencillos de crecimiento es que el aumento en cualquier momento de una variable es proporcional al tamaño ya alcanzado. En el crecimiento de un cultivo bacteriano, el número total de bacterias en un momento determinado se ajusta a esa ley. La relación queda bien ilustrada por el peso seco de embriones de pollo de 6 a 16 días del siguiente ejemplo.

Ejemplo.-

Se tiene el peso seco de embriones de pollo de 6 a 16 días de edad y se quiere saber si el aumento de peso con la edad se ajusta a una exponencial tipo

$$Y = \alpha \beta^X$$

Por tanto, lo que se tiene que comprobar es si se ajusta a la siguiente recta

$$\log Y = \log \alpha + X \log \beta$$

Edad en días X	*Peso seco* Y	*Logaritmos del peso* $\log Y$
6	0.029	-1.538
7	0.052	-1.284
8	0.079	-1.102
9	0.125	-0.903
10	0.181	-0.742
11	0.261	-0.583
12	0.425	-0.372
13	0.738	-0.132
14	1.130	0.053
15	1.882	0.275
16	2.812	0.449

$$\Sigma X = 121$$

$$\Sigma \log Y = -5.879$$

$$\Sigma \log YX = -43.12$$

$$SP = -43.121 - \frac{121 \times -5.879}{11} = 21.548$$

$$SC_{(X)} = 1441 - \frac{121^2}{11} = 110$$

$$b = \frac{21.548}{110} = 0.19589$$

$$a = -0.53455 - 0.19589 \times 11 = -2.68934$$

Por lo que la línea de regresión es

$$\log Y = -2.68934 + 0.19589\, X$$

Archivo del programa *SAS* (C14-1.SAS).-

```
title 'Curva exponencial';
optiones ls=75 ps=60;
data regres;
infile 'c14-1.dat';
input dias peso;
logpeso = log10(peso);
proc reg;
   model logpeso = dias;
run;
```

Archivo de datos (C14-1.DAT).-

```
 6       0.029
 7       0.052
 8       0.079
 9       0.125
10       0.181
11       0.261
12       0.425
13       0.738
14       1.130
15       1.882
16       2.812
```

Archivo de resultados (C14-1.LST).-

```
                          Curva exponencial

                          Sum of          Mean
Source                DF  Squares        Square     F Value      Prob>F
Model                  1  4.22063       4.22063    5384.937      0.0001
Error                  9  0.00705       0.00078
C Total               10  4.22768

    Root MSE          0.02800     R-square       0.9983
    Dep Mean         -0.53451     Adj R-sq       0.9981
    C.V.             -5.23775

                     Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|
INTERCEP   1       -2.689198     0.03055186       -88.021        0.0001
DIAS       1        0.195881     0.00266933        73.382        0.0001
```

Como se ve, el ajuste es muy bueno, por lo que se puede concluir que los embriones de pollo, cuando se miden por peso seco, crecen de acuerdo con una curva exponencial.

La metodología utilizada ha sido la misma estudiada más arriba, es decir, el de ajuste a una línea recta. Y ésta es la metodología para cualquier curva que pueda hacerse lineal por medio de logaritmos.

Si no se sabe, a priori, cual puede ser el tipo de curva, antes realizar los cálculos de ajuste se pueden representar los datos, con fines prácticos (no estéticos), con el mismo *SAS Estadístico,* para visualizar el tipo de curva.

Archivo del programa *SAS* (C14-2.SAS).-

```
title 'Representación de curva exponencial';
optiones ls=65 ps=30;
data regres;
infile 'c14-1.dat';
input dias peso;
proc plot;
plot peso * dias;
run;
```

Archivo de resultados (C14-2.LST).-

```
                Representación de curva exponencial

     Plot of PESO*DIAS.  Legend: A = 1 obs, B = 2 obs, etc.

     3 +                                                          A
       |
       |
       |
       |
       |
     2 +                                                    A
       |
       |
PESO   |
       |
       |                                                A
     1 +
       |
       |                                          A
       |                                    A
       |                              A
       |                        A     A
     0 +  A     A     A
       ---+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+--
          6     7     8     9     10    11    12    13    14    15    16
                                      DIAS
```

Se observa perfectamente que sigue una forma exponencial.

Crecimiento alométrico.-

Si se desea comparar los tamaños relativos de dos partes de un organismo, X e Y, una manera de hacerlo es mediante la ecuación

$$Y = a\,X^{b}$$

Si b es igual a cero, Y es siempre igual a a sea cual sea el valor de X. Por ejemplo, el tamaño de las células no difieren mucho si se comparan animales

corpulentos con animales pequeños. Si le damos a Y los valores de los diámetros de las células y a X las longitud global del cuerpo de los animales, se puede postular que b=0 y la ecuación anterior puede describir la relación entre X e Y bastante bien; Y sería igual a la constante a, para cualquier longitud corporal del animal.

Si ahora se pone en Y la extensión total de los brazos y en X la altura de los seres humanos adultos, La ecuación que mejor se ajustaría a los datos sería la anterior pero para b=1, puesto que en este caso la extensión de los brazos sería directamente proporcional a la altura.

Esta ecuación que relaciona X con Y puede describirse también de la forma ya conocida

$$\log Y = \log a + b \log X$$

Por tanto b será una constante en cada caso, esto es como consecuencia de que el cociente entre el incremento de estructuras de diferentes tamaños permanece aproximadamente constante, produciéndose un incremento relativamente grande de una variable con respecto a la otra en una escala lineal. Por ejemplo, el crecimiento de las astas de los ciervos en relación al tamaño del cuerpo sigue una relación alométrica.

Ejemplo.-

Se tienen mediciones de cráneos de diferentes tamaños. Se toman dos medidas concretas que forman parte de un índice. Se desea saber si siguen una relación alométrica. Los datos son

X	Y	log X	log Y
1	3.5	0.0000	0.5441
2	11.8	0.6931	1.0719
3	23.9	0.4771	1.3784
4	39.6	0.6021	1.5977
5	58.5	0.6990	1.7672
6	80.5	0.7781	1.9058
7	105.4	0.8451	2.0228
8	133.0	0.9031	2.1238
9	163.7	0.9542	2.2140
10	196.8	1.0000	2.2940
20	662.0	1.3010	2.8209
30	1345.9	1.4771	3.1290
40	2226.8	1.6021	3.3477
50	3290.5	1.6990	3.5173
60	4527.2	1.7781	3.6558
70	5929.1	1.8451	3.7730
80	7489.9	1.9031	3.8745
90	9204.3	1.9542	3.9640

$$\Sigma \log X = 20.1195$$
$$\Sigma \log Y = 45.0012$$
$$\Sigma \log Y \log X = 60.9066$$
$$SP = 60.9066 - \frac{20.1195 \times 45.0012}{18} = 10.6065$$
$$SC_{(\log X)} = 28.4598 - \frac{20.1195^2}{18} = 6.0612$$
$$b = \frac{10.6065}{6.0612} = 1.7499$$
$$a = 2.5 - 1.7499 \times 1.1177 = 0.5440$$

Por lo que la línea de regresión es

$$\log Y = 0.5440 + 0.17499 \log X$$

Archivo del programa *SAS* (C14-3.SAS).-

```
title 'Crecimiento alometrico';
option ls=75 ps=60;
data alometri;
infile 'c14-3.dat';
input X Y;
logx=log10(x);
logy=log10(Y);
proc reg;
  model logy=logx;
run;
```

Archivo de datos (C14-3.DAT).-

```
 1     3.5
 2    11.8
 3    23.9
 4    39.6
 5    58.5
 6    80.5
 7   105.4
 8   133.0
 9   163.7
10   196.8
20   662.0
30  1345.9
40  2226.8
50  3290.5
60  4527.2
70  5929.1
80  7478.9
90  9204.3
```

Archivo de resultados (C14-3.LST).-

```
                        Crecimiento alometrico

                                 Sum of          Mean
Source                    DF     Squares        Square      F Value       Prob>F
Model                      1    18.55997      18.55997 141432032.21       0.0001
Error                     16 2.0996623E-6  1.312289E-7
C Total                   17    18.55997

    Root MSE          0.00036     R-square        1.0000
    Dep Mean          2.50007     Adj R-sq        1.0000
    C.V.              0.01449

                    Parameter       Standard    T for H0:
 Variable  DF       Estimate          Error   Parameter=0    Prob > |T|
 INTERCEP   1       0.544135    0.00018531      2936.337        0.0001
 LOGX       1       1.749882    0.00014714     11892.520        0.0001
```

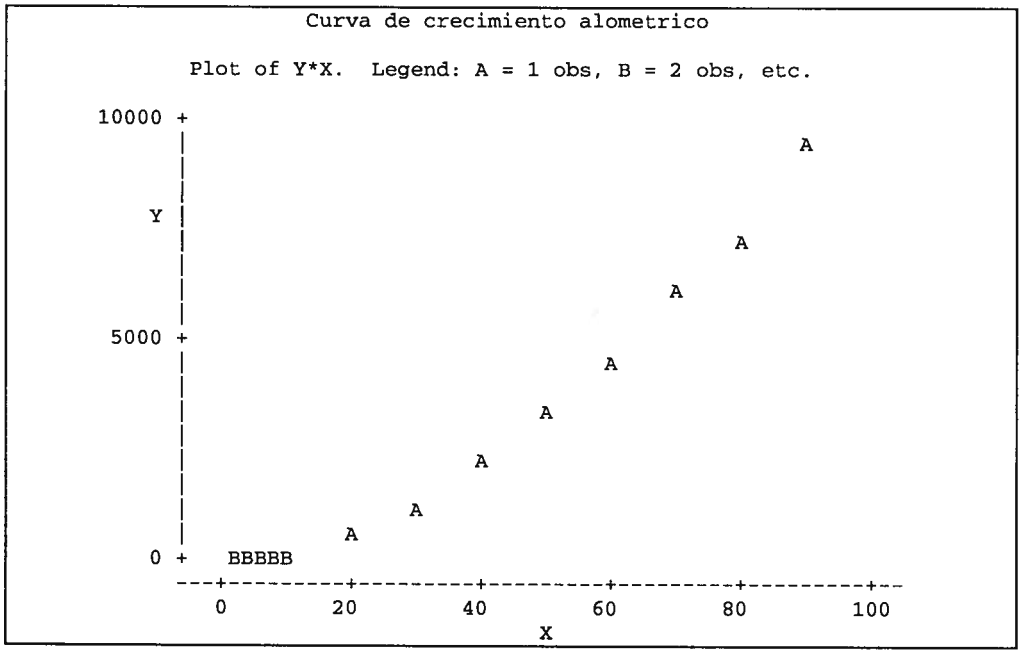
La pendiente es b =1.75, es decir, el crecimiento de Y se vería como el crecimiento de X elevado a 1.75.

Si, al igual que en el anterior ejemplo, se desea comprobar, previamente al ajuste de los datos, la línea que describen la nube de punto, se puede hacer con un programa *SAS* estadístico (no gráfico) de la siguiente manera.

Archivo del programa *SAS* (C14-4.SAS).-

```
title 'Curva de crecimiento alometrico';
option ls=75 ps=30;
data alometri;
infile 'c14-3.dat';
input X Y;
proc plot;
plot Y * X;
run;
```

Archivo de resultados (C14-4.LST).-



Curvas polinómicas.-

Podría ocurrir que nuestros datos no se ajustaran bien a una recta, por lo que tendríamos que rastrear si se ajustan a cualquiera de las curvas logarítmicas. Si el resultado fuera negativo tendríamos que probar con las llamadas curvas polinómicas cuya ecuación general es .

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + ... + \beta_n X^n + \varepsilon$$

Si la ecuación a la que se ajusta nuestros datos tiene solo los dos primeros sumandos de la derecha, sería una recta. Si tenemos que incluir el siguiente término sería un polinomio de segundo grado o curva cuadrática, *etc.*

Los polinomios son ampliamente utilizados para describir la relación entre dos variables, aunque estas relaciones no siempre estén apoyadas por alguna teoría o hipótesis. El motivo de este amplio uso es que para cualquier grupo de pares de observaciones siempre es posible encontrar un polinomio que se ajuste exactamente a los datos. El grado del polinomio que se requiere para que se realice este buen ajuste es, como máximo, uno menos que el número de pares de observaciones, aunque en la práctica, rara vez se utilizan polinomios mayores del tercer o cuarto grado, pues las curvas resultantes son auténticas montañas rusas sin sentido biológico.

La manera de encontrar una expresión que explique nuestros datos es hacer lo de siempre: los vamos ajustando a diferentes líneas y curvas hasta encontrar la que nos minimice, significativamente, el residuo.

Para el ajuste a polinomios, el problema se reduce a encontrar los coeficientes α, β_1, β_2, *etc.*, hasta encontrar un polinomio que minimice el residuo. Para esto, hacemos uso de lo que conocemos como ecuaciones normales; necesitamos tantas ecuaciones como coeficientes haya, o una más que el grado de la ecuación que deseamos ajustar.

Las ecuaciones normales son

$$\alpha n + \beta_1 \Sigma X + \beta_2 \Sigma X^2 + \beta_3 \Sigma X^3 + \ldots = \Sigma Y$$

$$\alpha \Sigma X + \beta_1 \Sigma X^2 + \beta_2 X \Sigma X^3 + \beta_3 \Sigma X^4 + \ldots = \Sigma X Y$$

$$\alpha \Sigma X^2 + \beta_1 \Sigma X^3 + \beta_2 \Sigma X^4 + \beta_3 X^5 + \ldots = \Sigma X^2 Y$$

$$\alpha \Sigma X^3 + \beta_1 \Sigma X^4 + \beta_2 \Sigma X^5 + \beta_3 \Sigma X^6 + \ldots = \Sigma X^3 Y$$

....

Sigamos la explicación con un ejemplo

Ejemplo.-

Se tiene la producción de judías verdes (*Y*) en seis tiempos diferentes de recolección (*X*). A la primera fecha le da el valor 0 y las fechas siguientes son el número de días transcurridos desde la fecha base.

	Tiempo *X*	*Producción* *Y*
	0	27.4
	4	39.3
	7	46.2
	10	47.8
	13	44.5
	18	24.5
Σ	52	229.7

Se comienza elaborando todas las posibles potencias y sumatorios que se pueden necesitar para la resolución de las ecuaciones normales

	X^2	X^3	X^4	X^5	X^6	XY	X^2Y	X^3Y
	0	0	0	0	0	0.0	0.0	0.0
	16	64	256	1024	4096	157.2	628.8	2515.2
	49	343	2401	16807	117649	323.4	2263.8	15846.6
	100	1000	10000	100000	1000000	478.0	4780.0	47800.0
	169	2197	28561	371293	4826809	578.5	7520.5	97766.5
	324	5832	104976	1889568	34012224	441.0	7938.0	142884.0
Σ	657	9436	146194	2378692	39960778	1978.1	23131.1	306812.3

Ahora se tiene todas las sumas que se necesitan para las ecuaciones normales hasta el tercer grado. Se comienza, como siempre, calculando la recta, bien por el método ya conocido de

$$b = \frac{SP}{SC_{(X)}}$$

$$a = \overline{Y} - b\,\overline{X}$$

o bien resolviendo las dos primeras ecuaciones normales

$$\alpha n + \beta_1 \Sigma X = \Sigma Y$$

$$\alpha \Sigma X + \beta_1 \Sigma X^2 = \Sigma XY$$

Sustituyendo con los valores de las anteriores tablas, se tiene el sistema

$$6\alpha + 52\,\beta_1 = 229.7$$

$$52\alpha + 658\,\beta_1 = 1978.1$$

Cuyas soluciones son

$$\beta_1 = -0.06093$$

$$\beta_2 = 38.8114$$

Se hace la prueba de ajuste para ver si es bueno y si el resultado fuera de no buen ajuste se seguiría buscando el ajuste a curvas logarítmicas y si estos tampoco fueran buenos se seguiría buscando el ajuste a polinomios.

Archivo del programa *SAS* (C14-5.SAS).-

```
title 'Ajustes curvilineos';
options ls=75 ps=60;
data regres;
infile 'c14-5.dat';
input dias prod;
logdias = log10(dias);
logprod = log10(prod);
dias2 = dias*dias;
dias3 = dias*dias*dias;
dias4 = dias*dias*dias*dias;
title 'Modelo:     Y = a + b X  ';
proc reg;
  model prod = dias;
run;
title 'Modelo:     Y = a + log X b  ';
proc reg;
  model prod = logdias;
run;
title 'Modelo:     log Y = a + b X  ';
proc reg;
  model logprod = dias;
run;
title 'Modelo:      log Y = a + log X b  ';
proc reg;
  model logprod = logdias;
run;
title 'Modelo: polinomio cuadrático  ';
proc reg;
  model prod = dias dias2;
run;
title 'Modelo: polinomio cúbico  ';
proc reg;
  model prod = dias dias2 dias3;
run;
title 'Modelo: polinomio cuarto grado  ';
proc reg;
  model prod =dias dias2 dias3 dias4;
run;
```

Archivo de datos (C14-5.DAT).-

```
  1    27.4
  5    39.3
  8    46.2
 11    47.8
 14    44.5
 19    24.5
```

Archivo de resultados (C14-5.LST).-

```
                        Modelo:     Y = a + b X

Model: MODEL1
Dependent Variable: PROD

                          Analysis of Variance
                             Sum of          Mean
Source              DF      Squares        Square      F Value       Prob>F
Model                1      0.76978       0.76978        0.006       0.9413
Error                4    500.57855     125.14464
C Total              5    501.34833

    Root MSE       11.18681     R-square       0.0015
    Dep Mean       38.28333     Adj R-sq      -0.2481
    C.V.           29.22109

                       Parameter Estimates
                    Parameter       Standard    T for H0:
Variable  DF         Estimate          Error   Parameter=0    Prob > |T|
INTERCEP   1       38.872347     8.78975042         4.422        0.0115
DIAS       1       -0.060932     0.77691152        -0.078        0.9413

                        Modelo:     Y = a + log X b

Model: MODEL1
Dependent Variable: PROD

                          Analysis of Variance
                             Sum of          Mean
Source              DF      Squares        Square      F Value       Prob>F
Model                1     50.04851      50.04851        0.444       0.5418
Error                4    451.29982     112.82496
C Total              5    501.34833

    Root MSE       10.62191     R-square       0.0998
    Dep Mean       38.28333     Adj R-sq      -0.1252
    C.V.           27.74552

                       Parameter Estimates
                    Parameter       Standard    T for H0:
Variable  DF         Estimate          Error   Parameter=0    Prob > |T|
INTERCEP   1       32.468274     9.74851422         3.331        0.0291
LOGDIAS    1        6.883989    10.33587296         0.666        0.5418

                        Modelo:     log Y = a + b X

Model: MODEL1
Dependent Variable: LOGPROD

                          Analysis of Variance
                             Sum of          Mean
Source              DF      Squares        Square      F Value       Prob>F
Model                1      0.00052       0.00052        0.027       0.8779
Error                4      0.07733       0.01933
C Total              5      0.07785
```

Root MSE	0.13904	R-square	0.0067
Dep Mean	1.56896	Adj R-sq	-0.2417
C.V.	8.86221		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.584236	0.10925043	14.501	0.0001
DIAS	1	-0.001581	0.00965647	-0.164	0.8779

Modelo: log Y = a + log X b

Model: MODEL1
Dependent Variable: LOGPROD

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.00611	0.00611	0.341	0.5907
Error	4	0.07174	0.01794		
C Total	5	0.07785			

Root MSE	0.13392	R-square	0.0785
Dep Mean	1.56896	Adj R-sq	-0.1519
C.V.	8.53574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.504702	0.12291014	12.242	0.0003
LOGDIAS	1	0.076066	0.13031561	0.584	0.5907

Modelo: polinomio cuadrático

Model: MODEL1
Dependent Variable: PROD

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	493.57666	246.78833	95.265	0.0019
Error	3	7.77167	2.59056		
C Total	5	501.34833			

Root MSE	1.60952	R-square	0.9845
Dep Mean	38.28333	Adj R-sq	0.9742
C.V.	4.20423		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	21.279578	1.79619211	11.847	0.0013
DIAS	1	5.321220	0.40591825	13.109	0.0010
DIAS2	1	-0.269021	0.01950494	-13.792	0.0008

Modelo: polinomio cúbico

Model: MODEL1
Dependent Variable: PROD

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	500.99528	166.99843	946.035	0.0011
Error	2	0.35305	0.17652		
C Total	5	501.34833			

Root MSE	0.42015	R-square	0.9993
Dep Mean	38.28333	Adj R-sq	0.9982
C.V.	1.09747		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	23.762843	0.60545741	39.248	0.0006
DIAS	1	3.606959	0.28487395	12.662	0.0062
DIAS2	1	-0.048047	0.03446469	-1.394	0.2980
DIAS3	1	-0.007361	0.00113547	-6.483	0.0230

Modelo: polinomio cuarto grado

Model: MODEL1
Dependent Variable: PROD

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	501.20600	125.30150	880.335	0.0253
Error	1	0.14233	0.14233		
C Total	5	501.34833			

Root MSE	0.37727	R-square	0.9997
Dep Mean	38.28333	Adj R-sq	0.9986
C.V.	0.98547		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	24.469156	0.79533474	30.766	0.0207
DIAS	1	2.795683	0.71415242	3.915	0.1592
DIAS2	1	0.141464	0.15879879	0.891	0.5367
DIAS3	1	-0.022481	0.01246861	-1.803	0.3224
DIAS4	1	0.000384	0.00031565	1.217	0.4380

Como se ve no se ajusta ni a la recta ni a ninguna ecuación lineal (*logarítmica*). Sin embargo sí se ajusta a un polinomio de segundo grado. Una vez hallada una curva a la que se ajusta no es preciso seguir buscando polinomios de mayor grado, pues éstos también se van a ajustar y hay que tomar el más simple.

Como se ha indicado anteriormente, la representación de los residuos con respecto a la variable independiente es un método efectivo de comprobar si es suficiente con esta variable o, por contra, se necesita introducir otros miembros en el modelo. Si el modelo es adecuado, como lo era el del primer ejemplo de este capítulo, los residuos se reparten aleatoriamente alrededor del cero. Si el modelo no es el

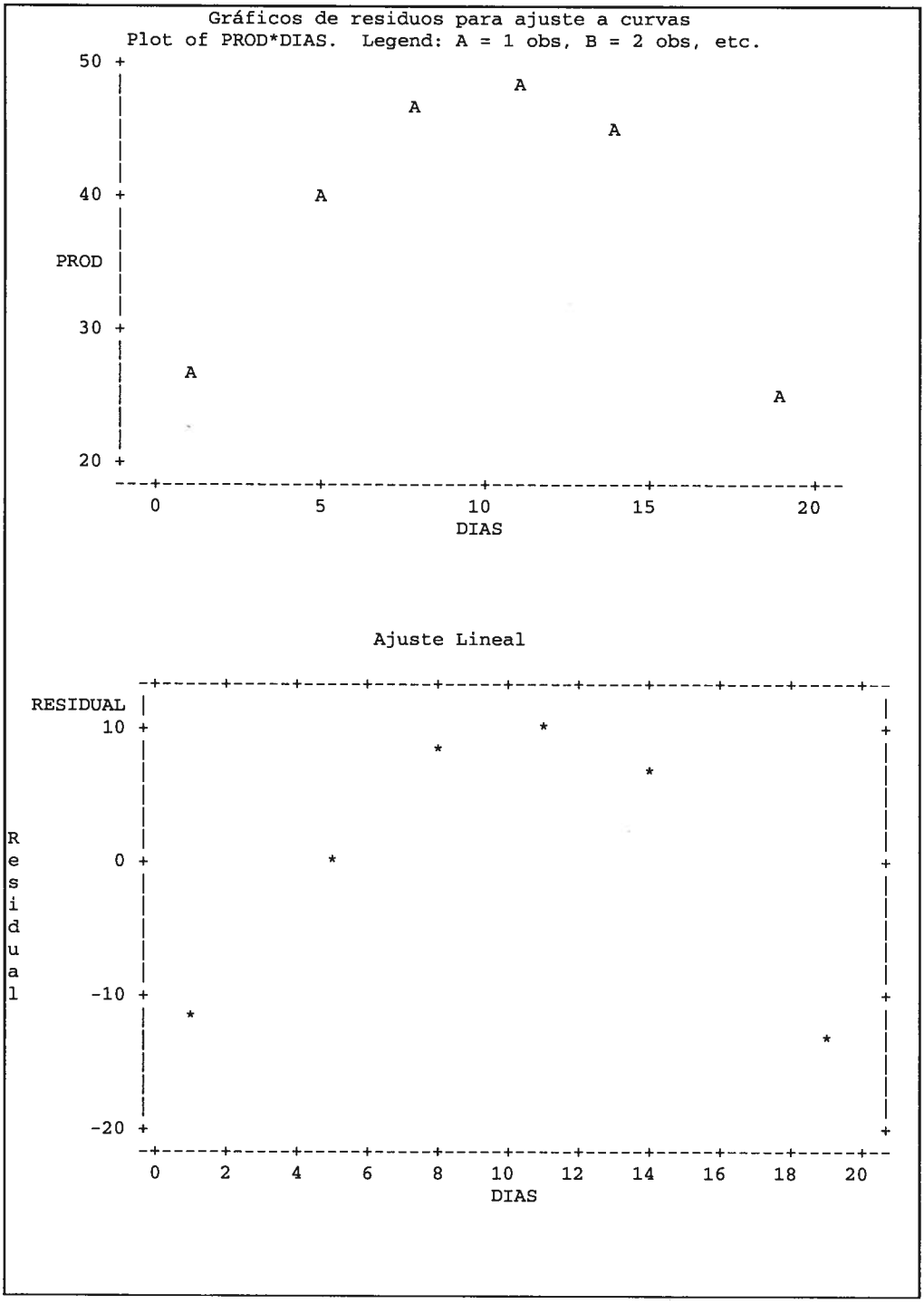
adecuado, los residuos mostraran una tendencia que nos orientaran hacia donde puede estar el buen ajuste.

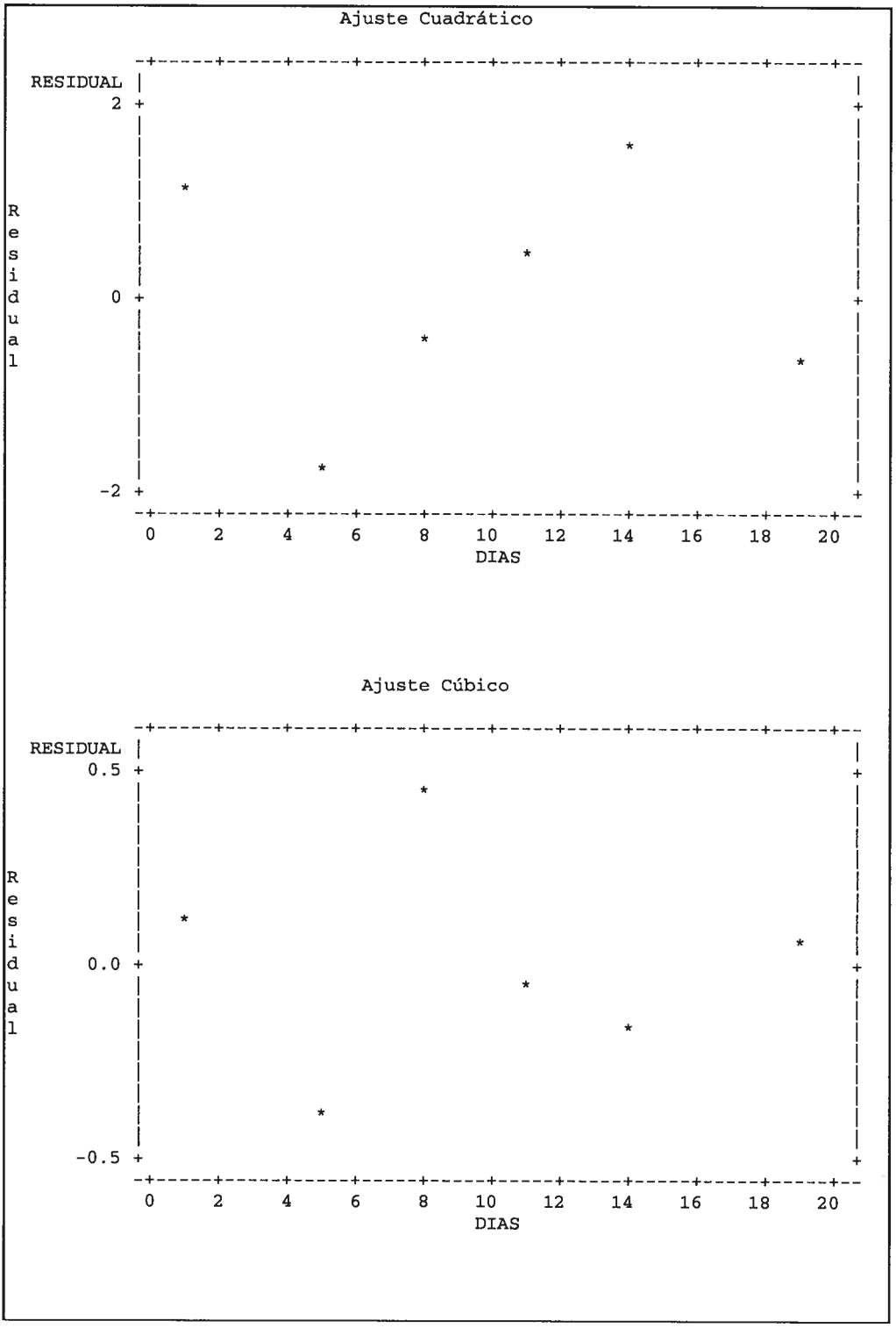
Si se hubiera hecho un estudio previo de *gráficos de residuos*, en el anterior problema, hubiera sido algo así.

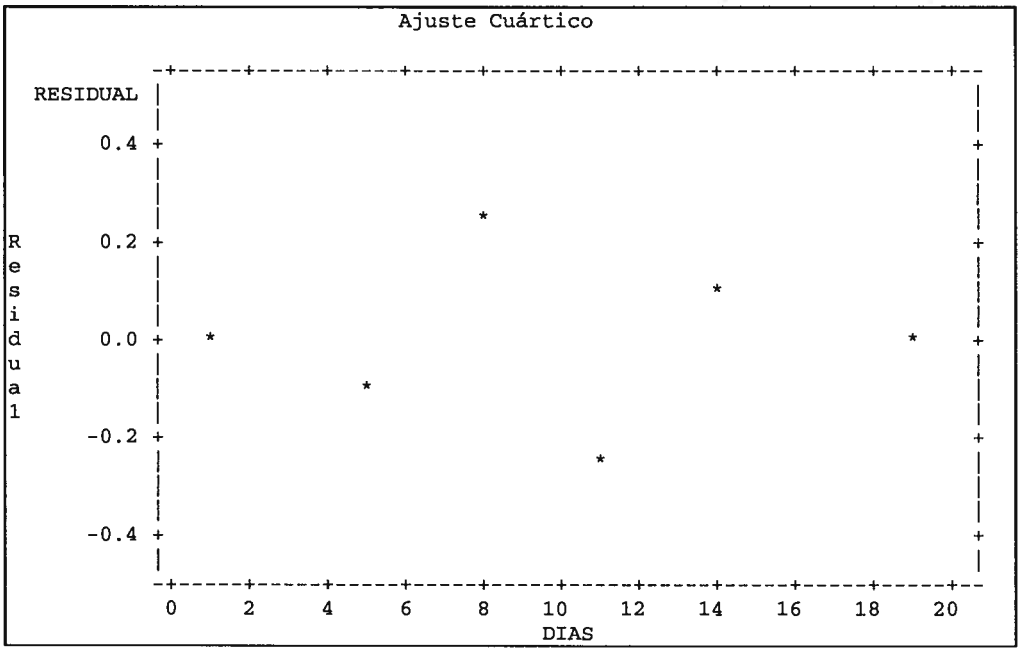
Archivo del programa *SAS* (C14-6.SAS).-

```
title 'Gráficos de residuos para ajuste a curvas';
options ls=75 ps=30;
data regres;
infile 'c14-5.dat';
input dias prod;
dias2 = dias*dias;
dias3 = dias*dias*dias;
dias4 = dias*dias*dias*dias;
proc plot;
 plot prod * dias;
run;
title 'Ajuste Lineal';
proc reg;
 model prod = dias / P noprint;
 plot residual.*dias='*';
run;
title 'Ajuste Cuadrático';
proc reg;
 model prod = dias dias2 / P noprint;
 plot residual.*dias='*';
run;
title 'Ajuste Cúbico';
proc reg;
 model prod = dias dias2 dias3 / P noprint;
 plot residual.*dias='*';
run;
title 'Ajuste Cuártico';
proc reg;
 model prod = dias dias2  dias3 dias4/ P noprint;
 plot residual.*dias='*';
```

Archivo de resultados (C14-6.LST).-




Ajuste Cuadrático
RESIDUAL
2
0
-2
Residual
0
2
4
6
8
10
12
14
16
18
20
DIAS
Ajuste Cúbico
RESIDUAL
0.5
0.0
-0.5
Residual
0
2
4
6
8
10
12
14
16
18
20
DIAS



Como se ve, la representación de los datos presenta una tendencia de subida y bajada. Esta tendencia es más acentuada cuando se estudia la gráfica de los residuos para el ajuste a la recta: comienza con valores de -10, sube a valores de +10 y vuelve a bajar a valores de -10. Esta tendencia indica que se necesita, por lo menos, un término cuadrático en el modelo. Se añade dicho término cuadrático y se analiza la gráfica de los residuos: se observa que estos oscilan alrededor del cero de una manera más aleatoria, por lo que podría ser este el primer ajuste a probar (en la resolución anterior de este mismo problema se comprobó que efectivamente este es un buen ajuste). De todas maneras, se ve una cierta tendencia, si añadimos un nuevo término cúbico, la gráfica de residuos se ve mas aleatoria alrededor del cero, y con un término cuártico, los residuos se distribuyen completamente horizontales alrededor del cero.

Polinomios ortogonales.-

Frecuentemente sucede que se hacen observaciones de una variable dependiente asociada con valores igualmente espaciados de una variable independiente; por ejemplo, si la variable independiente es el *tiempo* y se hacen lecturas de *Y* a intervalos diarios, semanales, mensuales o anuales, las *X* o tiempos son igualmente espaciados. Otros casos en los que a menudo se tienen intervalos igualmente espaciados de *X*, son los de experimentos que contemplan proporciones de fungicidas, insecticidas, antibióticos, hormonas, etc. Un experimento en el que las proporciones de tratamientos son igualmente espaciados presentan ventajas reales desde el punto de vista de la facilidad del análisis.

La base del método para el ajuste a curvas polinómicas se estudió en el Capítulo 10 y está basado en la descomposición de la suma de cuadrados por medio de los coeficientes expresados en la Tabla 6. Estas tablas pueden usarse para

encontrar las ecuaciones de regresión lineal, cuadrática, cúbica y de cuarto grado para cualquier número de observaciones.

En la parte superior de la tabla se encuentran los valores de n, es decir, el número de observaciones o tratamientos. Para cualquier problema dado, se necesita utilizar solamente la porción de la tabla por debajo del valor apropiado de n. La primera columna de coeficientes, encabezada por c_1, además de ser utilizada para diversos cálculos, consiste en valores codificados de X. La codificación se hace en forma tal que su resultados son números enteros lo más pequeños posibles. A pesar de que los valores de X estén igualmente espaciados, si n es impar, se puede tomar

$$X' = \frac{X - \overline{X}}{L}$$

donde L es el intervalo entre valores sucesivos de X. Si n es par, se toma

$$X' = \frac{(X - \overline{X})\,2}{L}$$

Esta transformación dará los valores de la columna c_1.

Los pasos para determinar las ecuaciones de regresión *lineal*, *cuadrática*, *cúbica* y *cuártica* son los siguientes:

1 Poner los valores de Y en una columna, de acuerdo con los valores ascendentes de las X, es decir, empezando con la Y correspondiente al menor valor de X.

2 Multiplicar los valores de Y por los coeficientes para c_1, c_2, c_3 y c_4 mostrados en la Tabla 6, obteniéndose cuatro columnas.

3 Encontrar la suma de cada columna, observando los signos más y menos. Estas sumas se denotan por ΣY, P_1, P_2, P_3 y P_4.

4 Aplicando los valores obtenidos de P_i y los valores de K provenientes de la Tabla 6, las ecuaciones *lineales*, *cuadráticas*, *cúbicas* y *cuárticas* pueden ser planteadas a partir de las siguientes relaciones:

Lineal $\hat{Y}_1 = \overline{Y} + (K_2 P_1) X'$

Cuadrática $\hat{Y}_2 = (\overline{Y} - K_1 P_2) + (K_2 P_1) X' + (K_4 P_2) X'$

Cúbica $\hat{Y}_3 = (\overline{Y} - K_1 P_2) + (K_2 P_1 - K_3 P_3) X' + (K_4 P_2) X'^2 + (K_5 P_3) X'^3$

Cuártica $\hat{Y}_4 = (\overline{Y} - K_1 P_2 + K_8 P_4) + (K_2 P_1 - K_3 P_3) X' + (K_4 P_2 + K_7 P_4) X'^2 + (K_5 P_3) X'^3 + (K_6 P_4) X'^4$

Nótese que estas ecuaciones están expresadas en términos de valores codificados de X.

5 Si los valores de Y en el primer paso fuesen el resultado de diferentes observaciones o repeticiones en cada nivel de X, y si se desea que las ecuaciones estén dadas en términos de las medias, se debe dividir cada término de las ecuaciones entre el número de repeticiones.

Ejemplo.-

Se tiene la producción total de leche diaria de 37 vacas controladas una vez al mes para los diez meses de lactación. (los datos están en el archivo C14-7.DAT, unas páginas más adelante)

Producción X	*Mes* Y
2442.3	1
2517.6	2
2334.4	3
2166.1	4
2030.0	5
1903.9	6
1779.5	7
1630.6	8
1485.7	9
1304.7	10
Σ 19594.8	

Aplicando los cinco pasos anteriormente expuestos, se comienza elaborando la siguiente tabla

c_1	c_1Y	c_2	c_2Y	c_3	c_3Y	c_4	c_4Y
-9	-21980.7	6	14653.8	-42	-102576.6	18	43961.4
-7	-17623.2	2	5035.2	14	35246.4	-22	-55387.2
-5	-11672.0	-1	-2334.4	35	81704.0	-17	-39684.8
-3	-6498.3	-3	-6498.3	31	67149.1	3	6498.3
-1	-2030.0	-4	-8120.0	12	24360.0	18	36540.0
1	1903.9	-4	-7615.6	-12	-22846.8	18	34270.2
3	5338.5	-3	-5338.5	-31	-55164.5	3	5338.5
5	8153.0	-1	-1630.6	-35	-57071.0	-17	-27720.2
7	10399.9	2	2971.4	-14	-20799.8	-22	-32685.4
9	11742.3	6	7828.2	42	54797.4	18	23484.6
P_1 = -22266.6		P_2 = -1048.8		P_3 = 4798.2		P_4 = -5384.6	

Los coeficientes c_1, c_2, c_3 y c_4 son los de la Tabla 6 y han sido multiplicados por los valores correspondientes de Y (producción de leche) para obtener los totales de dichas columnas, teniendo de esta manera ΣY, P_1, P_2, P_3 y P_4. Por lo que aplicando el paso 4 se obtienen las ecuaciones

$$\hat{Y}_1 = 1959.48 + \left(\frac{1}{330} \times -22266.6\right) X' = 1959.48 - 67.475 X'$$

$$\hat{Y}_2 = \left[1959.48 + \left(\frac{1}{32} \times -1048.8\right)\right] - 67.475 X' + \left(\frac{1}{1056} \times -1048.8\right) X'^2$$

$$= 1992.26 - 67.475 X' - 0.9932 X'^2$$

$$\hat{Y}_3 = 1992.26 + \left[-67.475 - \left(\frac{293}{205920} \times -4798.2\right)\right] X' -$$

$$- 0.9932 X'^2 + \left(\frac{1}{41184} \times 4798.2\right) X'^3$$

$$= 1992.26 - 74.302 X' - 0.9932 X'^2 + 0.1165 X'^3$$

$$\hat{Y}_4 = \left[1992.26 + \left(\frac{1}{1280} \times -5384.6\right)\right] - 7.4302 X' +$$

$$+ \left[-0.9932 - \left(\frac{41}{54912} \times -5384.6\right)\right] X'^2 +$$

$$+ 0.11651 X'^3 + \left(\frac{1}{109824} \times -5384.6\right) X'^4$$

$$= 1954.40 - 74.302 X' + 3.0272 X'^2 + 0.1165 X'^3 + 0.049029 X'^4$$

Cómo separar la suma de cuadrados.-

Ya se vio en el Capítulo 10 en el epígrafe *Contrastes ortogonales*, como obtener la suma de cuadrados asociada con un solo grado de libertad a partir de un conjunto de coeficientes, mediante la fórmula general

$$SC = \frac{(\Sigma c_i Y_i)^2}{n \Sigma c_i^2}$$

Tal como se han calculado las P, se tiene que $P_1 = \Sigma c_i Y_i$ cuando las c son de los coeficientes lineales. $P_2 = \Sigma c_i Y_i$ cuando se usan los coeficientes cuadráticos, *etc.*

Los divisores mostrados en la Tabla 6 son las sumas de cuadrados de los coeficientes, por tanto, la suma de cuadrados debida a la regresión cuadrática es P_1 dividida por el número de replicas; la suma de cuadrados para la regresión cuadrática es P_2, y así sucesivamente hasta el componente de cuarto grado.

Después de calcular las sumas de cuadrados para cada componente, se puede encontrar la suma de cuadrados residual al sustraer las sumas de cuadrados de los componentes a la suma de cuadrados total. Esta suma de cuadrados residual es igual

a la suma de cuadrados de las desviaciones con respecto a la curva de los datos observados.

Ejemplo.-

Aplíquese este método al ejemplo anterior.

El valor de P_1 es de -22266.6 de modo que la *SC lineal* es

$$SC_{(\text{Lineal})} = \frac{-22266.6^2}{330 \times 37} = 40606.18$$

La suma de cuadrados total de Y (*entre meses*) es

$$SC_{(\text{Meses})} = 41343.01$$

de manera que la suma de cuadrados debida a la desviación de la línea recta es

$$SC_{(\text{desviación lineal})} = SC_{(\text{Meses})} - SC_{(\text{Lineal})} = 41342.74 - 40606.18 = 736.56$$

Como P_2 es igual a -1048.8, la suma de cuadrados para el componente cuadrático es

$$SC_{(\text{Cuadrático})} = \frac{-1048.8^2}{132 \times 37} = 225.22$$

Restando este resultado de la suma de cuadrados de la desviación de la línea, que es una suma de cuadrados residual, se obtiene la desviación con respecto a la línea cuadrática

$$SC_{(\text{desviación cuadrática})} = SC_{(\text{desv. lineal})} - SC_{(\text{Cuadrática})} = 736.53 - 225.22 = 511.34$$

$P_3 = 4798.2$ de modo que la suma de cuadrados del componente cúbico es

$$SC_{(\text{Cúbica})} = \frac{4798.2^2}{8580 \times 37} = 72.52$$

Y el residuo o desviación es 511.34-72.52 = 438.82.

Finalmente, P_4 es igual a -5384.6 de manera que la suma de cuadrados para el componente cuártico es

$$SC_{(\text{Cuártico})} = \frac{-5384.6^2}{2860 \times 37} = 273.99$$

Y el residuo o desviación es 438.82-273.99=164.83

Todos estos resultados se pueden resumir en la siguiente tabla

FV	*gl*	*SC*	*CM*	*F*
Vaca	36	22946.35	637.40	19.55***
Mes	9	41342.74	4593.64	140.91***
Lineal	1	40606.18	40606.18	1245.58**
desv. lineal	8	736.56	92.07	2.82**
Cuadrática	1	225.22	225.22	6.91*
desv. cuadráti	7	511.34	73.05	2.24*
Cúbica	1	72.52	72.52	2.22*ns*
desv. cúbica	6	438.82	73.14	2.24*
Cuártica	1	273.99	273.99	8.40**
desv. cuártica	5	164.83	32.97	1.01*ns*
Error	324	10562.11	32.60	
Total	369	74851.21		

Como se puede ver, existe una diferencia altamente significativa entre vacas y entre meses. Este resultado es de esperar; lo que realmente se desea saber es el patrón de cambio de la producción de leche mes a mes. El alto valor de *F* para el componente *lineal* indica la existencia de una tendencia descendente altamente significativa, pero la desviación significativa del componente lineal indica que una línea recta no explica correctamente la variación mensual. El componente *cuadrático* significativo muestra que una curva simple representa una mejora sobre la recta, pero aún persiste una cantidad significativa de variación residual. El ajuste a una curva *cúbica* no mejora significativamente y el residuo dejado es significativo. Sin embargo, el componente cuártico explica una proporción muy elevada de la suma de cuadrados restante, en el sentido de que la *F* es altamente significativa, y la desviación de la componente cuártica es no significativa. Por tanto se tiene que la componente cuártica es la que mejor explica la variación mensual de la producción de leche en estos rebaños.

Archivo del programa *SAS* (C14-7.SAS).-

```
title 'Ajuste a Polinomios';
options ls=75 ps=60;
data poli;
infile 'c14-7.dat';
input vaca mes prod @@;
proc glm;
 class vaca mes;
  model prod=vaca mes;
  contrast 'Lineal ' mes  -9  -7  -5 -3 -1   1   3   5   7  9;
  contrast 'Cuadrát' mes   6   2  -1 -3 -4  -4  -3  -1   2  6;
  contrast 'Cúbica ' mes -42  14  35 31 12 -12 -31 -35 -14 42;
  contrast 'Cuártic' mes  18 -22 -17  3 18  18   3 -17 -22 18;
run;
```

Archivo de datos (C14-7.DAT).-

```
 1  1 81.58  2  1 86.79  3  1 85.21  4  1 40.82  5  1 45.93
 6  1 50.48  7  1 67.34  8  1 59.91  9  1 69.18 10  1 65.34
11  1 71.73 12  1 64.93 13  1 69.27 14  1 68.57 15  1 59.74
16  1 67.87 17  1 60.15 18  1 71.11 19  1 71.46 20  1 70.52
21  1 67.61 22  1 63.26 23  1 71.86 24  1 71.99 25  1 63.45
26  1 71.63 27  1 65.70 28  1 66.90 29  1 59.89 30  1 69.49
31  1 67.44 32  1 61.57 33  1 61.14 34  1 64.27 35  1 60.45
36  1 69.26 37  1 58.50
 1  2 82.71  2  2 90.39  3  2 92.35  4  2 41.90  5  2 47.10
 6  2 53.64  7  2 68.08  8  2 61.77  9  2 71.08 10  2 68.82
11  2 65.94 12  2 65.79 13  2 68.19 14  2 73.24 15  2 67.02
16  2 73.49 17  2 67.15 18  2 65.69 19  2 69.19 20  2 70.54
21  2 64.94 22  2 62.48 23  2 72.51 24  2 63.29 25  2 64.13
26  2 70.11 27  2 68.82 28  2 73.64 29  2 71.03 30  2 70.15
31  2 64.29 32  2 71.51 33  2 62.17 34  2 70.98 35  2 69.79
36  2 62.93 37  2 70.74
 1  3 77.68  2  3 83.31  3  3 80.29  4  3 41.12  5  3 46.35
 6  3 49.75  7  3 60.29  8  3 69.46  9  3 62.50 10  3 67.38
11  3 67.73 12  3 66.13 13  3 61.01 14  3 57.35 15  3 65.18
16  3 67.06 17  3 61.11 18  3 62.62 19  3 63.86 20  3 63.64
21  3 61.66 22  3 68.41 23  3 61.63 24  3 56.69 25  3 68.05
26  3 65.10 27  3 59.69 28  3 61.47 29  3 59.01 30  3 57.15
31  3 62.05 32  3 64.77 33  3 62.85 34  3 59.16 35  3 63.46
36  3 64.75 37  3 64.64
 1  4 81.72  2  4 80.91  3  4 72.59  4  4 32.38  5  4 35.91
 6  4 39.57  7  4 54.27  8  4 58.42  9  4 60.27 10  4 55.57
11  4 55.23 12  4 54.26 13  4 62.19 14  4 60.93 15  4 60.15
16  4 57.39 17  4 58.58 18  4 57.88 19  4 59.03 20  4 55.90
21  4 64.76 22  4 54.12 23  4 59.09 24  4 59.62 25  4 64.16
26  4 56.19 27  4 63.51 28  4 53.25 29  4 52.97 30  4 59.84
31  4 55.47 32  4 54.01 33  4 57.35 34  4 55.29 35  4 52.11
36  4 63.11 37  4 88.12
 1  5 76.80  2  5 80.87  3  5 69.96  4  5 31.21  5  5 35.31
 6  5 37.85  7  5 55.87  8  5 57.39  9  5 59.54 10  5 52.70
11  5 57.75 12  5 61.10 13  5 50.03 14  5 54.88 15  5 58.22
16  5 49.97 17  5 53.27 18  5 59.17 19  5 56.99 20  5 52.42
21  5 59.00 22  5 48.96 23  5 52.56 24  5 54.40 25  5 54.43
26  5 56.39 27  5 55.37 28  5 53.42 29  5 48.43 30  5 51.30
31  5 58.17 32  5 49.32 33  5 58.63 34  5 58.86 35  5 50.58
36  5 55.57 37  5 53.33
 1  6 75.19  2  6 71.59  3  6 76.54  4  6 34.40  5  6 35.53
 6  6 29.98  7  6 57.33  8  6 48.94  9  6 51.39 10  6 49.28
11  6 57.70 12  6 54.59 13  6 52.86 14  6 49.14 15  6 56.05
16  6 52.92 17  6 45.32 18  6 57.49 19  6 54.80 20  6 57.44
21  6 56.67 22  6 47.24 23  6 45.65 24  6 50.80 25  6 46.16
26  6 55.72 27  6 45.07 28  6 52.18 29  6 45.69 30  6 50.90
31  6 51.05 32  6 54.94 33  6 49.47 34  6 55.42 35  6 51.97
36  6 56.16 37  6 20.33
 1  7 62.07  2  7 68.48  3  7 65.98  4  7 26.05  5  7 27.57
 6  7 22.97  7  7 52.21  8  7 48.04  9  7 44.81 10  7 48.71
11  7 45.54 12  7 47.36 13  7 53.10 14  7 45.33 15  7 41.63
16  7 50.75 17  7 43.90 18  7 45.30 19  7 48.77 20  7 47.91
21  7 50.60 22  7 50.61 23  7 49.32 24  7 48.15 25  7 49.79
26  7 43.47 27  7 41.86 28  7 53.13 29  7 47.74 30  7 45.19
31  7 42.36 32  7 52.15 33  7 44.39 34  7 42.90 35  7 52.69
36  7 46.38 37  7 82.27
 1  8 67.68  2  8 57.66  3  8 61.97  4  8 24.40  5  8 29.54
```

```
 6   8 21.84   7   8 39.61   8   8 43.21   9   8 43.74 10   8 40.54
11   8 42.02 12   8 41.99 13   8 43.45 14   8 45.64 15   8 41.11
16   8 37.83 17   8 41.93 18   8 43.72 19   8 42.44 20   8 46.87
21   8 42.19 22   8 38.72 23   8 45.71 24   8 48.34 25   8 45.06
26   8 43.20 27   8 49.93 28   8 40.45 29   8 42.49 30   8 49.80
31   8 44.04 32   8 47.78 33   8 40.71 34   8 39.66 35   8 42.10
36   8 45.04 37   8 68.20
 1   9 57.10   2   9 54.10   3   9 58.14   4   9 16.12   5   9 25.73
 6   9 15.34   7   9 34.09   8   9 41.05   9   9 45.13 10   9 42.31
11   9 36.17 12   9 40.49 13   9 33.87 14   9 34.59 15   9 45.64
16   9 39.79 17   9 37.68 18   9 46.71 19   9 44.42 20   9 37.17
21   9 35.27 22   9 42.40 23   9 39.54 24   9 35.63 25   9 39.79
26   9 45.58 27   9 35.34 28   9 45.71 29   9 40.60 30   9 42.12
31   9 39.17 32   9 41.15 33   9 45.46 34   9 46.64 35   9 37.27
36   9 43.23 37   9 45.18
 1 10 63.81   2 10 62.24   3 10 63.31   4 10 42.44   5 10 42.88
 6 10 10.41   7 10 40.48   8 10 40.38   9 10 40.93 10 10 37.92
11 10 28.77 12 10 31.94 13 10 34.61 14 10 37.91 15 10 32.82
16 10 31.50 17 10 30.37 18 10 28.89 19 10 37.61 20 10 37.36
21 10 25.77 22 10 30.28 23 10 31.71 24 10 30.66 25 10 33.55
26 10 21.29 27 10 33.29 28 10 29.42 29 10 30.71 30 10 32.64
31 10 24.54 32 10 31.37 33 10 46.40 34 10 35.40 35 10 48.15
36 10 22.14 37 10 20.80
```

Archivo de resultados (C14-7.LST).-

```
                              Ajuste a Polinomios

Dependent Variable: PROD

                                        Sum of          Mean
Source                      DF          Squares        Square  F Value     Pr > F
Model                       45      64289.0962    1428.6466    43.82     0.0001
Error                      324      10562.1088      32.5991
Corrected Total            369      74851.2050

                       R-Square          C.V.      Root MSE          PROD Mean
                       0.858892      10.78109       5.70956            52.9590

Source                      DF       Type I SS  Mean Square  F Value     Pr > F
VACA                        36      22946.3517     637.3987    19.55     0.0001
MES                          9      41342.7445    4593.6383   140.91     0.0001

Source                      DF      Type III SS  Mean Square  F Value     Pr > F
VACA                        36      22946.3517     637.3987    19.55     0.0001
MES                          9      41342.7445    4593.6383   140.91     0.0001

Contrast                    DF      Contrast SS  Mean Square  F Value     Pr > F
Lineal                       1      40606.3273   40606.3273  1245.63     0.0001
Cuadrát                      1        225.1312     225.1312     6.91     0.0090
Cúbica                       1         72.4500      72.4500     2.22     0.1370
Cuártic                      1        273.8534     273.8534     8.40     0.0040
```

Bibliografía

Dagnelie, P. 1970. THÉORIE ET MÉTHODES STATISTIQUES. Ed J. Duculot, S.A. Gembloux.

Freund, R.J., and Littel, R.C. 1991. *SAS* SYSTEM FOR REGRESION. SAS Institute Inc., Cary, NC, USA.

*Infante Gil, S. y Zárate De Lara, G.P.*1984. METODOS ESTADISTICOS. Ed. TRILLAS. México.

Lite, TM, y Jackson Hills, F. 1987. METODOS ESTADISTICOS PARA LA INVESTIGACION EN LA AGRICULTURA. Ed TRILLAS. México.

Ostle, B. 1965. ESTADISTICA APLICADA. Ed. Limusa-Wiley. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

Littel, R.C., Freund, R.J. and Spector, P.C. 1991. *SAS* FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. 1990. *SAS* PROCEDURE GUIDE. Cary, NC, USA.

CAPÍTULO 15

Correlación simple

CAPÍTULO 15

Correlación simple

Introducción.-

Se continua en éste, como en los tres últimos capítulo, con el estudio de estadísticos de dos dimensiones o distribuciones bivariantes. En los capítulos anteriores se estudió que la regresión trata de la relación funcional de una variable sobre la otra. En el presente capítulo se va a tratar de ***medir el grado de asociación o de variación conjunta de dos variables*** cualesquiera. Esta materia se conoce en Estadística con el nombre de ***correlación lineal***.

Correlación y regresión.-

Si se tienen pares aleatorios de observaciones, es corriente encontrarse con la duda de cuál técnica estadística utilizar, si la regresión o la correlación, y es corriente, así mismo, encontrarse con una utilización incorrecta o insuficiente de estas técnicas.

Existe mucha confusión acerca de la regresión y la correlación, de manera que es frecuente encontrarlos confundidos. Esta confusión viene motivada, en primer lugar, por la semejanza en los cálculos para ambos coeficientes, puesto que para ambos, la cantidad fundamental es la *covarianza* o *suma de los productos*.

Con la regresión, lo que se intenta es describir la *dependencia* de una variable Y de una variable independiente X. Trata, sobre todo, de las medias de una variable (la dependiente) y cómo éstas medias cambian su localización cuando cambia el valor de la otra variable (la independiente). La ecuación de regresión se puede emplear con alguno de estos tres fines

Apoyar las hipótesis que postulan la posible causalidad de los cambios de Y en los cambios de X

Para propósitos de predicción de Y en términos de X

Para propósitos de explicar qué parte de la variación de *Y* es debida a *X*, utilizando esta variable como control estadístico.

Algunos ejemplos de regresión pueden ser: los estudios de los efectos de la temperatura sobre los latidos cardíacos, el contenido de nitrógeno en el suelo sobre el crecimiento de una planta, la edad de un animal sobre la presión sanguínea o la dosis de un determinado insecticida sobre la mortalidad de una población de insectos. Estos son ejemplos típicos de regresión para los propósitos señalados más arriba.

Con la correlación, por el contrario, se investiga si dos variables son independientes o covarían, esto es, si varían conjuntamente. No se expresa una como función de la otra, así como tampoco se hace distinción alguna entre variables dependientes e independientes. Podría ocurrir que de una pareja de variables, cuya correlación se estudia, una sea causa de la otra pero no se sabe ni se sospecha.

Una hipótesis importante, aunque no esencial, para la correlación, es que las dos covariables sean efectos de una causa común, y lo que se desea conocer es el grado en que ambas variables varían conjuntamente.

Algunos ejemplos de correlación pueden se: la correlación entre las longitudes de piernas y brazos en una población de mamíferos, o entre el peso del cuerpo y la producción de huevos en insectos, o entre los días necesarios para la madurez y el número de semillas en un cultivo. Las razones por las que se quiere demostrar y medir la asociación entre pares se estudiará más adelante.

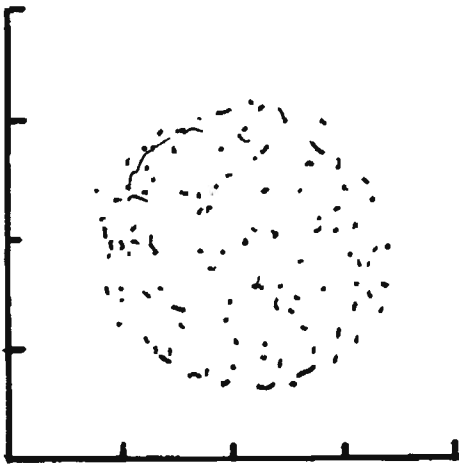
Otra dificultad es que parezca que se esta utilizando el método correcto, con arreglo a lo dicho anteriormente, pero puede surgir complicaciones con la toma de datos. Así, por ejemplo, si se desea establecer el contenido de colesterol en la sangre como función del peso, se podría llevar a cabo el experimento tomando una muestra aleatoria de hombres de la misma edad, obteniendo al mismo tiempo el contenido de colesterol en la sangre, así como el peso de cada individuo, y calculando la regresión. Sin embargo, ambas variables habrán sido medidas con error. Los valores individuales de la variable supuestamente independiente, *X*, no fueron deliberadamente escogidos o controlados por el experimentador. Las condiciones básicas del *Modelo I* de regresión no se cumplen, y el ajuste de los datos a un Modelo I no es, por tanto, correcto. Lo que no quita para que existan multitud de ejemplos de este tipo que usan la regresión.

También se presenta la dificultad inversa, la de tratar de obtener un coeficiente de correlación a partir de datos que se han tomado para un uso apropiado de la regresión, es decir, con *X* fijo. Un ejemplo serían los latidos de corazón de un animal poiquilotermo como función de la temperatura. Este tipo de coeficiente de correlación se obtiene matemáticamente de manera sencilla, pero sería simplemente un valor numérico y no una estima de la correlación paramétrica.

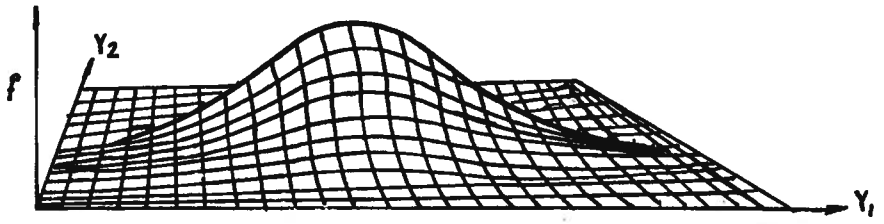
Función de densidad bivariante y supuestos paramétricos.-

No se va estudiar la función de densidad de la distribución normal bidimensional o bivariante, pero sí se va a hacer una aproximación intuitiva a ella.

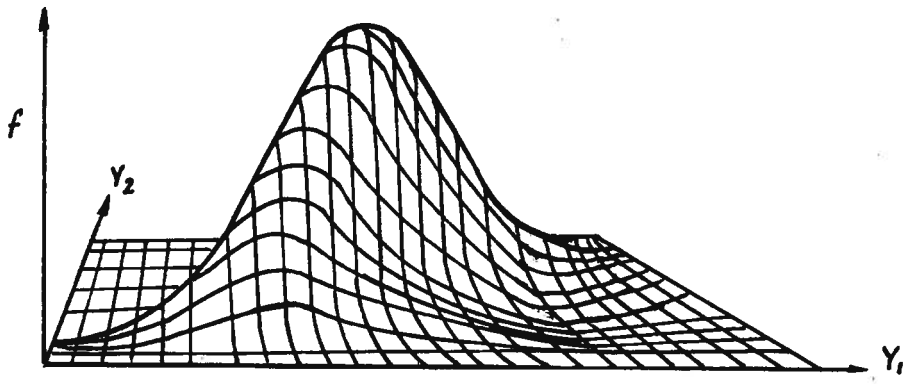
Supóngase que se ha muestreado un centenar de unidades experimentales y que se han medido dos variables de cada unidad, obteniendo de esta forma dos muestras de cien datos cada una. Supóngase, así mismo, que ambas variables están distribuidas normalmente y también que son independientes una de otra, de manera que el hecho de que una unidad experimental presente un valor para la variable *Y* mayor que la media de esta variable, no influye sobre el valor de la variable *X*, de manera que esta misma unidad experimental puede tener un valor de la variable *X* por encima o por debajo de la media de esta variable, con la misma probabilidad. Si no existe relación alguna entre las variables *Y* e *X* y si las dos variables están tipificadas a fin de hacer sus escalas comparables, se encontrará que el perfil del *diagrama de dispersión* es aproximadamente circular. Por lo que si se representan estos cien datos sobre un gráfico, en el que la variable *X* e *Y* son la coordenadas, se obtendrá un diagrama de puntos similar a este



Desde luego, para una muestra de cien datos, el diagrama poseerá un perfil circular imperfecto, pero a medida que la muestra sea más grande, más claramente se marcará un circulo alrededor de la intersección $\overline{X}, \overline{Y}$. Si se sigue muestreando se tendrá que superponer nuevos puntos en los mismos lugares que ya tienen puestos otros puntos; si en lugar de puntos fueran, por ejemplo, granos de arena, lo que ocurriría sería que se irían amontonando de manera que darían lugar (conforme la *muestra* se aproxima a la *población*) a una formación semejante a una campana maciza como la de la siguiente gráfica



Supóngase, ahora, que las dos variables, X e Y, no son independientes sino que están correlacionadas o asociadas directamente. De esta manera, si una unidad experimental dada posee un valor de X por encima de su media, la probabilidad de que el valor de Y, de la misma unidad, este por encima de su media es mayor que la probabilidad de que este por debajo de su media. De manera análoga, un valor pequeño de X estará asociado con más probabilidad a un valor pequeño de Y que a un valor grande. Si se muestreasen datos de tal población, el diagrama de puntos resultante sería alargado, adoptando forma de elipse. Esto es de esta manera como consecuencia de que aquellas partes del circulo de puntos que en un principio incluían individuos con un alto valor para una variable y bajo para la otra (y viceversa) están ahora escasamente representados. Si se hace la experiencia anterior con los granos de arena, lo que se tendría ahora, si se muestrea la población, sería una campana maciza de forma elíptica, tal como se representa en la siguiente figura



Si la correlación o asociación entre los valores de ambas variables fuera perfecta, todos los datos caerían a lo largo de una línea de regresión única (esta línea describiría tanto la regresión $b_{X.Y}$ como la regresión $b_{Y.X}$) y si se siguen acumulando granos de arena se obtendrían una curva normal plana apoyada en dicha recta de regresión.

El tamaño del perfil, más o menos circular o más o menos elíptico, del diagrama de puntos y del montón de granos de arena, es claramente una función del grado de correlación entre las dos variables, y esta correlación viene medida por el parámetro ρ de la distribución normal bivariante. Este parámetro se define como

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

donde σ_{XY} es la covarianza paramétrica de las variables X e Y, y σ_X y σ_Y son las desviaciones típicas paramétricas de las variables X e Y.

Cuando las dos variables se distribuyen mediante la función normal bidimensional, el coeficiente de correlación muestral, r, permite estimar el coeficiente de correlación paramétrica ρ pudiéndose realizar, por tanto, afirmaciones acerca de la

distribución muestral de ρ, pruebas de hipótesis y establecer *intervalos de confianza* para dicho coeficiente.

Los supuestos para la estima y pruebas de hipótesis de ρ son

Un supuesto importante, aunque no esencial, para la correlación, es que las dos variables son efectos de una causa común, y lo que se desea conocer es el grado en que ambas variables varían conjuntamente.

Otro supuesto es el de la existencia de una *relación lineal* entre las dos variables en la población.

Otro supuesto es el de normalidad de las dos variables, es decir, el de una distribución normal bivariante.

A diferencia de la varianza o del coeficiente de regresión, el coeficiente de correlación es independiente de las unidades de medida; es una cantidad absoluta sin dimensión. Por lo que el coeficiente de correlación, ρ, puede variar desde +1 para la asociación directa perfecta, hasta -1 para la asociación indirecta perfecta, pasando por 0 para la ausencia de asociación.

Coeficiente de correlación.-

La correlación, como la covarianza y como la suma de productos, es una medida del grado en que dos variables varían conjuntamente o una medida de la intensidad de asociación. Por lo que se espera que haya simetría en las dos variables. El *coeficiente de correlación lineal muestral*, también llamado *correlación simple*, *correlación total* y *correlación momento-producto*, se usa tanto con propósitos descriptivos como de inferencia y viene dado por la fórmula

$$r = \frac{SP}{\sqrt{SC_{(X)}\, SC_{(Y)}}}$$

Es decir, es igual a la suma de productos dividida por el producto de las raíces cuadradas de las sumas de cuadrados de X e Y, o lo que es lo mismo, la covarianza dividida por el producto de las desviaciones típicas. El uso de Y y X no implica una variable dependiente y otra independiente. Se sigue utilizando esta terminología por ser la mas extendida, pero una terminología más correcta sería simbolizar las dos variables como X_1 y X_2.

Si se usa con propósitos inferenciales, el coeficiente de correlación r es una estima no sesgada del correspondiente coeficiente de correlación poblacional ρ sólo cuando el parámetro poblacional, ρ, es cero y existe, en la población, una relación lineal entre las dos variables.

También con propósitos inferenciales, hay que estar seguro, al menos, de la normalidad de una de las variables. Se puede calcular r para cualquier par de valores, (X, Y), cualquiera que sea la distribución, e incluso si se trata de datos que deberían

tratarse mediante regresión. Pero en tales casos, el coeficiente de correlación solo es un índice matemático, no un *estadístico muestral* que *estima* un *parámetro* desconocido, siendo, por tanto, de escaso interés. Si las variables se ajustan a una distribución normal bidimensional o bivariante, el coeficiente de correlación muestral, *r*, corresponderá a la estima de un parámetro de esta distribución simbolizado por ρ.

Al igual que ocurre con el parámetro, y a diferencia de la varianza o del coeficiente de regresión, el coeficiente de correlación es independiente de las unidades de medida; es una cantidad absoluta sin dimensión. Por lo que el coeficiente de correlación, *r*, puede variar desde +1 para la asociación directa perfecta, hasta -1 para la asociación indirecta perfecta, pasando por 0 para la ausencia de asociación. La correlación directa perfecta se ve claramente si se calcula para una misma variable *X*

$$r = \frac{SP_{(XX)}}{\sqrt{SC_{(X)}\ SC_{(X)}}} = \frac{SC_{(X)}}{\sqrt{SC_{(X)}^2}} = \frac{SC_{(X)}}{SC_{(X)}} = 1$$

Si el tamaño de muestra es pequeño, *r* es una estima ligeramente parcial y subestima ρ en pequeñas muestras. En el caso de pequeñas muestras, una estima imparcial de ρ sería

$$r^* = r\left(1 + \frac{1 - r^2}{2(n-4)}\right)$$

Aunque para la mayoría de las aplicaciones prácticas puede ignorarse esta parcialidad.

Coeficiente de alineación o factor de mejoramiento.-

Si se eleva al cuadrado el coeficiente de correlación muestral definido anteriormente, resulta

$$r^2 = \frac{SP^2}{SC_{(X)}\ SC_{(Y)}} = \frac{SP^2}{SC_{(X)}} \times \frac{1}{SC_{(Y)}}$$

El término de la izquierda de la segunda expresión es la suma de cuadrados debida a la regresión, es decir, es la fracción de la suma de cuadrados total de la variable *Y* que es debida a la variación de la variable *X*, por lo que el cuadrado del coeficiente de correlación es el cociente de la suma de cuadrados debida a la regresión dividido por la suma de cuadrados total de la variable *Y*

$$r^2 = \frac{\frac{SP^2}{SC_{(X)}}}{SC_{(Y)}} = \frac{SC_{(\text{regresión})}}{SC_{(Y)}}$$

Como la denominación de *X* y de *Y* no implica dependencia o independencia de

una variable sobre otra, lo mismo da tener Y explicada sobre X que lo contrario

$$r^2 = \frac{\frac{SP^2}{SC_{(X)}}}{SC_{(Y)}} = \frac{\frac{SP^2}{SC_{(Y)}}}{SC_{(X)}}$$

Este cociente es una proporción entre cero y uno, puesto que la suma de cuadrados explicada de cualquier variable tiene que ser menor que su suma de cuadrados total, o, en caso extremo, si explica toda la variación de una variable, puede ser tan grande como la suma de cuadrados total, pero no mayor.

Ya se vio que esta cantidad r^2 se denomina *coeficiente de determinación* que, efectivamente, varía de 0 a +1 y no puede ser negativo, independientemente de que lo sea r o b, pues todas las sumas de cuadrados son positivas y el numerador es una expresión elevada al cuadrado.

Puesto que el valor del coeficiente de determinación oscila entre 0 y +1, se le puede restar a 1 con el fin de, conocido el coeficiente de correlación de una muestra r y la suma de cuadrados total de la variable Y, poder calcular la fracción de la suma de cuadrados explicada y no explicada, es decir, la fracción de la suma de cuadrados total de Y explicada por la variabilidad de X (debida a la regresión) es

$$SC_{(Y.X)} = \frac{SP^2}{SC_{(X)}} = r^2 SC_{(Y)}$$

Y la suma de cuadrados no explicada por X es

$$SC_{(\text{error})} = SC_{(Y)} - \frac{SP^2}{SC_{(X)}} = (1 - r^2) SC_{(Y)}$$

La cantidad $1\text{-}r^2$ se denomina *coeficiente de indeterminación* o de no determinación y expresa la proporción de la varianza de una variable que no ha sido explicada por la otra variable. La raíz cuadrada de este coeficiente de indeterminación

$$\sqrt{1 - r^2}$$

se denomina *coeficiente de alineación* o *factor de mejoramiento* y mide la falta de asociación entre las variables X e Y.

Relación entre los coeficientes de correlación y regresión.-

Esta relación puede deducirse fácilmente a partir de la expresión ya conocida

$$r = \frac{SP}{\sqrt{SC_{(X)}\ SC_{(Y)}}} = \frac{SP}{\sqrt{SC_{(X)}}} \times \frac{1}{\sqrt{SC_{(Y)}}}$$

Multiplicando el numerador y denominador por la raíz cuadrada de la suma de cuadrados de X se obtiene

$$r = \frac{SP}{\sqrt{SC_{(X)}}\sqrt{SC_{(X)}}} \times \frac{\sqrt{SC_{(X)}}}{\sqrt{SC_{(Y)}}} = \frac{SP}{SC_{(X)}} \times \frac{\sqrt{SC_{(X)}}}{\sqrt{SC_{(Y)}}}$$

Dividiendo el numerador y denominador del término de la derecha de esta expresión por la raíz cuadrada de n-1 se obtiene

$$r = \frac{SP}{SC_{(X)}} \times \frac{\sqrt{\frac{SC_{(X)}}{n-1}}}{\sqrt{\frac{SC_{(Y)}}{n-1}}} = b_{X.Y} \frac{S_X}{S_Y}$$

De manera análoga se puede deducir

$$r = b_{X.Y} \frac{S_Y}{S_X}$$

por tanto

$$b_{Y.X} = r \frac{S_Y}{S_X}$$

y

$$b_{X.Y} = r \frac{S_X}{S_Y}$$

Se tiene, entonces, que el coeficiente de correlación puede considerarse como si fuese un coeficiente de regresión tipificando. Si las dos desviaciones típicas son idénticas, ambos coeficientes, regresión y correlación, tendrán igual valor.

Una segunda relación entre los dos coeficientes es la siguiente. Multiplicando las dos regresiones, $b_{Y.X}$ y $b_{X.Y}$, se obtiene

$$b_{Y.X}\ b_{X.Y} = r \frac{S_Y}{S_X} r \frac{S_X}{S_Y} = r^2$$

Por lo que

$$r = \pm \sqrt{b_{Y.X}\ b_{X.Y}}$$

Es decir, que el coeficiente de correlación es igual a la media geométrica de los dos coeficientes de regresión.

Si las dos líneas de regresión se representan sobre el mismo gráfico, la correlación se puede considerar como la media del ángulo que forman dichas líneas de regresión. Si este ángulo es recto, no existirá correlación en los datos y se obtendrá un diagrama de dispersión circular. Si el ángulo es muy pequeño, entonces la correlación es muy alta, y en el caso de una correlación perfecta, las dos líneas de regresión coinciden por lo que el ángulo entre ellas será cero.

Pruebas de hipótesis e intervalos de confianza para ρ=0 y tamaños de muestra superior a 50.-

Como el rango de *r* es

$$-1 \leq r \leq 1$$

no se puede esperar que la distribución muestral de *r* sea simétrica cuando el parámetro poblacional ρ es diferente de cero. La simetría ocurre solamente para ρ=0, y la asimetría aumenta al acercarse ρ a +1 o a -1.

La prueba de significación más común consiste en determinar si un coeficiente de correlación muestral puede provenir de una población con un coeficiente de correlación paramétrica igual a cero. Las hipótesis, por tanto, pueden ser

Cola derecha	Cola izquierda	Dos colas
$H_o: \rho \leq 0$	$H_o: \rho \geq 0$	$H_o: \rho = 0$
$H_1: \rho > 0$	$H_1: \rho < 0$	$H_1: \rho \neq 0$

Es decir, la H_o tiene el término ρ=0. Esto implica que las dos variables no están correlacionadas. Si la muestra proviene de una distribución normal de dos variables y ρ=0 el error típico del coeficiente de correlación es

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

Por lo que la hipótesis se puede probar mediante una *t*

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}$$

que se contrasta con la *t* de *Student*, $t_{(n-2;\ \alpha/2)}$ para la hipótesis de dos colas y con la $t_{(n-2;\ \alpha)}$ para las hipótesis de una cola.

Esta prueba *t* es equivalente a la prueba de significación de *b*, midiéndose en ambos casos la fuerza de asociación lineal entre las dos variables. Y el cuadrado de esta *t* es análogo a la prueba *F* de ajuste de la regresión.

Siguiendo métodos similares a los estudiados en capítulos anteriores, resulta fácil calcular los límites o intervalo de confianza para ρ, de la siguiente manera

$$LC_{(\rho)} = r \pm S_r\, t_{(n-2;\alpha/2)}$$

Hay que insistir que esta prueba se aplica únicamente para H_0:ρ=0 y tamaño de muestra grande, de manera que no debe aplicarse para comprobar la hipótesis de que ρ posea un valor específico distinto de cero o con un tamaño de muestra pequeño. Esos caso se estudiará en el siguiente epígrafe.

También se podría haber hecho esta prueba por medio de una Z, tal como se verá, también, en el siguiente epígrafe.

Ejemplo.-

Se tiene el aumento de peso (X_1) y el peso total de alimento consumido (X_2) de nueve lechones en el mismo tiempo. Se quiere saber si existe una relación lineal positiva entre ambas variables.

Lechón	X_1	X_2
1	26.8	236
2	26.0	241
3	24.3	239
4	29.0	285
5	29.4	282
6	27.0	273
7	26.6	258
8	29.8	289
9	28.5	278

$$n = 9$$

$$\Sigma X_1 = 247.4000 \qquad \Sigma X_2 = 2381.0000$$

$$\overline{X}_1 = 27.4889 \qquad \overline{X}_2 = 254.5556$$

$$SC_{(X_1)} = 26.1880 \qquad SC_{(X_2)} = 3638.2500$$

$$SP = 272.2500$$

La correlación es, por tanto

$$r = \frac{272.25}{\sqrt{26.188 \times 3638.25}} = 0.8820$$

La prueba para contrastar si es estadísticamente diferente de cero, es

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0$$
$$S_r = \sqrt{\frac{1-0.882^2}{9-2}} = 0.1781$$
$$t_o = \frac{0.882}{0.1781} = 4.9518^{**}$$
$$t_{(7;0.01/2)} = 3.4995$$

Aunque, en este ejemplo, la hipótesis del experimentador, es que $\rho>0$, es decir, que a mayor consumo de alimento, mayor incremento de peso, por lo que la hipótesis a probar es de la cola derecha y la prueba sería

$$H_o : \rho \leq 0$$
$$H_1 : \rho > 0$$
$$S_r = \sqrt{\frac{1-0.882^2}{9-2}} = 0.1781$$
$$t_o = \frac{0.882}{0.1781} = 4.9518^{***}$$
$$t_{(7;0.001)} = 4.03$$

Si se hubiera realizado con intervalo de confianza, este sería

$$LC_{(\rho)} = 0.882 \pm 0.1781 \times 2.3646 = 0.882 \pm 0.4211$$
$$L_i = 0.4609$$
$$L^s > 1$$

no incluye el cero, luego es estadísticamente diferente de cero. El límite superior sobrepasa el rango paramétrico, esto es como consecuencia de que, como se ha indicado anteriormente, esta prueba *t* y este intervalo es para el caso en que el tamaño de muestra es grande. Se ha puesto un ejemplo con pocos datos con objeto de que cupiera el desarrollo completo de los datos en un espacio razonable. Pero un ejemplo con este tamaño de muestra habría que haberlo resuelto con la transformación *Z* como así se hará en el siguiente epígrafe.

Archivo del programa *SAS* (C15-1.SAS).-

```
title 'Correlación';
options ls=75 ps=60;
data correla;
infile 'c15-1.dat';
input x1 x2;
proc corr csscp cov;
run;
```

Archivo de datos (C15-1.DAT).-

```
 26.8    236
 26.0    241
 24.3    239
 29.0    285
 29.4    282
 27.0    273
 26.6    258
 29.8    289
 28.5    278
```

Archivo de resultados (C15-1.LST).-

```
                              Correlación

                          Correlation Analysis
                     2 'VAR' Variables:  X1       X2

                Corrected Sum-of-Squares and Crossproducts

                                    X1                  X2
                  X1          26.188889          272.255556
                  X2         272.255556         3638.222222

                       Covariance Matrix     DF = 8

                                    X1                  X2
                  X1           3.2736111          34.0319444
                  X2          34.0319444         454.7777778

                                Simple Statistics
Variable            N        Mean     Std Dev         Sum     Minimum     Maximum
X1                  9     27.4889      1.8093    247.4000     24.3000     29.8000
X2                  9    264.5556     21.3255        2381    236.0000    289.0000

  Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 9

                                    X1                  X2
                  X1           1.00000             0.88201
                                   0.0                  0.0017
                  X2           0.88201             1.00000
                                   0.0017               0.0
```

Pruebas de hipótesis para $\rho \neq 0$ y para $\rho = 0$ en muestras pequeñas.-

Cuando $\rho \neq 0$, la distribución de los valores muestrales de r son asimétricos y, aunque se pueda calcular el error típico de r, este no debe aplicarse, a menos que el tamaño de muestra sea muy grande (n>500). Para soslayar este problema se puede transformar r en valores Z de la distribución normal típica. Esta transformación consiste en

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

y por tanto

$$r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}$$

Observando estas expresiones se comprueba que para r=0, Z_r=0, ya que el logaritmo neperiano de uno es cero. Sin embargo cuando r se aproxima a uno, $(1+r)/(1-r)$ se aproxima al infinito, en consecuencia Z_r se aproxima al infinito. Por tanto irán apareciendo grandes diferencias entre los valores de r y de Z_r conforme r aumente de valor.

Por tanto, la ventaja de la transformación Z de r es que mientras que los coeficientes de correlación se distribuyen de manera asimétrica para valores de $\rho \neq 0$, los valores de Z_r están distribuidos de manera aproximadamente normal para cualquier valor del parámetro, al que se identificará con la zeta griega ξ. La varianza esperada de Z_r es

$$\sigma^2_{Z_r} = \frac{1}{n-3}$$

Esta es una aproximación adecuada para muestras de tamaño $n \geq 50$ e, incluso, una aproximación tolerable para $n \geq 25$. Como se puede comprobar, esta varianza es independiente de la magnitud de r, siendo simplemente una función del tamaño de la muestra. Por tanto, si el tamaño de muestra es superior a 50 se pueden realizar las pruebas de hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_o: \rho \leq 0$	$H_o: \rho \geq 0$	$H_o: \rho = 0$
$H_1: \rho > 0$	$H_1: \rho < 0$	$H_1: \rho \neq 0$

mediante la transformación Z con el siguiente estadístico

$$H_o: Z = 0$$
$$H_1: Z \neq 0$$
$$Z_o = \frac{Z_r}{\frac{1}{\sqrt{n-3}}} = Z_r\sqrt{n-3}$$

Dado que Z_o está distribuida normalmente, y que se está utilizando una desviación típica paramétrica, se puede utilizar esta Z_o para realizar cualquier prueba de hipótesis contrastándola con la Z normal típica (Tabla 1).

Todas las posibles pruebas de hipótesis son

Cola derecha	Cola izquierda	Dos colas
$H_o: \rho \leq \rho_o$	$H_o: \rho \geq \rho_o$	$H_o: \rho = \rho_o$
$H_1: \rho > \rho_o$	$H_1: \rho < \rho_o$	$H_1: \rho \neq \rho_o$

Que si se realiza la transformación Z equivalen a

Cola derecha	Cola izquierda	Dos colas
$H_o: Z \leq \zeta$	$H_o: Z \geq \zeta$	$H_o: Z = \zeta$
$H_1: Z > \zeta$	$H_1: Z < \zeta$	$H_1: Z \neq \zeta$

Y la prueba sería

$$Z_o = \frac{Z-\zeta}{\frac{1}{\sqrt{n-3}}} = (Z-\zeta)\sqrt{n-3}$$

Que se contrastará con la Z_α o la $Z_{\alpha/2}$ de la tabla 1, según sea una hipótesis de una cola o de las dos colas, respectivamente.

Esta transformación Z sirve, se insiste, para tamaños de muestra superior a 50, si bien puede ser valida para n mayor de 25.

Si se desea una más exacta aproximación, se puede utilizar la Z^*

$$Z^* = Z_r - \frac{3Z_r + r}{4n}$$

Cuya varianza viene dada por

$$\sigma^2_{Z^*} = \frac{1}{n-1}$$

Esta es una aproximación válida para muestras de tamaño n<50.

Las diferencias entre Z_r y Z^* son muy pequeñas pero serán importantes para las pruebas que den valores cercanos a los de significación.

Las pruebas para las hipótesis siguientes son

Cola derecha	Cola izquierda	Dos colas
$H_o: \rho \leq 0$	$H_o: \rho \geq 0$	$H_o: \rho = 0$
$H_1: \rho > 0$	$H_1: \rho < 0$	$H_1: \rho \neq 0$

$$Z_o = Z^* \sqrt{n-1}$$

Que se contrastará con la Z_α o la $Z_{\alpha/2}$ de la tabla 1, según sea una hipótesis de una cola o de las dos colas, respectivamente.

Esta transformación Z, al igual que la anterior, servirá, también, para probar cualquier hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_o : Z \leq \zeta$	$H_o : Z \geq \zeta$	$H_o : Z = \zeta$
$H_1 : Z > \zeta$	$H_1 : Z < \zeta$	$H_1 : Z \neq \zeta$

$$Z_o = (Z^* - \zeta)\sqrt{n-1}$$

Es decir, para el caso en que el coeficiente de correlación paramétrico tenga un valor diferente de cero.

Se puede, así mismo, construir límites de confianza para *r* utilizando cualquiera de las dos transformaciones *Z*, según los tamaños de muestra que se tenga.

Primeramente se pasa la *r* muestral a *Z*, se construye el intervalo de confianza para esta *Z* y después se vuelve a transformar estos límites a valores de *r*

Si el tamaño de muestra es superior a 25, el intervalo es

$$L_i = Z_r - \sigma_{Z_r} Z_{(\alpha/2)}$$
$$L^s = Z_r + \sigma_{Z_r} Z_{(\alpha/2)}$$

Si el tamaño de muestra es inferior a 25, el intervalo es

$$L_i = Z^* - \sigma_{Z^*} Z_{(\alpha/2)}$$
$$L^s = Z^* + \sigma_{Z^*} Z_{(\alpha/2)}$$

Ejemplo.-

Resuélvase el ejemplo anterior por medio de la transformación *Z* adecuada. Al tener un tamaño de muestra inferior a 50, se tendría que utilizar la transformación denominada *Z**.

La prueba para contrastar el valor paramétrico ρ=0, es

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0$$
$$r = 0.882$$
$$n = 9$$
$$Z_r = 0.5 \ln\left(\frac{1+0.882}{1-0.882}\right) = 1.3847$$
$$Z^* = 1.3847 - \frac{3 \times 1.3847 + 0.882}{4 \times 9} = 1.2448$$
$$Z_o = 1.2448 \sqrt{9.1} = 3.521^{***}$$
$$Z_{(0.05/2)} = 1.96$$
$$Z_{(0.01/2)} = 2.58$$
$$Z_{(0.001/2)} = 3.33$$

Se rechaza la hipótesis nula, luego el valor paramétrico es significativamente diferente de cero.

Si se tiene que probar la hipótesis de si ρ=0.5 sería de esta manera

$$H_o : \rho = 0.5$$
$$H_1 : \rho \neq 0.5$$
$$r = 0.882; \quad n = 9$$
$$Z_r = 0.5 \ln\left(\frac{1+0.882}{1-0.882}\right) = 1.3847$$
$$\zeta = 0.5 \ln\left(\frac{1+0.5}{1-0.5}\right) = 0.5493$$
$$Z^* = 1.3847 - \frac{3 \times 1.3847 + 0.882}{4 \times 9} = 1.2448$$
$$\zeta^* = 0.5493 - \frac{3 \times 0.5493 + 0.5}{4 \times 9} = 0.4896$$
$$Z_o = (1.2448 - 0.4896) \sqrt{9-1} = 2.136^{*}$$
$$Z_{(0.05/2)} = 1.96$$

El valor de r=0.882 es significativamente diferente de 0.5, al nivel de significación de α=0.05; por lo que es poco probable que la correlación entre el aumento de peso y la cantidad de alimento consumido sea 0.5.

Si se hubiese resuelto por intervalo de confianza, este sería al 95%

$$LC_{(Z')} = 1.2448 \pm \frac{1}{\sqrt{9-1}}\,1.96 = 1.2448 \pm 0.69296$$

$$L_i = 0.5518$$

$$L^s = 1.9378$$

Y transformando estos valores Z en valores de correlación

$$r_i = \frac{2.71827^{2\times0.5518}-1}{2.71827^{2\times0.5518}+1} = 0.5019$$

$$r^s = \frac{2.71827^{2\times1.9378}-1}{2.71827^{2\times1.9378}+1} = 0.9594$$

Como se ve, estos límites del valor autentico del parámetro, no incluyen el cero, ni el 0.5, con este nivel de confianza.

Ejemplo.-

Se continúa con el mismo ejemplo pero suponiendo que se ha obtenido el valor de r=0.882 en una muestra de n=60.

Se puede probar, primero, si ρ=0

$$H_o : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$r = 0.882; \quad n = 60$$

$$Z_r = 0.5 \ln\left(\frac{1+0.882}{1-0.882}\right) = 1.3847$$

$$Z_o = 1.3847\sqrt{60-3} = 10.454 ***$$

$$Z_{(0.05/2)} = 1.96; \; Z_{(0.01/2)} = 2.58; \; Z_{(0.001/2)} = 3.33$$

Se rechaza la hipótesis nula, luego el valor de r es significativamente diferente de cero.

Si, por ejemplo, se tiene que probar la hipótesis de si ρ=0.5 sería de esta manera

$$H_o : \rho = 0.5$$
$$H_1 : \rho \neq 0.5$$
$$r = 0.882; \quad n = 60$$
$$Z_r = 0.5 \ln\left(\frac{1+0.882}{1-0.882}\right) = 1.3847$$
$$\zeta = 0.5 \ln\left(\frac{1+0.5}{1-0.5}\right) = 0.5493$$
$$Z_o = (1.3847 - 0.5493)\sqrt{60-3} = 6.307 \text{ * * *}$$
$$Z_{(0.05/2)} = 1.96; \quad Z_{(0.01/2)} = 2.58; \quad Z_{(0.001/2)} = 3.33$$

El valor de r=0.882 es significativamente diferente de 0.5, por lo que es poco probable que la correlación entre el aumento de peso y el peso del alimento consumido sea 0.5.

Si se desea resolver por los límites de confianza, estos serían, al 95%

$$LC_{(Z')} = 1.3847 \pm \frac{1}{\sqrt{60-3}}\ 1.96 = 1.3847 \pm 0.2596$$
$$L_i = 1.1251$$
$$L^s = 1.6443$$

Y transformando estos valores Z en valores de correlación

$$r_i = \frac{2.71827^{2\times1.1251} - 1}{2.71827^{2\times1.1251} + 1} = 0.8093$$
$$r^s = \frac{2.71827^{2\times1.6443} - 1}{2.71827^{2\times1.6443} + 1} = 0.9281$$

Como se ve, estos límites del valor autentico del parámetro, no incluyen el cero, ni el 0.5, con este nivel de confianza.

Homogeneidad de dos coeficientes de correlación.-

Hay ocasiones en que se miden dos variables en varias muestras, obteniéndose los coeficientes de correlación de cada muestra. En casos como estos puede ser interesante constatar si estos coeficientes de correlación pueden considerarse como estimas de un mismo valor paramétrico (ρ) o son diferentes poblaciones. Una forma de establecer la hipótesis nula es afirmar que los t coeficientes de correlación son homogéneos y calcular un valor paramétrico común, ρ.

Para las pruebas de comparación de coeficientes de correlación, se hace previamente la transformación en el valor de Z adecuado al número de datos, y a partir

de ahí se trata como comparaciones de poblaciones normales.

Por lo tanto, para el caso concreto de la prueba de homogeneidad de dos coeficientes de correlación, se calcula el error típico para la diferencia de dos Z que, tal como se estudió en el Capítulo 4, no es sino la raíz cuadrada de la suma de las varianzas de las dos muestras (puesto que estas son independientes). Con este error típico se puede hacer la prueba de diferencias de valores Z y se contrata con el valor de la tabla de áreas de la curva normal típica (Tabla 1). Puesto que lo que se pretende con esta prueba, corrientemente, es conocer si se puede utilizar las dos estimas para tener una solo estima más precisa, las hipótesis suelen ser de dos colas

$$H_o : \rho_1 = \rho_2$$
$$H_1 : \rho_1 \neq \rho_2$$

La prueba es, si los tamaños de muestra son superiores a 25

$$Z_o = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

En el caso de que los tamaños de muestra sean pequeños

$$Z_o = \frac{Z_1^* - Z_2^*}{\sqrt{\frac{1}{n_1-1} + \frac{1}{n_2-1}}}$$

siendo Z_r y Z^* los definidos en páginas anteriores.

La diferencia de las Z, lógicamente, se distribuye normalmente. Y puesto que se está utilizando una desviación típica paramétrica, se puede contrastar este valor Z_o Zo con $Z_{\alpha/2}$ de la tabla 1.

Ejemplo.-

Se ha estudiado la correlación entre el aumento de peso y el alimento consumido en dos razas de cerdos. Los resultados son

Raza	*n*	*r*	Z_r	Z^*
1	22	0.310	0.32055	0.30610
2	21	0.542	0.60699	0.57886

$$H_o : \rho_1 = \rho_2$$

$$H_1 : \rho_1 \neq \rho_2$$

$$Z_{r_1} = 0.5 \ln\left(\frac{1+0.31}{1-0.31}\right) = 0.32055$$

$$Z_1^* = 0.32055 - \frac{3 \times 0.32055 + 0.310}{4 \times 22} = 0.30610$$

$$Z_{r_2} = 0.5 \ln\left(\frac{1+0.542}{1-0.542}\right) = 0.60699$$

$$Z_2^* = 0.60699 - \frac{3 \times 0.60699 + 0.542}{4 \times 21} = 0.57886$$

$$Z_o = \frac{0.30610 - 0.57886}{\sqrt{\frac{1}{22-1} + \frac{1}{21-1}}} = 0.873 ns$$

$$Z_{(0.05/2)} = 1.96$$

Por lo que no hay suficiente evidencia que induzca a rechazar la hipótesis nula; se puede concluir que ambas correlaciones son estimas de un mismo parámetro.

Ejemplo.-

Supóngase que los tamaños de muestra hubieran sido 62 y 61, respectivamente. En ese caso se tendría

Raza	*n*	*r*	Z_r
1	62	0.310	0.32055
2	61	0.542	0.60699

$$H_o : \rho_1 = \rho_2$$

$$H_1 : \rho_1 \neq \rho_2$$

$$Z_{r_1} = 0.5 \ln\left(\frac{1+0.31}{1-0.31}\right) = 0.32055$$

$$Z_{r_2} = 0.5 \ln\left(\frac{1+0.542}{1-0.542}\right) = 0.60699$$

$$Z_o = \frac{0.32055 - 0.60699}{\sqrt{\frac{1}{62-3} + \frac{1}{61-3}}} = 1.549 ns$$

$$Z_{(0.05/2)} = 1.96$$

Por lo que no hay suficiente evidencia que induzca a rechazar la hipótesis nula; se puede concluir que ambas correlaciones son estimas de un mismo parámetro.

Homogeneidad de varios coeficientes de correlación.-

También se puede presentar el caso, más general, de probar la homogeneidad de varios coeficientes de correlación y obtener, en su caso, un solo coeficiente de correlación si resultan ser homogéneos. Por ejemplo, se puede tener mediciones de dos características de un cultivo o de una especie animal para varias cepas o razas. Las varianzas de estas cepas o razas pueden ser no homogéneas, así que la combinación de las sumas de productos y el cálculo de un solo coeficiente de correlación no sería válido.

Los cálculos son muy sencillos y de hecho consisten en el cálculo de una suma de cuadrados ponderada de los valores de Z_r o de Z^*, correspondientes, de los coeficientes de correlación.

Esta Z ponderada es, para el caso de tamaño de muestra superior a 25

$$\overline{Z} = \frac{\Sigma_i\,(n_i-3)Z_{ri}}{\Sigma_i(n_i-3)}$$

Y para tamaño de muestra inferior a 25

$$\overline{Z} = \frac{\Sigma_i(n_i-1)Z_i^*}{\Sigma_i(n_i-1)}$$

Si se observa esta fórmula se comprueba que es como la fórmula definición de la distribución θ^2, por lo que la, anteriormente citada, suma de cuadrados ponderada se distribuye mediante una θ^2 y viene dada por

$$\chi^2 = \Sigma_i\left(\frac{Z_{ri}-\overline{Z}}{\frac{1}{\sqrt{n_i-3}}}\right)^2 = \Sigma_i(n_i-3)(Z_{ri}-\overline{Z})^2$$

El denominador del paréntesis es la desviación típica paramétrica de Z, σ_Z.

Como las desviaciones típicas son de t-1 coeficientes de correlación son independientes, se tendrán $gl = t$-1.

Ejemplo.-

Se ha estudiado la correlación entre el aumento de peso y el alimento consumido en tres razas de cerdos.

Raza	*n*	*r*	Z_{ri}	$Z_{ri}-\overline{Z}$	$(n_i-3)(Z_{ri}-\overline{Z})$
1	50	0.362	0.3792	-0.0913	0.392
2	50	0.419	0.4465	-0.0243	0.028
3	50	0.527	0.5860	+0.1157	0.629
				$\chi^2=$	1.049

$$\overline{Z} = \frac{47\times0.3792+47\times0.4465+47\times0.5860}{47\times3} = 0.4703$$

$$\chi^2 = 1.049ns$$

$$\chi^2_{(2;\,0.05/2)} = 5.9915$$

Por lo que no se tiene evidencia suficiente para rechazar la hipótesis nula de homogeneidad de coeficientes de correlación.

Se puede, por tanto, utilizar un valor único de correlación entre estas dos variables, que sería un *r* ponderado que no es otro que el correspondiente a $\overline{Z}$, haciendo la transformación inversa

$$r = \frac{2.71828^{2\times0.4703}-1}{2.71828^{2\times0.4703}+1} = 0.4384$$

Pudiéndose, ahora, establecer límites de confianza en torno a este coeficiente *r*.

Correlación intraclase o repetibilidad.-

Esta medida de correlación se ideó para indicar el grado de asociación o similitud entre individuos dentro de clases o grupos, de ahí su nombre.

En ocasiones se desea un coeficiente de correlación sin que se tenga un criterio para asignar las diferentes observaciones a un miembro o a otro del par, o simplemente se desea saber la correlación entre individuos que no constituidos en parejas. Esto puede ser así cuando se pretende medir la correlación de un carácter (una única variable) en parientes, en hermanos, en gemelos o en el mismo individuo. Por ejemplo, considérese el problema de medir la correlación entre la estatura de hermanos; debido a que todo lo que se desea es una medida de similitud entre estaturas de hermanos, cualquier intento de indicar a unos como *X* y a otros como *Y* (como puede ser, por ejemplo, por edad) sería introducir un elemento adulterado en la correlación. Este elemento adulterado, por supuesto, sería que una correlación ordinaria (lineal simple) mediría la correlación entre las estaturas de los hermanos mayores y menores, más bien, que simplemente indicar la similitud de estatura de hermanos.

En casos como estos, se puede obtener un valor del coeficiente a partir del

cálculo de ciertas varianzas. Al coeficiente resultante se le denomina correlación intraclase (r_I) y se calcula mediante la fórmula

$$r_I = \frac{CM_{Familias} - CM_{error}}{CM_{Familias} + (n-1)CM_{error}}$$

Si el experimento es desequilibrado, el cálculo es

$$r_I = \frac{CM_{Familias} - CM_{error}}{CM_{Familias} + \frac{N^2 - \Sigma_i \, n_i^2}{N(t\text{-}1)} - 1 CM_{error}}$$

Que es el resultado de la definición de parámetro

$$\rho_I = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$$

siendo σ^2 y σ_T^2 estimaciones de las correspondientes componentes de la varianza de un *ANOVA* de una vía.

Recuérdese que el modelo de este *ANOVA* es

$$X_{ij} = \mu + T_i + e_{ij}$$

Y que las estimas de cada componente son

$$\mathrm{E}\, CM_T = n\,\sigma_T^2 + \sigma^2$$
$$\mathrm{E}\, CM_e = \sigma^2$$

Pues estas estimas en términos de coeficiente de correlación intraclase, son

$$\mathrm{E}CM_T = \sigma^2 (1 - \rho_I)$$
$$\mathrm{E}CM_e = \sigma^2 \, [1 + (n-1)\rho_I]$$

Si este término es significativo, la correlación intraclase es un medida del parecido entre parientes.

Considérese, por ejemplo, el agrupamiento de los individuos en familias de hermanos carnales; la componente debida a la *familia* es la varianza de los individuos con respecto a la media de la familia a la que pertenezcan y la componente del *error* o *entre familias* es la varianza de las medias de las familias con respecto a la media de la población. El parecido entre parientes, es decir, entre hermanos, en este caso, puede verse, bien como la similitud existente entre individuos de la misma familia bien como la diferencia existente entre individuos de diferentes familias. Mientras mayor sea la similitud dentro de las familias, mayor será la proporción de diferencia que exista entre

las familias. El grado de parecido puede por lo tanto expresarse como la componente *familia* en proporción a la varianza total, es decir, el coeficiente de correlación intraclase.

Ejemplo.-

Se quiere saber el grado de parecido en base al número de crestas digitales en ambas manos de 12 pares de gemelos homozigotos femeninos

Familia											
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
71	79	105	115	76	83	114	57	114	94	75	76
71	82	99	114	70	82	113	44	113	91	83	72

El análisis de la varianza es

FV	*gl*	*SC*	*CM*	*F*	*ECM*
Familias	11	8990.458	817.3144	57.19**	$\sigma^2+2\sigma_F$
Error	12	171.500	14.2917		σ^2
Total	23	9161.958			

$$\sigma^2 = 1.42917$$

$$\sigma_F^2 = \frac{8.173144 - 1.42917}{2} = 4.015114$$

La correlación intraclase sería

$$r_I = \frac{8.173144 - 1.42917}{8.173144 + 1 \times 1.42917} = 0.9656$$

O bien

$$r_I = \frac{4.015114}{4.015114 + 1.42917} = 0.9656$$

La significación de la F_o indica que esta correlación es significativamente diferente de cero, con lo que se concluye que el número de crestas digitales es casi el mismo para los dos miembros de cada par de gemelos, pero diferente entre pares.

Esta misma correlación se puede hallar para familias de hermanos no gemelos, y por lo tanto, para familias de tamaños diferentes y números de hermanos diferentes de dos.

Pruebas de asociación no paramétricas.-

Muchas veces una población bivariante se sabe que no se distribuye normalmente, por lo que en este caso no tiene sentido el cálculo de *r* como una estima del parámetro ρ. En algunos casos la transformación de las variables (ver el epígrafe *Transformaciones* y siguientes del Capítulo 6) puede acercar su distribución conjunta a la distribución normal bivariante, haciendo posible la estima de ρ en la nueva escala. En la prueba de hipótesis nula (no correlación) la *r* puede utilizarse siempre que una de las variables sea normal. Si esto falla, no se puede expresar el grado de correlación de datos no normales por medio de un parámetro como ρ.

Sin embargo, tal vez se necesite examinar si dos variables son independientes, o si varían en la misma o en opuesta dirección. Si ninguna de las dos variables es normal, el procedimiento mejor conocido es aquél en el que ambas variables están ordenadas y se trabaja con estas ordenaciones. Por ejemplo, se quiere saber si las valoraciones que hacen dos evaluadores de los sementales, por su estampa morfológica, existe algún grado de concordancia o no. Problemas como estos lo que pretenden, en definitiva, es conocer si los sementales valorados en los primeros lugares son los mismos en los dos evaluadores, así como los peores y los intermedios.

Coeficiente de correlación de rangos de *Spearman*.-

Es el más antiguo y conocido de los coeficientes de correlación de clasificaciones u ordenaciones. Es una medida de la correlación entre dos clasificaciones y viene dado por la expresión

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

siendo

d_i la diferencia entre las dos clasificaciones de la unidad *i-ésima.*

n es el tamaño de muestra.

Al igual que *r*, la correlación de rangos puede oscilar entre +1, cuando las dos clasificaciones son idénticas; hasta -1, cuando las clasificaciones son tan diferentes como es posible, es decir, cuando la unidad clasificada en lo más alto en una ocasión se clasifica en lo más bajo en la otra ocasión, la siguiente más alta sería la segunda más baja y así sucesivamente.

Este coeficiente de correlación puede calcularse con la fórmula de la correlación ya conocida, pues de hecho, la formula anterior procede del coeficiente de correlación simple. Es decir, la fórmula de este coeficiente de correlación es una manera fácil de calcular el coeficiente de correlación de *Pearson* pero con los ordinales o clasificaciones en lugar de con los valores originales.

Para realizar la prueba de hipótesis H_o: $r_s=0$, para un tamaño de muestra $n<10$ pares de datos, los niveles de significación vienen dados en la siguiente tabla

Tamaño muestra	α=0.05	α=0.01
4 *o menos*	*ninguno*	*ninguno*
5	1.000	*ninguno*
6	0.886	1.000
7	0.750	0.893
8	0.714	0.857
9	0.683	0.833
10	0.648	0.794
11 *o más*	*utilizar como prueba para la r*	

Para n>10 se realiza la misma prueba que si fuera la *r*.

Coeficiente de correlación de clasificación de *Kendall*.-

Otra medida del grado de correlación o de concordancia, muy semejante a r_s es la τ (*tau*) de *Kendall*. Puesto que la distribución del muestreo de esta medida, τ, es conocida resulta generalmente más satisfactoria que r_s para fines de inferencias. Además en el caso de no nulidad, es decir, cuando se ha confirmado la hipótesis nula de τ=0, tiene una interpretación en función de la probabilidad de que dos unidades seleccionadas al azar tienen las mismas clasificaciones en *a* y en *b*.

La τ se define como

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{2S}{n(n-1)}$$

siendo

n es el número de unidades clasificadas
S=*P*-*Q*
P el número de pares que tienen el mismo ordinal en las dos clasificaciones
Q es el número de pares para los cuales los ordinales no concuerdan.

De la fórmula se desprende que el coeficiente, τ, es la diferencia entre *P* y *Q* expresado como una fracción de $n(n-1)/2$ que es el número total de formas en las cuales las *n* unidades pueden ser comparadas dos a dos a la vez.

Para la obtención de los valores de *P* y *Q* hay que reacomodar las clasificaciones de manera que una de ellas esté en orden creciente para en la segunda clasificación contar para cada uno de las *i* unidades el número de unidades (a la derecha o abajo) que tienen una clasificación mayor que ésta (P_i) y el número de unidades que tienen una clasificación inferior a ésta (Q_i); sumando al final todas las *P* y las *Q* para calcular *S* y sustituir en la fórmula de la correlación τ.

Para la significación de τ, si n<10 se puede usar la tabla anterior y para n>10 se

puede considerar que τ se distribuye normalmente con media cero y varianza igual a

$$\sigma^2 = \frac{4n+10}{9n(n-1)}$$

por lo que la raíz cuadrada de esta varianza será el *error típico* de τ.

Al hacer la prueba de significación es esencial, sin embargo, hacer una corrección por continuidad consistente en restarle uno al valor absoluto de S para calcular la τ ajustada. Si este valor, τ_{adj}, se divide por el *error típico* se tendrá un valor de Z_o que se podrá contrastar con el valor de la tabla Z (Tabla 1) para el nivel de significación que se considere adecuado. Se tiene, por tanto

$$T_{adj} = \frac{S}{\frac{n(n-1)}{2}} = \frac{2(S-1)}{n(n-1)}$$

$$\sigma = = \sqrt{\frac{4n+10}{9n(n-1)}}$$

$$Z_o = \frac{T_{adj}}{\sigma}$$

Las cantidades r_s y τ pueden usarse como medida de la habilidad para apreciar o detectar valores en unidades con finalidad clasificatoria. Por ejemplo, se tiene un determinado grupo de aspirantes a catadores, se le podría dar a cada uno varios frascos conteniendo aceite con cuatro grados de acidez diferentes e indicarles que coloquen los frascos en orden, de acuerdo con el grado de acidez. Si X_1 representa la clasificación correcta de los grados de acidez y X_2 la clasificación dada por un aspirante, el valor de r_s o de τ para esta persona medirá el éxito en su oficio. De los resultados obtenidos con una muestra de mujeres y hombres podríamos investigar si las mujeres son mejores (o no) que los hombres en este menester. La diferencia entre τ o r_s para mujeres y hombres podría compararse aproximadamente por medio de una prueba *t*.

Ejemplo.-

Se tiene la clasificación de siete ovejas, hecha por dos observadores diferentes, sobre su condición corporal después de tres semanas bajo una dieta deficiente

Oveja	*Clasificación*		*diferencia*	
nº	*a*	*b*	*d*	d^2
1	4	4	0	0
2	1	2	-1	1
3	6	5	1	1
4	5	6	-1	1
5	3	1	2	4
6	2	3	-1	1

7	7	7	0	0
		Σ	0	8

$$r_s = 1 - \frac{6 \times 8}{7(19-1)} = 0.857$$

Como $n<10$ la significación se vería comparando con la tabla de significación expuesta un par de páginas atrás. Si $n>10$ el contraste de hipótesis sería

$$H_0 : r_s = 0$$
$$H_1 : r_s \neq 0$$
$$t = 0.857 \sqrt{\frac{7-2}{1-0.875^2}} = 3.72086 *$$
$$t_{(5;\, 0.05/2)} = 2.576$$

La clasificación hecha por ambos observadores se puede considerar como la misma con un α=0.05 de confianza.

Para el cálculo de τ de *Kendall* se reordena la clasificación con respecto al primer observador, por ejemplo

Oveja	*Clasificación*			
nº	*a*	*b*	*P*	*Q*
2	1	2	5	1
6	2	3	4	1
5	3	1	4	0
1	4	4	3	0
4	5	6	1	1
3	6	5	1	0
7	7	7	0	0
		Σ	18	3

$$S = 18 - 3 = 15$$
$$\tau = \frac{2 \times 15}{7(7-1)} = 0.7143$$

Como $n<10$ la significación se vería comparando este valor con los de la tabla expuesta tres páginas anterior. Si $n>10$ el contraste de hipótesis sería

$$\sigma^2 = \frac{4\times7+10}{9\times7(7-1)} = 0.10053$$

$$\sigma = 0.31706$$

$$\tau_{adj} = \frac{2\times14}{7\times6} = 0.66667$$

$$Z_o = \frac{0.66667}{0.31706} = 2.1027\ *$$

$$Z_{0.05/2} = 1.96$$

La clasificación hecha por ambos observadores se puede considerar como la misma con un α=0.05 de confianza.

Archivo del programa *SAS* (C15-2.SAS).-

```
title ' Correlaciones no paramétricas';
option ls=75 ps=60;
data corrnp;
infile 'c15-2.dat';
input rata cla1 cla2;
proc corr spearman kendall;
var cla1 cla2;
run;
```

Archivo de datos (C15-2.DAT).-

```
   1       4       4
   2       1       2
   3       6       5
   4       5       6
   5       3       1
   6       2       3
   7       7       7
```

Archivo de resultados (C15-2.LST).-

```
                    Correlaciones no paramétricas
                        Correlation Analysis
                      2 'VAR' Variables:  CLA1      CLA2

                             Simple Statistics
Variable            N          Mean       Std Dev        Median       Minimum       Maximum
CLA1                7        4.0000        2.1602        4.0000        1.0000        7.0000
CLA2                7        4.0000        2.1602        4.0000        1.0000        7.0000

 Spearman Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 7
                                  CLA1                  CLA2
                     CLA1         1.00000               0.85714
                                   0.0                   0.0137
                     CLA2         0.85714               1.00000
                                   0.0137                0.0

Kendall Tau b Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 7
                                  CLA1                  CLA2
                     CLA1         1.00000               0.71429
                                   0.0                   0.0243
                     CLA2         0.71429               1.00000
                                   0.0243                0.0
```

Coeficiente de correlación serial.-

Este coeficiente de correlación se usa para probar si existe independencia de errores en los datos de una muestra. Es decir, para probar si el valor de un dato viene influenciado o no por el valor del dato tomado inmediatamente antes.

Para el estudio de este coeficiente de correlación consúltese el Capítulo 6.

Bibliografía

Dixon, J.D. y Manssey, F.J. 1974. INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO. Ed. Del Castillo. Madrid.

Gilbert, N. 1976. ESTADÍSTICA. Ed. Interamericana. México.

Lite, TM, y Jackson Hills, F. 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed TRILLAS. México.

Mills, F.S. 1969. MÉTODOS ESTADÍSTICOS. Ed. Aguilar. Madrid.

Milton, J.S. 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Spiegel M.R. 1990. ESTADÍSTICA. Ed. McGraw-Hill. Madrid.

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

SAS Institute Inc. 1990. *SAS/STAT* USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.

SAS Institute Inc. 1990. *SAS PROCEDURE* GUIDE. Cary, NC, USA.

CAPÍTULO 16

Correlación multivariante

CAPÍTULO 16

Correlación multivariante

Correlación parcial.-

Como se ha visto en el capítulo anterior, el coeficiente de correlación simple está basado en la suposición de la aproximación a la distribución *normal bivariante*. Si se tiene más de dos variables, el modelo básico para la correlación múltiple, sería una ampliación de esta distribución, denominada distribución *normal multivariante*.

Si hay tres variables, habrá tres correlaciones simples entre ellas, ρ_{12}, ρ_{13} y ρ_{23}. Estos coeficientes miden la relación lineal que existen entre estas variables, dos a dos, sin tener en cuenta la posible influencia de la tercera.

La correlación parcial se define como la correlación entre dos variables si las demás variable no varían, es decir, el valor de las demás variables son fijos. Por ejemplo, el coeficiente de correlación parcial $\rho_{12.3}$, es la correlación entre la variable 1 y 2 siendo constante el valor de la variable 3; o el coeficiente de correlación parcial $\rho_{23.1}$ es la correlación entre la variable 2 y 3 siendo constante el valor de la variable 1.

El mantener constante una variable puede hacerse experimentalmente o estadísticamente, debiendo dar en ambos casos resultados equivalentes. Para ver claro el porqué se necesita hallar una correlación haciendo constante el valor de otra u otras variables supóngase que se está interesado en conocer la correlación entre la longitud del brazo y de la pierna cuando el tamaño total del organismo permanece constante. Está claro que la longitud del brazo y de la pierna estarán altamente correlacionados debido al tamaño general; así, un individuo alto tendrá brazos y piernas largos, mientras que un individuo bajo tendrá extremidades cortas. Sin embargo, si este estudio se seleccionan individuos del mismo tamaño se puede esperar que exista alguna correlación residual entre la longitud del brazo y de la pierna. Esto es muy probable en vertebrados, debido a que ambas extremidades están determinadas embriológicamente con mecanismos homólogos responsables de la diferenciación y determinación. Por tanto existirá alguna correlación entre éstas dos longitudes, incluso en ausencia de una causa común como es el tamaño del individuo.

Si una correlación significativa entre dos variables se convierte en correlación parcial no significativa cuando una tercera variable permanece constante, esto sugiere, aunque no prueba, que la variable que permanece constante es la causa común de la correlación de las otras dos.

Correlación parcial o correlación de lo residuos.-

La correlación parcial $r_{12.3}$, sería la correlación lineal entre la variable 1 y 2 dejando como constante la variable 3. Esto quiere decir que hay que medir la correlación entre la variable 1 y 2 que no sea un reflejo de sus relaciones con la variable 3. Por tanto, se puede obtener una estima muestral $r_{12.3}$ calculando la desviación o residuo e_{13}, de la regresión de la variable 1 sobre la variable 3, y la desviación o residuo e_{23}, de la regresión de la variable 2 sobre la variable 3. Y $r_{12.3}$ es el coeficiente de correlación simple entre e_{13} y e_{23}.

Ejemplo.-

Se tiene el rendimiento (X_1) de una línea de trigo observado en 11 años sucesivos y los datos meteorológicos correspondientes: precipitación en Noviembre y Diciembre (X_2), temperatura media en Julio (X_2), precipitación en Julio (X_3) y radiación solar en Julio (X_4).

¿Cuál sería la correlación entre la producción y la precipitación en Noviembre si la temperatura media de Julio hubiera sido la misma todos los años?

Archivo del programa *SAS* (C16-1.SAS).-

```
title 'Correlación entre los residuos de las dos rectas de regresión';
options ls=75 ps=60;
data corpar;
infile 'c16-1.dat';
input prod preNov tmjul preJul radJul;
Proc reg noprint;
  Model prod = tmjul;
   output out=residuos R=rprod;
run;
proc reg noprint;
  model prenov = tmjul;
   output out=residuos R=rprenov;
run;
proc corr;
var rprod rprenov ;
run;
```

Archivo de datos (C16-1.DAT).-

```
 87.9  19.6   1.0  1661  28.37
 89.9  15.2  90.1   986  23.77
153.0  19.7  56.6  1353  26.04
132.1  17.0  91.0  1293  25.74
 88.8  18.3  93.7  1153  26.68
220.9  17.8 106.9  1286  14.29
117.7  17.8  65.5  1104  28.00
109.0  18.3  41.8  1574  28.37
```

```
156.1  17.8  57.4  1222  24.96
181.5  16.8 140.6   902  21.66
181.4  17.0  74.3  1150  24.37
```

Archivo de resultados (C16-1.LST).-

```
      Correlación entre los residuos de las dos rectas de regresión
                          Correlation Analysis
                    2 'VAR' Variables:  RPROD    RPRENOV

                           Simple Statistics
 Variable                 N            Mean         Std Dev             Sum
 RPROD                   11               0       38.712229               0
 RPRENOV                 11               0        1.000125               0

                    Simple Statistics
 Variable          Minimum          Maximum      Label
 RPROD         -60.657747        63.606063      Residual
 RPRENOV        -2.214239         1.557521      Residual

 Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 11
                                   RPROD          RPRENOV
               RPROD              1.00000          0.32184
               Residual            0.0              0.3345
               RPRENOV            0.32184          1.00000
               Residual            0.3345           0.0
```

La correlación de los dos residuos es 0.3218, que como se verá en el siguiente ejemplo, es la correlación parcial de la producción con la precipitación en Noviembre si la temperatura media de Julio se hubiera mantenido constante.

Cálculo de la correlación parcial.-

Puede demostrarse que $r_{12.3}$ satisface la siguiente fórmula

$$R_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

El error típico de esta estima es

$$S_{r_{12.3}} = \sqrt{\frac{1-r_{12.3}^2}{n-3}}$$

Por lo que podemos probar H_0: $r_{12.3} = 0$ por medio de la t

$$t = \frac{r_{12.3}\sqrt{n-3}}{\sqrt{1-r_{12.3}^2}}$$

Que se contrasta con la $t_{(n-3;\, \alpha/2)}$.

De la misma manera se puede hallar la regresión entre la variable 1 y la variable 3 dejando constante la variable 2; o la correlación entre la variable 2 y la variable 3

dejando constante la variable 1.

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}; \quad S_{r_{13.2}} = \sqrt{\frac{1-r_{13.2}^2}{n-3}}; \quad t = \frac{r_{13.2}\sqrt{n-3}}{\sqrt{1-r_{13.2}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}; \quad S_{r_{23.1}} = \sqrt{\frac{1-r_{23.1}^2}{n-3}}; \quad t = \frac{r_{23.1}\sqrt{n-3}}{\sqrt{1-r_{23.1}^2}}$$

El coeficiente de correlación simple entre dos variables se le puede denominar *coeficiente de orden cero* y se simboliza por medio de una *r* con dos subíndices que hacen referencia a las variables de las que se está hallando la correlación. Los coeficientes de correlación parcial que se refieren a la correlación de dos variables dejando fija una tercera se denominan *coeficientes de primer orden* y se representan con la *r* con tres subíndices, los dos primeros separados del tercero por un punto, es decir, los dos primeros hacen referencia a las variables para las que se ha hallado la correlación y el tercero la variable que se ha hecho constante. De forma análoga se puede obtener coeficientes de *segundo*, *tercer*, *cuarto* o *n-ésimo* orden, dependiendo del número de variables que se mantienen constantes mientras se mide la correlación entre dos variables.

Los coeficientes de correlación parcial de un orden determinado pueden deducirse partiendo de los de orden inmediatamente inferior. Así, se puede obtener un coeficiente de primer orden aplicando la relación, ya conocida, de los de orden cero

$$R_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Para un coeficiente de segundo orden la relación es con coeficientes de primer orden, esta es

$$R_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}}$$

Y una correlación parcial de orden *k-ésimo*

$$R_{12.34...k} = \frac{r_{12.34..(k\text{-}1)} - r_{1k.34...(k\text{-}1)}\, r_{2k.34...(k\text{-}1)}}{\sqrt{(1-r_{1k.34...(k\text{-}1)}^2)(1-r_{2k.34...(k\text{-}1)}^2)}}$$

La prueba $\underline{t}$ para contrastar si esta *r* es diferente de cero, es

$$t = \frac{r_{12.34...k}\sqrt{n-k}}{\sqrt{1-r_{12.34..k}^2}}$$

que se contrasta con la $t_{(n\text{-}k;\ \alpha/2)}$.

Por tanto, es posible, partiendo de los coeficientes de correlación de orden cero, calcular sucesivamente todos los coeficientes de orden más elevado.

Ejemplo.-

Siguiendo con el ejemplo del trigo calcúlese, a modo de ejemplo: (a) las correlaciones simples entre todas las variables; (b) correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio fuera constante; (c) la correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio y la precipitación de Julio fueran constantes; (d) la correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio, la precipitación de Julio y la radiación solar de Julio fueran constantes.

Archivo del programa *SAS* (C16-2.SAS).-

```
title 'Correlaciones parciales';
Options ls=75 ps=60;
Data corpar;
Infile 'C16-1.dat';
Input prod preNov tmjul preJul radJul;
Proc corr;
run;
title 'Correlación Producción-Precipitación Noviembre, fija tm Julio';
proc corr;
var prod prenov;
partial tmjul;
run;
title 'Correlación Producción-Precipitación Noviembre, fija tm y Pre
Julio';
proc corr;
var prod prenov;
partial tmjul preJul;
run;
title 'Correlación Producción-Precipitación Noviembre, fija tm, pre y Rad
Julio';
proc corr;
var prod prenov;
partial tmjul preJul radJul;
run;
```

Archivo de resultados (C16-2.LST).-

Correlaciones parciales

5 'VAR' Variables: PROD PRENOV TMJUL PREJUL RADJUL

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PROD	11	138.0273	44.4218	1518	87.9000	220.9000
PRENOV	11	17.7545	1.2793	195.3000	15.2000	19.7000
TMJUL	11	74.4455	36.6996	818.9000	1.0000	140.6000
PREJUL	11	1244	227.9395	13684	902.0000	1661
RADJUL	11	24.7500	4.0404	272.2500	14.2900	28.3700

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 11

	PROD	PRENOV	TMJUL	PREJUL	RADJUL
PROD	1.00000	-0.08658	0.49045	-0.24947	-0.78050
	0.0	0.8002	0.1256	0.4594	0.0046
PRENOV	-0.08658	1.00000	-0.62360	0.72905	0.32679
	0.8002	0.0	0.0404	0.0109	0.3267
TMJUL	0.49045	-0.62360	1.00000	-0.81223	-0.63754
	0.1256	0.0404	0.0	0.0024	0.0348
PREJUL	-0.24947	0.72905	-0.81223	1.00000	0.34355
	0.4594	0.0109	0.0024	0.0	0.3009
RADJUL	-0.78050	0.32679	-0.63754	0.34355	1.00000
	0.0046	0.3267	0.0348	0.3009	0.0

Correlación Producción-Precipitación Noviembre, fija tm Julio

Correlation Analysis

1 'PARTIAL' Variables: TMJUL

2 'VAR' Variables: PROD PRENOV

Simple Statistics

Variable	N	Mean	Std Dev	Sum
TMJUL	11	74.445455	36.699574	818.900000
PROD	11	138.027273	44.421821	1518.300000
PRENOV	11	17.754545	1.279346	195.300000

Variable	Minimum	Maximum	Partial Variance	Partial Std Dev
TMJUL	1.000000	140.600000	.	.
PROD	87.900000	220.900000	1665.151848	40.806272
PRENOV	15.200000	19.700000	1.111390	1.054225

Pearson Partial Correlation Coefficients / Prob > |R| under Ho: Partial Rho=0 / N = 11

	PROD	PRENOV
PROD	1.00000	0.32184
	0.0	0.3645
PRENOV	0.32184	1.00000
	0.3645	0.0

Correlación Producción-Precipitación Noviembre, fija tm y Pre Julio

Correlation Analysis

2 'PARTIAL' Variables: TMJUL PREJUL

2 'VAR' Variables: PROD PRENOV

Simple Statistics

Variable	N	Mean	Std Dev	Sum
TMJUL	11	74.445455	36.699574	818.900000
PREJUL	11	1244.000000	227.939466	13684
PROD	11	138.027273	44.421821	1518.300000
PRENOV	11	17.754545	1.279346	195.300000

Variable	Minimum	Maximum	Partial Variance	Partial Std Dev
TMJUL	1.000000	140.600000	.	.
PREJUL	902.000000	1661.000000	.	.
PROD	87.900000	220.900000	1712.596716	41.383532
PRENOV	15.200000	19.700000	0.952545	0.975984

```
            Pearson Partial Correlation Coefficients
            / Prob > |R| under Ho: Partial Rho=0 / N = 11
                                  PROD            PRENOV
              PROD              1.00000           0.21437
                                 0.0               0.5797
              PRENOV            0.21437           1.00000
                                 0.5797            0.0

Correlación Producción-Precipitación Noviembre, fija tm, pre y Rad Julio
                        Correlation Analysis
            3 'PARTIAL' Variables:  TMJUL    PREJUL    RADJUL
            2 'VAR'     Variables:  PROD     PRENOV

                         Simple Statistics

Variable                N            Mean         Std Dev             Sum
TMJUL                  11       74.445455       36.699574       818.900000
PREJUL                 11     1244.000000      227.939466           13684
RADJUL                 11       24.750000        4.040408       272.250000
PROD                   11      138.027273       44.421821      1518.300000
PRENOV                 11       17.754545        1.279346       195.300000

                                                   Partial         Partial
Variable         Minimum         Maximum          Variance        Std Dev
TMJUL           1.000000      140.600000                 .               .
PREJUL        902.000000     1661.000000                 .               .
RADJUL         14.290000       28.370000                 .               .
PROD           87.900000      220.900000      1100.163972       33.168720
PRENOV         15.200000       19.700000         1.079971        1.039217

            Pearson Partial Correlation Coefficients
            / Prob > |R| under Ho: Partial Rho=0 / N = 11
                                  PROD            PRENOV
              PROD              1.00000           0.36608
                                 0.0               0.3725
              PRENOV            0.36608           1.00000
                                 0.3725            0.0
```

En el archivo de salida (**C16-2.LST**) se observa que en la salida del primer procedimiento la correlación simple entre la producción y la precipitación en Noviembre es negativa, –0.08658, mientras que en la salida del segundo procedimiento, esta correlación es positiva, 0.32184, al igual que las siguientes correlaciones parciales entre la producción y la precipitación en noviembre. No llega a ser significativa ninguna, tal vez por el pequeño tamaño de muestra, pero está claro que se pasa de interpretar que la hay una correlación negativa de la producción con la precipitación de noviembre, si bien muy pequeña, a que hay una correlación positiva, alta, de la producción con la precipitación de noviembre, valor lo suficientemente alto como para que de significativo con una muestra más grande.

Se comprueba también que el valor de esta correlación parcial es el mismo que el de la correlación simple entre los residuos del ejemplo anterior.

Correlación múltiple.-

Como ya se ha afirmado, la correlación parcial no involucra la noción de variables independientes y dependientes sino que es una medida de interdependencia. Por otro lado, el coeficiente de correlación múltiple se aplica a la situación en que una variable, a la que se puede seguir llamando *Y*, ha sido aislada para examinar su

relación con el conjunto de las otras variables. Este coeficiente de correlación viene determinado por la expresión

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

Al igual que ocurría con el coeficiente de correlación simple, R^2 es el ***coeficiente de determinación múltiple***.

El valor de un coeficiente de correlación múltiple, R, se encuentra entre cero y uno. Cuanto más se acerque a uno mayor es el grado de asociación entre las variables. Y cuanto más se acerca a 0 la relación lineal es peor.

Existe una relación entre el coeficiente de correlación múltiple y los diferentes coeficientes de correlación parcial, que puede facilitar el cálculo de aquél, esta es

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$$
$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)$$

Una correlación múltiple de orden *k-ésimo*

$$1 - R_{1.23...k}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)...(- r_{1k.23...(k-1)}^2)$$

Como se ve, la generalización de todas estas fórmulas para k variables es automática.

Así como en la prueba de ajuste de una regresión múltiple, R^2 es la fracción de la suma de cuadrados de las desviaciones de Y de su media, atribuible a la regresión, en tanto que $(1\text{-}R^2)$ es la fracción no asociada a la regresión; ahora, la prueba de hipótesis nula, de que la correlación múltiple en la población es cero, es idéntica a la prueba F de la hipótesis nula que $\beta_1=\beta_2=...=\beta_k=0$; y ésta es

$$F = \frac{(n-k)\, R^2}{(k-1)\,(1-R^2)}$$

Siendo R el coeficiente de determinación múltiple. Esta F_o se contrasta con la $F_{(k-1,n-k;\alpha)}$.

Ejemplo.-

Siguiendo con el ejemplo anterior, el coeficiente de correlación múltiple de la variable producción con las variables: precipitación en Noviembre, temperatura media en Julio, precipitación en Julio y radiación en Julio

Archivo del programa *SAS* (C16-3.SAS).-

```
title 'Correlación múltiple';
Options ls=75 ps=60;
Data cormul;
Infile 'C16-1.dat';
Input prod preNov tmjul preJul radJul;
title 'Correlación múltiple por el procedimiento correlaciones canónicas';
Proc cancorr;
Var prod;
With preNov tmJul preJul radJul;
run;
title 'Correlación múltiple, raíz cuadrada del coeficiente de determinación
de la regresión ';
proc reg;
model prod = preNov tmJul preJul radJul;
run;
```

Archivo de resultados (C16-3.DAT).-

```
   Correlación múltiple por el procedimiento correlaciones canónicas
                    Canonical Correlation Analysis

                                 Adjusted       Approx       Squared
                      Canonical  Canonical      Standard     Canonical
                      Correlation Correlation    Error        Correlation
        1              0.813654    0.765099     0.106874      0.662034

                                Eigenvalues of INV(E)*H
                                  = CanRsq/(1-CanRsq)
                      Eigenvalue  Difference   Proportion    Cumulative
        1              1.9589          .         1.0000        1.0000

                    Test of H0: The canonical correlations in the
                      current row and all that follow are zero
                   Likelihood
                     Ratio       Approx F     Num DF      Den DF    Pr > F
        1         0.33796638      2.9383          4           6     0.1153

                        NOTE: The F statistic is exact.
                    Multivariate Statistics and Exact F Statistics
                                 S=1    M=1    N=2

Statistic                      Value          F       Num DF    Den DF  Pr > F
Wilks' Lambda               0.33796638     2.9383          4         6  0.1153
Pillai's Trace              0.66203362     2.9383          4         6  0.1153
Hotelling-Lawley Trace      1.95887416     2.9383          4         6  0.1153
Roy's Greatest Root         1.95887416     2.9383          4         6  0.1153

Correlación múltiple, raíz cuadrada del coeficiente de determinación de la
regresión
Model: MODEL1
Dependent Variable: PROD

                              Analysis of Variance

                                 Sum of          Mean
  Source              DF        Squares        Square      F Value       Prob>F
  Model                4    13063.89732    3265.97433        2.938       0.1153
  Error                6     6669.08450    1111.51408
  C Total             10    19732.98182

       Root MSE       33.33938      R-square       0.6620
       Dep Mean      138.02727      Adj R-sq       0.4367
       C.V.           24.15420
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	194.400955	256.79114975	0.757	0.4777
PRENOV	1	11.684177	12.12557742	0.964	0.3725
TMJUL	1	0.034548	0.65127765	0.053	0.9594
PREJUL	1	-0.037992	0.09826225	-0.387	0.7124
RADJUL	1	-8.853754	3.68913828	-2.400	0.0533

Como se ve, la correlación múltiple es un caso particular de la correlación canónica en la que hay una sola variable (**VAR**).

En el archivo de resultado se observa que la correlación múltiple vale 0.8136, y la ***F*** de la prueba de significación vale 2.938, que es no significativa (volvemos a insistir que el tamaño de muestra es pequeño).

El mismo resultado se obtiene con la regresión múltiple, la correlación múltiple es igual a la raíz cuadrada del coeficiente de determinación, esto es

$$R_{1.234} = \sqrt{6.662} = 0.8136$$

con el mismo valor y significación de la ***F***, la del modelo.

ANÁLISIS DE CORRELACIÓN CANÓNICA

La técnica del análisis de la correlación canónica se entiende mejor considerándola como una extensión de la regresión múltiple y de la correlación. El análisis de regresión múltiple consiste en encontrar la mejor combinación lineal de p variables independientes, X_1, X_2,..., X_p, para predecir la variable dependiente Y. La correlación múltiple es la correlación simple entre Y y sus valores estimados por la ecuación de regresión, $\hat{Y}$. Por tanto, el objetivo en los análisis de regresión y correlación múltiple está en examinar la relación entre varias variables, X, y una variable, Y.

El análisis de correlación canónica se aplica a situaciones donde es apropiada la técnica de la regresión pero para más de una variable dependiente. Aunque otra aplicación del análisis de correlación canónica es como un método para determinar la asociación entre dos grupos de variables. Es una generalización de la regresión múltiple al caso de más de una variable dependiente.

Este análisis esta íntimamente relacionado con el análisis canónico discriminante y tiene ciertas propiedades análogas al análisis de componentes principales y al análisis factorial, en el que en lugar de tratar de estudiar las dependencias internas entre las variables de un mismo grupo, en el caso de la correlación canónica lo que se estudia es la relación o dependencia entre dos grupos de variables.

Recuérdese que el análisis de regresión múltiple trataba de encontrar la combinación lineal de p variables, X_1, X_2,..., X_p, que mejor predigan la variable dependiente Y. El coeficiente de correlación múltiple es la correlación simple entre Y y su predicción por medio de la ecuación de regresión.

En el análisis de correlación canónica se examina la relación lineal entre un grupo de variables, *X*, y un grupo, o más de un grupo, de variables *Y*. Por lo que la diferencia es que ahora se tiene más de una variable *Y*. La técnica consiste en encontrar una combinación lineal de las variables *X* ($V_1=b_1X_1+b_2X_2+...+b_pX_p$) y otra combinación lineal de las variables *Y* ($U_1=a_1Y_1+a_2Y_2+...+a_qY_q$) de tal manera que la correlación entre *U* y *V* sea máxima. Después encontrar otras dos combinaciones lineales para cada grupo de variable que tenga correlación máxima y así sucesivamente se encuentran un conjunto de combinaciones lineales para cada grupo de variables que tienen correlación máxima. A estas combinaciones lineales se denominan *variables canónicas*, y las correlaciones entre los correspondientes pares de variables canónicas se denominan *correlaciones canónicas*.

En una aplicación común de esta técnica las *Y* se interpretan como variable *respuesta* o variables *dependiente*, mientras que las variables *X* representan variables *predictivas* o variables *independientes*. Las variables *Y* pueden ser más dificultoso de medir que *X* como ocurre con los problemas de calibración.

El análisis de correlación canónica se aplica a situaciones en las que es adecuada la técnica de la regresión pero existe más de una variable dependiente. Otra aplicación útil es para probar la independencia entre los dos grupos de variables, *Y* y *X*, como se verá dentro de un momento.

Ejemplos de aplicaciones de la correlación canónica pueden ser el estudio para relacionar las características de ciertas variedades de trigo y características de las harinas resultantes. En este estudio fue posible concluir que el trigo deseable es el que tiene valores altos de textura, densidad y contenido en proteínas y el que tiene valores bajos en granos deteriorados y en productos extraños. Similarmente, una harina buena debe tener un alto contenido en proteína y bajos valores de ceniza. La correlación canónica también puede usarse en psicología para calibrar dos grupos de pruebas de inteligencia hechas a los mismos individuos. También, ha sido usado para relacionar las combinaciones lineales de las escalas de personalidad con las combinaciones lineales de las pruebas psicológicas realizadas.

El análisis de correlación canónica es, de las técnicas multivariantes, uno de los menos utilizados. Esto es debido, en parte, a la dificultad que se puede encontrar a la hora de interpretar los resultados.

Ejemplo hipotético.-

Supóngase cierta especie, por ejemplo, caprino, en la que se toma una muestra de diez individuos en los que se miden dos variables productivas de leche, como pueden ser *producción máxima diaria* (Y_1, *producción* en adelante) y *porcentaje de nitrógeno total* (Y_2, *porcentaje* en adelante), y dos variables de conformación, como pueden ser *longitud total del cuerpo* (X_1, *longitud* en adelante) y *anchura de las caderas* (X_2, *anchura* en adelante). Los datos son

Individuo	Y_1	Y_2	X_1	X_2

1	122	40	332	116
2	120	42	320	107
3	126	44	339	119
4	125	39	336	114
5	120	38	321	106
6	127	45	336	119
7	128	49	347	128
8	130	39	349	129
9	123	41	338	111
10	124	42	333	112

Las variables de conformación están medidas en la misma escala pero en diferentes unidades.

Los estadísticos básicos de estas variables son

Variable	$\overline{X}$	S
Y_1	124.50	3.3416
Y_2	41.90	3.3483
X_1	335.10	9.4334
X_2	116.10	7.8662

y la matriz de correlaciones es

	Y_1	Y_2	X_1	X_2
Y_1	1.00000	0.41212	0.92525	0.92782
Y_2		1.00000	0.38379	0.46868
X_1			1.00000	0.90874
X_2				1.00000

Como se observa en esta matriz, la variable *producción* está altamente correlacionada tanto con la *longitud* como con la *anchura*, mientras que la variable *proporción* no esta tan correlacionada con las variables de conformación. Las dos variables de conformación están, lógicamente, muy correlacionadas entre ellas, mientras que las dos variables de producción, siendo una de cantidad y la otra de proporción, están medianamente correlacionadas.

Conceptos básicos de la correlación canónica.-

Supóngase que se va ha estudiar la relación entre un grupo de variables, x_1, x_2,..., x_p y otro grupo de variables, y_1, y_2,..., y_q. Las variables x pueden ser vistas como variables independientes o predictoras, mientras que las variables y se pueden considerar como variables dependientes o variables respuesta. Se asume que, en una muestra dada, se le resta a los datos originales la media de cada variable, por lo que la media de todas las x y todas las y valen cero.

Primera correlación canónica.-

La idea básica del análisis de correlación canónica comienza buscando una combinación lineal de las *y*, tal como

$$U_1 = a_1y_1+a_2y_2+...+a_qy_q$$

y una combinación lineal de las *x*, tal como

$$V_1 = b_1x_1+b_2x_2+...+b_px_p$$

Para cualquier elección de los coeficientes, *a* y *b*, se puede calcular los valores U_1 y V_1 de cada individuo de la muestra. Para los *N* individuos de la muestra se puede calcular la correlación simple de los *N* pares, U_1 y V_1, de la manera usual. La correlación resultante dependerá de la elección de los valores de *a* y *b*.

En el análisis de correlación canónica, se seleccionan los valores de los coeficientes *a* y *b* de manera que *maximice* la correlación entre U_1 y V_1. Como consecuencia de esta particular elección de los coeficientes, a la combinación lineal U_1 se le denomina *primera variable canónica* de las *y*, y a la combinación lineal V_1 se le denomina *primera variable canónica* de las *x*. Nótese que tanto U_1 como V_1 tienen media cero. La correlación entre U_1 y V_1 se le denomina *primera correlación canónica*.

La primera correlación canónica es, por tanto, la correlación mayor posible entre la combinación lineal de las *x* y la combinación lineal de las *y*. En este sentido, es la correlación lineal máxima entre el grupo de las *x* y el grupo de las *y*. La primera correlación canónica es análoga al coeficiente de correlación múltiple entre una variable *Y* y un grupo de variables *X*. La diferencia es que en la correlación canónica hay varias *y* por lo que también hay que encontrar una combinación lineal de ellas.

El *SAS* provee los coeficientes, *a* y *b*, estos son

Coeficientes		*Coeficientes tipificados*	
a_1=0.29107	b_1=0.04948	a_1=0.9727	b_1=0.4667
a_2=0.01864	b_2=0.07077	a_2=0.0624	b_2=0.5567

Como se ve, los coeficientes tipificados se calculan multiplicando los coeficientes por la desviación típica de la variable, así 0.9729=0.29107 x 3.34166.

Las variables canónicas se calculan con los coeficientes no tipificados y, tal como se dijo anteriormente, con la diferencia de cada dato original con su media, de manera que las primeras variables canónicas (U_1 y V_1) para, por ejemplo, el primer individuo de la tabla de datos es

$$U_1 = 0.29107(122-124.5) + 0.01864(40-41.9) = -0.76309$$
$$V_1 = 0.04948(332-335.1) + 0.07077(116-116.1) = -0.16045$$

Si se calcula U_1 y V_1 para todos los individuos y se estima la correlación lineal simple de estas dos variables se obtiene la primera correlación canónica, que en este

caso vale 0.9499. Este valor representa la correlación mayor posible entre cualquier combinación lineal de las variables independientes y cualquier combinación lineal de las variables dependientes. Particularmente, es mayor que cualquier correlación entre una *X* y una *Y*, como se puede comprobar con la matriz de correlaciones de las variables expuesta anteriormente.

Un método para interpretar el valor relativo de cada variable en la combinación lineal canónica, es viendo el valor de los coeficientes tipificados. Así, para las *Y*, la primera variable canónica viene determinada fundamentalmente por la variable *producción*, lo que quiere decir que un individuo que produzca relativamente mucho tendrán un alto valor de la primera variable canónica U_1. Mientras que en el valor de la primera variable canónica de las *X* tiene una influencia ligeramente superior la *anchura* que la *longitud*, se puede considerar que en ambas la influencia es la misma.

Otro método para interpretar el valor relativo de cada variable en la combinación lineal canónica, es viendo el valor de la correlación de cada variable original con su variable canónica (o con la variable canónica del otro grupo de variables). Estos valores los da el *SAS* por defecto y son, en el caso del ejemplo hipotético

	U_1	V_1
Y_1	0.9984	0.9484
Y_2	0.4633	0.4400
X_1	0.9239	0.9726
X_2	0.9317	0.9808

Como se ve, la primera variable canónica de las *Y* (U_1) está altamente correlacionada con la Y_1 y medianamente correlacionada con la Y_2 por lo que se puede determinar que la primera variable canónica viene determinada fundamentalmente por la variable *producción*, lo que quiere decir que un individuo que produzca relativamente más, tendrán un alto valor de la primera variable canónica U_1. Mientras que en el valor de la primera variable canónica de las *X* (V_1) tiene una correlación ligeramente superior con *anchura* que la *longitud*, pero en ambas es elevada.

De todo esto se deduce que los individuos muy *anchos* y *largos* tienen una elevada *producción* pero no indica nada de la *proporción*. Lógicamente, en un ejemplo real con más variables el análisis de estos coeficientes puede ser gran utilidad para conducir a múltiples y variadas conclusiones.

Segunda (y sucesivas) correlación canónica.-

Se puede realizar interpretaciones adicionales de la relación entre las *X* y las *Y* obteniendo otro conjunto de variables canónicas y su correspondiente correlación canónica. Concretamente, se puede hallar la segunda variable canónica V_2 (combinación lineal de las *x*) y la correspondiente variable canónica U_2 (combinación lineal de las *y*). Los coeficientes de estas combinaciones lineales se eligen teniendo en cuenta las siguientes condiciones:

1. V_2 esta incorrelacionada con V_1 y U_1.
2. U_2 esta incorrelacionada con V_1 y U_1.
3. Una vez cumplidas las condiciones anteriores, U_2 y V_2 tienen la máxima correlación posible.

La correlación entre U_2 y V_2 se denomina *segunda correlación canónica* y necesariamente es menor o igual que la primera correlación canónica.

Este paso se repite para calcular la tercera, cuarta, *etc.*, variables canónicas. El número máximo de correlaciones canónicas y sus correspondientes variables canónicas es igual al número mínimo de variable en los grupos, esto es, si hay por ejemplo 10 variables X y 5 variables Y el número de correlaciones canónicas que se podrán calcular será de 5.

Para el ejemplo hipotético, la segunda correlación canónica vale 0.2147. Los coeficientes de las segundas variables canónicas son

Coeficientes		*Coeficientes tipificados*	
a_1=-0.15215	b_1=-0.24912	a_1=-0.5084	b_1=-2.3501
a_2= 0.32726	b_2= 0.29625	a_2= 1.0958	b_2= 2.3304

y las correlaciones con las variables originales son

	U_2	V_2
Y_1	-0.0569	-0.0122
Y_2	0.8862	0.1903
X_1	-0.0499	-0.2323
X_2	0.0418	0.1948

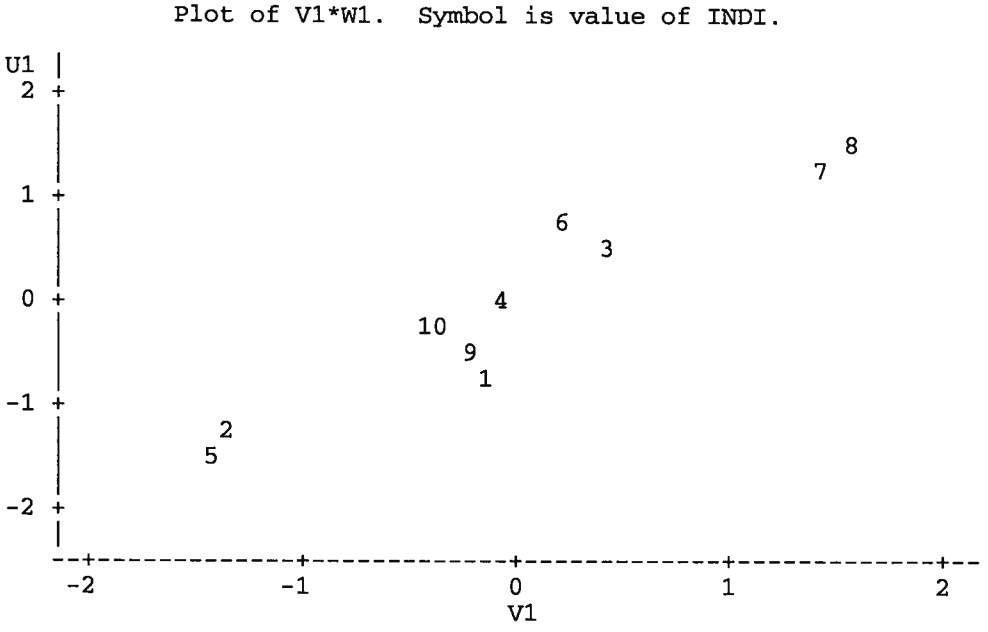
Tanto con los coeficientes tipificados como con las correlaciones se observa que en la segunda variable canónica de las Y (U_2) tiene una influencia negativa la variable Y_1, baja en valor absoluto comparada con la influencia positiva de la variables Y_2, esto significa que en la segunda variable canónica de las Y tiene una gran influencia positiva la variable *proporción* y una leve influencia negativa la variable *producción*. Mientras que en el valor de la segunda variable canónica de las X (V_2) tienen prácticamente la misma influencia las dos variables pero en sentido contrario, esto es, la *longitud* tiene una influencia negativa en el valor se su segunda variable canónica y la *anchura* tiene una influencia positiva, pero en ambas variables esta influencia es más bien baja.

Pruebas de hipótesis.-

El paquete *SAS* realiza la prueba de razón de verosimilitud para probar las correlaciones canónicas. En el caso del ejemplo hipotético, la razón de verosimilitud para la primera correlación canónica es de 0.09318 que corresponde a una *F* aproximada de 6.8276 que para 4 y 12 grados de libertad es significativa al 0.01 por lo que se puede concluir que la primera correlación canónica es significativa, por lo que están correlacionadas ambos tipos de variables (las de producción con las de conformación) y son significativamente ciertas las conclusiones que obtuvimos al analizar los coeficientes tipificados y las correlaciones de las primeras variables canónicas. Mientras que la razón de verosimilitud de la segunda correlación canónica es 0.95389 (prácticamente 1), cuya *F* aproximada es 0.3383 (inferior a 1) por lo que la segunda correlación canónica es no significativa, y por tanto no son significativas ciertas las conclusiones que obtuvimos al analizar los coeficientes y las correlaciones de las segundas variables canónicas.

Representación de los valores de las variables canónicas.-

Puede ser útil la representación, en unos ejes cartesianos, de los individuos siendo el valor de sus coordenadas la de los valores de las variables canónicas, U_1 y V_1. Para el ejemplo hipotético esta representación sería



Como se observa, existe una pendiente positiva acentuada como era de esperar por la significación de la primera correlación canónica (0.9499). Para datos multivariante normal, esta gráfica será una elipse de dispersión que podrá ser útil para detectar posibles datos erróneos o individuos peculiares.

Archivo del programa *SAS* (C16-4.SAS).-

Para obtener los resultados y análisis del ejemplo hipotético. el programa *SAS* es

```
title 'correlación canónica';
options ls=64 ps=30;
data corrcan;
infile 'c16-4.dat';
input indi y1 y2 x1 x2;
proc cancorr corr out=canonica;
var y1 y2;
with x1 x2;
run;
proc print;
run;
proc plot;
plot V1*W1=indi;
run;
```

Archivo de datos (C16-4.DAT).-

```
 1 122 40 332 116
 2 120 42 320 107
 1 126 44 339 119
 4 125 39 336 114
 5 120 38 321 106
 6 127 45 336 119
 7 128 49 347 128
 8 130 39 349 129
 9 123 41 338 111
10 124 42 333 112
```

Archivo de resultados (C16-4.LST).-

```
                    correlación canónica

          Correlations Among the Original Variables
            Correlations Among the 'VAR' Variables

                              Y1                Y2
            Y1            1.0000            0.4121
            Y2            0.4121            1.0000

           Correlations Among the 'WITH' Variables

                              X1                X2
            X1            1.0000            0.9087
            X2            0.9087            1.0000

           Correlations Between the 'VAR' Variables
                   and the 'WITH' Variables

                              X1                X2
            Y1            0.9252            0.9278
            Y2            0.3838            0.4687

                        Adjusted        Approx        Squared
          Canonical     Canonical      Standard      Canonical
        Correlation   Correlation         Error    Correlation
   1       0.949900      0.943569      0.032563       0.902310
   2       0.214725             .      0.317964       0.046107
```

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	9.2365	9.1882	0.9948	0.9948
2	0.0483	.	0.0052	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.09318546	6.8276	4	12	0.0042
2	0.95389323	0.3383	1	7	0.5790

Multivariate Statistics and F Approximations

S=2 M=-0.5 N=2

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.093185	6.8276	4	12	0.0042
Pillai's Trace	0.948417	3.1566	4	14	0.0480
Hotelling-Lawley Trace	9.284839	11.606	4	10	0.0009
Roy's Greatest Root	9.236503	32.328	2	7	0.0003

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Raw Canonical Coefficients for the 'VAR' Variables

	V1	V2
Y1	0.2910727943	-0.15215139
Y2	0.0186370971	0.3272589661

Raw Canonical Coefficients for the 'WITH' Variables

	W1	W2
X1	0.0494764796	-0.249124834
X2	0.0707716177	0.2962555648

Standardized Canonical Coefficients for the 'VAR' Variables

	V1	V2
Y1	0.9727	-0.5084
Y2	0.0624	1.0958

Standardized Canonical Coefficients for the 'WITH' Variables

	W1	W2
X1	0.4667	-2.3501
X2	0.5567	2.3304

Canonical Structure
Correlations Between the 'VAR' Variables and Their Canonical Variables

	V1	V2
Y1	0.9984	-0.0569
Y2	0.4633	0.8862

Correlations Between the 'WITH' Variables and Their Canonical Variables

	W1	W2
X1	0.9726	-0.2323
X2	0.9808	0.1948

Correlations Between the 'VAR' Variables and the Canonical Variables of the 'WITH' Variables

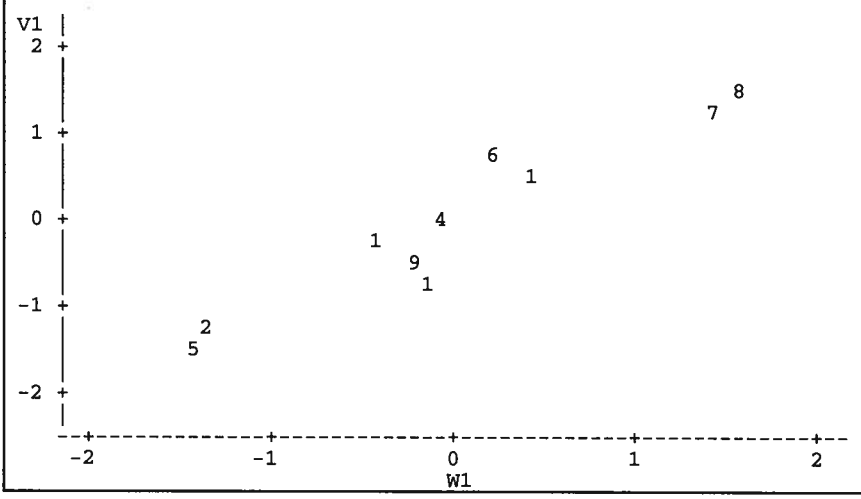
	W1	W2
Y1	0.9484	-0.0122
Y2	0.4400	0.1903

Canonical Structure
Correlations Between the 'WITH' Variables and the Canonical Variables of the 'VAR' Variables

	V1	V2
X1	0.9239	-0.0499
X2	0.9317	0.0418

OBS	INDI	Y1	Y2	X1	X2	V1	V2	W1	W2
1	1	122	40	332	116	-0.76309	-0.24141	-0.16045	0.74266
2	2	120	42	320	107	-1.30796	0.71741	-1.39112	1.06586
3	1	126	44	339	119	0.47575	0.45902	0.39820	-0.11245
4	4	125	39	336	114	0.09149	-1.02513	-0.10409	-0.84635
5	5	120	38	321	106	-1.38251	-0.59163	-1.41241	0.52048
6	6	127	45	336	119	0.78546	0.63412	0.24977	0.63493
7	7	128	49	347	128	1.15108	1.79101	1.43095	0.56086
8	8	130	39	349	129	1.54685	-1.78588	1.60068	0.35886
9	9	123	41	338	111	-0.45338	-0.06631	-0.21745	-2.23337
10	10	124	42	333	112	-0.14367	0.10880	-0.39406	-0.69149

Plot of V1*W1. Symbol is value of INDI.



Correlación canónica usando los componentes principales.-

Otro método para examinar la relación entre un grupo de variables X y de variables Y es el siguiente:

1. Obtener los componentes principales de y_1, y_2,..., y_q y simbolizarlos por D_1, D_2,..., D_q.
2. Obtener los componentes principales de x_1, x_2,..., x_q y simbolizarlos por C_1, C_2,..., C_q.
3. Elegir las primeras m componentes principales de cada grupo.

4. Calcular la correlación entre C_1 y D_1; entre C_2 y D_2; ...,; C_m y D_m.

Las correlaciones calculadas en el paso cuatro son, por lo general, menores que las correlaciones canónicas, pero las componentes principales pueden ser interesantes por ellas mismas. Las componentes principales explican el máximo de varianza *dentro* del grupo de variables, mientras que las variables canónicas maximizan la correlación *entre* los dos grupos de variables.

Puesto que, por ejemplo, C_1 y D_2 pueden tener una correlación diferente de cero, puede ser útil calcular dichas correlaciones además de las descritas en el paso 4.

Archivo del programa *SAS* (C16-5.SAS).-

```
Title 'Correlaciones canónicas usando componentes principales';
Options ls=64 ps=30;
Data corrcan;
Infile 'c16-4.dat';
Input indi y1 y2 x1 x2;
Proc princomp out=ccom;
var y1 y2;
run;
proc princomp
      data=ccom(rename=(prin1=priny1
prin2=priny2))
      out=ccom;
var x1 x2;
run;
proc corr data=ccom(keep=priny1 priny2 prin1
prin2);
run;
```

Archivo de resultados (C16-5.LST).-

```
    Correlaciones canónicas usando componentes principales

                Principal Component Analysis
      10 Observations
       2 Variables

                      Simple Statistics
                              Y1                 Y2
           Mean        124.5000000        41.90000000
           StD           3.3416563         3.34829973

                      Correlation Matrix
                              Y1                 Y2
                  Y1        1.0000             0.4121
                  Y2        0.4121             1.0000

             Eigenvalues of the Correlation Matrix

          Eigenvalue     Difference     Proportion     Cumulative
PRIN1       1.41212        0.824232       0.706058        0.70606
PRIN2       0.58788              .        0.293942        1.00000

                          Eigenvectors

                             PRIN1              PRIN2
                  Y1      0.707107           0.707107
                  Y2      0.707107           -.707107
```

10 Observations
2 Variables

Simple Statistics

	X1	X2
Mean	335.1000000	116.1000000
StD	9.4333922	7.8662429

Correlation Matrix

	X1	X2
X1	1.0000	0.9087
X2	0.9087	1.0000

Principal Component Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	1.90874	1.81748	0.954370	0.95437
PRIN2	0.09126	.	0.045630	1.00000

Eigenvectors

	PRIN1	PRIN2
X1	0.707107	0.707107
X2	0.707107	-.707107

Correlation Analysis

4 'VAR' Variables: PRINY1 PRINY2 PRIN1 PRIN2

Simple Statistics

Variable	N	Mean	Std Dev	Sum
PRINY1	10	0	1.188325	0
PRINY2	10	0	0.766736	0
PRIN1	10	0	1.381571	0
PRIN2	10	0	0.302094	0

Simple Statistics

Variable	Minimum	Maximum
PRINY1	-1.775834	2.240018
PRINY2	-0.973335	1.776253
PRIN1	-1.964808	2.201512
PRIN2	-0.313853	0.675823

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0
/ N = 10

	PRINY1	PRINY2	PRIN1	PRIN2
PRINY1	1.00000	0.00000	0.82398	-0.12183
	0.0	1.0000	0.0034	0.7374
PRINY2	0.00000	1.00000	0.47229	0.17771
	1.0000	0.0	0.1681	0.6233
PRIN1	0.82398	0.47229	1.00000	0.00000
	0.0034	0.1681	0.0	1.0000
PRIN2	-0.12183	0.17771	0.00000	1.00000
	0.7374	0.6233	1.0000	0.0

Aplicación al análisis discriminante.-

Cuando se estudie el Análisis Discriminante (ver Capítulo 19), se virá que consiste en la clasificación de un individuo en una de k>=2 poblaciones en base a las medidas X_1, X_2,..., X_p. Para clasificar un individuo se calcula cada una de las k funciones discriminantes y se asigna el individuo a la población para la cual ha dado un mayor valor la función discriminante. Este proceso es el aspecto predictivo de la clasificación. Para fines descriptivos está el Análisis Canónico Discriminante. Recuerde que en este análisis las variables canónicas se deducían de manera que fuera máxima la diferencia entre los grupos. Conocidas las correlaciones canónicas se puede estudiar el planteamiento de estas funciones discriminantes .

Se puede comenzar definiendo un conjunto de nuevas variables Y_1. Y_2,..., Y_{k-1}, que serán variables de diseño o de incidencia, indicando el valor de cada variable la pertenencia o no a uno de los grupos. Recuerdes cuando se estudio el análisis discriminante que se dijo que se necesitan k-1 variables de diseño para describir k grupos. Por ejemplo, supóngase que se han medido p variables en k=4 grupos, el valor de las variables Y_1. Y_2 e Y_3 serán

grupo	Y_1	Y_2	Y_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Si un individuos pertenece al grupo 1 se le da el valor de 1 a la variable Y_1 y cero a las otras dos variables Y, *etc*, con lo que se tendrá q=k-1 variables Y y p variables X. Con estas variables se realiza el análisis de correlaciones canónicas en el que se obtendrán las variables canónicas U_1 y V_1, cuyo número es, tal como se dijo antes, igual al mínimo valor de p o de q.

La variable canónica V_1 es la función discriminante que se vio en el análisis discriminante. Como V_1 es la combinación lineal de las X que maximiza la correlación con U_1, en este sentido, V_1 maximiza la correlación con las variables de diseño que representan los grupos y además maximiza la diferencia entre los grupos. Similarmente, V_2 exhibe la máxima diferencia entre grupos con la condición previa de que este incorrelacionada con V_2.

Ejemplo.-

Se ha medido en 29 cabras dos tipos de medidas. medidas productivas de la leche total y medidas productivas de una fracción de la leche. Las medidas productivas de la leche total son; producción total en *Kg* (**prod**), porcentaje de proteína total (**prot**), porcentaje de caseína total (**cas**), porcentaje de grasa (**gras**) y porcentaje de lactosa (**lac**). Las medidas productiva de la fracción de la caseína son: porcentaje de caseína α (**alfa**), porcentaje de la caseína β (**beta**) y porcentaje de la caseína κ (**kapa**).

Se quiere saber si las variables de la fracción de las caseínas influyen en la producción total

Archivo del programa *SAS* (C16-6.SAS).-

```
title 'Correlaciones canónicas y análisis discriminante';
options ls=75 ps=30 ;
data coca;
infile 'c16-6.dat';
input prod prot cas gras lac alfa beta kapa;
proc cancorr corr out=canonica;
  var prod prot cas gras lac;
  with alfa beta kapa;
run;
proc plot;
plot V1*W1;
run;
```

Archivo de datos (C16-6.DAT).-

```
188.60   5.64   4.92  12.23   6.92   1.87   2.81   0.61
288.18   9.17   8.06  20.16  10.68   2.82   4.64   1.45
346.25   9.85   8.74  19.14  13.41   3.23   4.49   1.39
402.53  10.69   9.55  22.80  15.51   3.55   5.30   1.03
331.38  11.85   9.68  24.44  12.37   3.25   4.78   2.02
282.25   9.89   7.90  17.87  10.04   2.72   4.27   1.16
278.90   8.22   7.66  16.47  10.39   2.70   4.11   1.23
304.75   8.92   7.56  18.32  11.70   2.88   4.13   0.87
335.95   9.98   9.49  24.26  12.00   2.94   5.98   1.09
340.23  11.79   9.78  23.34  12.06   3.16   5.43   1.65
365.45  11.95   9.86  24.89  12.79   3.53   4.92   1.98
283.40   8.68   7.41  13.13  12.98   2.07   4.60   1.04
420.98  14.37  11.07  24.90  15.53   4.05   5.18   2.04
507.65  19.12  13.10  28.08  19.84   4.31   7.00   2.15
531.58  19.95  13.81  28.91  20.92   4.32   7.49   2.56
467.10  15.04  12.87  30.97  17.20   4.11   7.18   1.92
606.23  23.05  15.76  30.96  24.21   4.60   9.03   2.64
336.20  10.58   8.56  21.57  12.56   3.21   4.15   1.43
266.55   8.69   6.87  18.15   9.94   2.55   3.39   1.12
266.95   8.96   7.05  17.46   9.99   2.55   3.26   1.30
337.85  11.26   8.78  21.53  12.76   3.22   3.94   1.91
332.90  10.85   8.70  22.19  12.44   3.18   4.16   1.75
252.25   8.67   6.86  16.87   8.98   2.42   3.76   0.77
194.08   5.91   6.28  15.78   5.58   1.44   4.67   0.37
244.85   7.27   6.27  14.57   9.17   2.33   3.06   1.01
580.10  19.95  14.75  29.79  23.07   4.61   8.96   1.80
274.20   9.20   6.94  19.42  10.40   2.65   3.13   1.37
702.75  24.17  19.29  40.45  26.09   5.63  12.13   2.30
249.25   7.57   6.16  16.46   9.10   2.38   3.05   1.17
```

Archivo de resultados (C16-6.LST).-

Correlaciones canónicas y análisis discriminante

Correlations Among the Original Variables
Correlations Among the 'VAR' Variables

	PROD	PROT	CAS	GRAS	LAC
PROD	1.0000	0.9803	0.9884	0.9401	0.9901
PROT	0.9803	1.0000	0.9768	0.9218	0.9722
CAS	0.9884	0.9768	1.0000	0.9615	0.9661
GRAS	0.9401	0.9218	0.9615	1.0000	0.8934
LAC	0.9901	0.9722	0.9661	0.8934	1.0000

Correlations Among the Original Variables
Correlations Among the 'WITH' Variables

	ALFA	BETA	KAPA
ALFA	1.0000	0.8685	0.8562
BETA	0.8685	1.0000	0.6633
KAPA	0.8562	0.6633	1.0000

Correlations Between the 'VAR' Variables and the 'WITH' Variables

	ALFA	BETA	KAPA
PROD	0.9715	0.9415	0.8120
PROT	0.9465	0.9193	0.8568
CAS	0.9581	0.9672	0.8138
GRAS	0.9512	0.9090	0.8178
LAC	0.9495	0.9167	0.8049

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.999190	0.999045	0.000306	0.998381
2	0.837445	0.808430	0.056446	0.701314
3	0.690151	0.686679	0.098968	0.476308

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	616.6637	614.3157	0.9947	0.9947
2	2.3480	1.4385	0.0038	0.9985
3	0.9095	.	0.0015	1.0000

Test of H0: The canonical correlations in the
current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.00025324	74.2574	15	58.37315	0.0001
2	0.15641949	8.4065	8	44	0.0001
3	0.52369207	6.9730	3	23	0.0017

Multivariate Statistics and F Approximations
S=3 M=0.5 N=9.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00025324	74.2574	15	58.3732	0.0001
Pillai's Trace	2.17600293	12.1476	15	69	0.0001
Hotelling-Lawley Trace	619.92119	812.786	15	59	0.0001
Roy's Greatest Root	616.663673	2836.65	5	23	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Raw Canonical Coefficients for the 'VAR' Variables

	V1	V2	V3
PROD	0.001167721	-0.098884689	-0.033446343
PROT	-0.020687319	0.3954195116	0.7841976653
CAS	0.2850130903	1.974711243	-1.84978534
GRAS	0.0041180613	-0.045852499	0.5286837139
LAC	-0.00274858	0.8486001878	0.6719671304

Raw Canonical Coefficients for the 'WITH' Variables

	W1	W2	W3
ALFA	0.3768546412	-3.278532477	0.2321208871
BETA	0.2730021517	0.7747679119	-0.567918656
KAPA	0.2172365833	2.8895596259	2.0782198692

Standardized Canonical Coefficients for the 'VAR' Variables

	V1	V2	V3
PROD	0.1470	-12.4462	-4.2097
PROT	-0.1020	1.9499	3.8671
CAS	0.9426	6.5306	-6.1174
GRAS	0.0260	-0.2896	3.3389
LAC	-0.0140	4.3108	3.4135

Standardized Canonical Coefficients for the 'WITH' Variables

	W1	W2	W3
ALFA	0.3507	-3.0506	0.2160
BETA	0.5809	1.6487	-1.2085
KAPA	0.1242	1.6519	1.1881

Canonical Structure

Correlations Between the 'VAR' Variables and Their Canonical Variables

	V1	V2	V3
PROD	0.9893	-0.0836	0.0533
PROT	0.9731	0.0519	0.1614
CAS	0.9997	0.0194	0.0072
GRAS	0.9640	-0.0621	0.1136
LAC	0.9663	-0.0663	0.0775

Correlations Between the 'WITH' Variables and Their Canonical Variables

	W1	W2	W3
ALFA	0.9616	-0.2043	0.1835
BETA	0.9679	0.0948	-0.2329
KAPA	0.8097	0.1335	0.5714

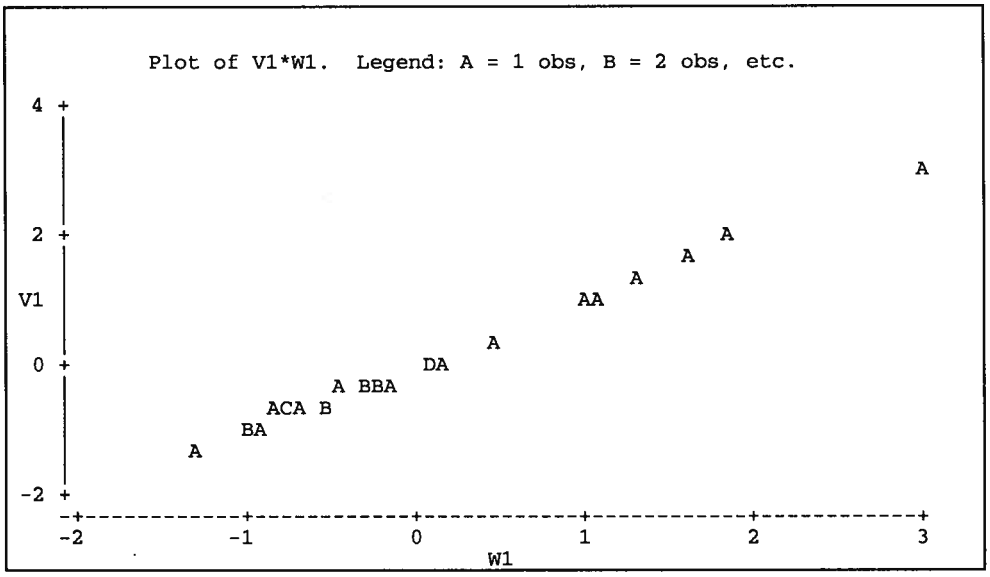
Canonical Structure

Correlations Between the 'VAR' Variables and the Canonical Variables of the 'WITH' Variables

	W1	W2	W3
PROD	0.9885	-0.0700	0.0368
PROT	0.9723	0.0434	0.1114
CAS	0.9989	0.0162	0.0050
GRAS	0.9632	-0.0520	0.0784
LAC	0.9655	-0.0555	0.0535

Correlations Between the 'WITH' Variables and the Canonical Variables of the 'VAR' Variables

	V1	V2	V3
ALFA	0.9608	-0.1711	0.1267
BETA	0.9671	0.0794	-0.1608
KAPA	0.8091	0.1118	0.3944



. Como se ha puesto la opción **CORR** la primera salida es la de las correlaciones entre las variables originales, que son muy elevadas como era de esperar por las características de los datos.

. Como el grupo más pequeño de variables es el de tres variables, solo se puede estimar tres correlaciones canónicas. Esta es la siguiente salida, la primera correlación canónica vale 0.99919, la segunda vale 0.83744 y la tercera 0.69015.

. Después de los valores propios vienen las tres pruebas de hipótesis para las tres correlaciones canónicas, indicando que las tres son significativamente diferentes de cero.

. Después vienen los coeficientes de las variables canónicas, los coeficientes tipificados y las correlaciones de cada variable original con las variables canónicas. Estudiando los coeficientes tipificados y las correlaciones se observa que para la primera variable canónica de las variables productivas totales, la que más influye en ella es la caseína y la que menos la producción total, mientras que en la primera variable canónica de las variables de la fracción de caseína la que más influye es la α-caseína y la que menos la κ-caseína.

. En la salida del procedimiento **PLOT** se observa claramente que ambos grupos de variables están altamente correlacionadas, pues la nube de dispersión es prácticamente una recta de pendiente muy marcada.

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

Lite, TM, y Jackson Hills, F. 1987. MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA. Ed TRILLAS. México.

Milton, J.S. 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Spiegel M.R. 1990. ESTADÍSTICA. Ed. McGraw-Hill. Madrid.

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

SAS Institute Inc. 1990. *SAS/STAT* USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.

SAS Institute Inc. 1990. *SAS PROCEDURE* GUIDE. Cary, NC, USA.

CAPÍTULO 17

Análisis de la Covarianza

CAPÍTULO 17

Análisis de la Covarianza

Introducción.-

El análisis de covarianza es una técnica estadística que combina los aspectos del análisis de varianza y del análisis de regresión. El análisis de covarianza trata de dos o más variables medidas, siendo *Y* la variable que se investiga, midiéndose también los valores de una (o unas) variable, *X*, *independiente* (o *covariable* o *variable concomitante*) que no se encuentra a niveles predeterminados, como ocurriría en un diseño factorial y donde los valores de esta covariable no están influidos por los tratamientos. Se va a estudiar la covarianza lineal solamente. Aunque como ya se ha estudiado en el Capítulo 11, una relación lineal es una aproximación razonablemente buena para una relación no lineal con tal de que, dado el caso, los valores de las variables independientes no cubran un intervalo muy amplio.

Usos del análisis de covarianza.-

Los usos más importantes del análisis de covarianza son

1) Control el error y aumento de la precisión de los experimentos.
2) Ajuste de las medias de tratamientos de la variable dependiente teniendo en cuenta el efecto de las variables independientes correspondientes. Es decir, hace el contraste con las medias minimocuadráticas y no, como se hacía en el análisis de varianza, con las medias aritméticas.
3) Ayuda a la interpretación de los datos, especialmente en lo concemiente a la naturaleza de los efectos de los tratamientos.
4) Divide la covarianza o suma de productos cruzados en componentes, como se hizo con la varianza.
5) Compara las regresiones.

Desarrollemos brevemente estos usos.

Control del error.-

Como se vio en los Capítulos 4 y siguientes, la varianza de la media de un tratamiento es

$$\sigma^2_{(\bar{Y})} = \frac{\sigma^2_y}{n}$$

Por tanto, para disminuir esta varianza, sólo se tiene dos posibilidades: aumentar el tamaño de la muestra o disminuir, por medio de un mayor control, la varianza de la muestra.

El control de σ_y^2 se logra mediante un diseño experimental adecuado o mediante el uso de una o más covariables o con ambos métodos a la vez. Cuando se usa la covarianza como método para reducir el error, esto es, de controlar σ_y^2, se hace aceptando el hecho de que la variación observada en la variable dependiente *Y* es parcialmente atribuible a la variación de la variable independiente *X*, por tanto, se le puede restar a la varianza del error de *Y* (la que queda después de restarle a la varianza total la debida a los tratamientos) la parte de varianza debida a la covariable. El uso de ésta o estas covariables exige el uso de las técnicas de regresión estudiadas en los Capítulos 11 y siguientes.

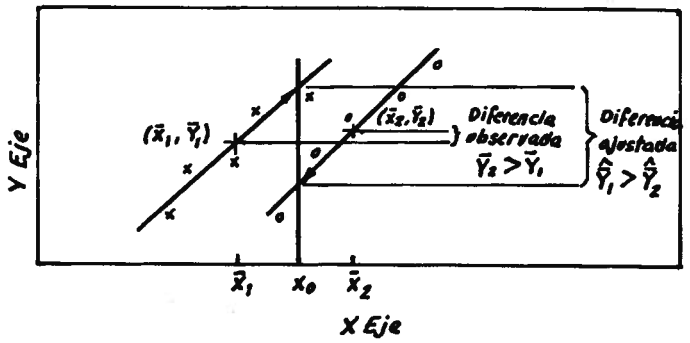
El uso de la covarianza para controlar el error es un medio de aumentar la precisión de la estima de los efectos de los tratamientos, eliminando, por regresión, ciertos efectos conocidos que no pueden ser o no han sido convenientemente controlados mediante el diseño experimental. La covariable (*X*) se mide en cada unidad experimental antes de aplicar los tratamientos. Por ejemplo, en un experimento de nutrición animal para comparar el efecto de varias dietas en el *incremento de peso*, los animales asignados a un bloque o tratamiento variarán en su *peso inicial* o *peso al nacimiento* y si el peso inicial está correlacionado con el incremento de peso, una fracción muy importante del error experimental, en la ganancia de peso, se deberá a las diferencias en los pesos iniciales. Mediante el análisis de covarianza, esta porción del error debida a la diferencia en el peso inicial puede calcularse y eliminarse del error experimental de la variable *incremento de peso*. Se conseguiría el mismo resultado si fuera posible comenzar esta experiencia con individuos, todos, del mismo peso, controlando, por tanto, la σ^2 mediante el diseño experimental. Como se ve, es mucho más fácil y factible aumentar la precisión restándole al error la variabilidad debida a la covariable. Este es, probablemente, el uso más común del análisis de covarianza.

Ajuste de medias de tratamientos.-

Cuando parte de la variación observada en la variable dependiente (*Y*) puede atribuirse a la variación de la variable independiente (*X*), también se ve afectada la variación de *Y* entre los tratamientos por la variación de *X* entre los tratamientos. Para que sean comparables los tratamientos, las $\bar{Y}$ de los tratamientos tienen que ajustarse y estimarlas como si las $\bar{X}$ de los tratamientos hubiesen sido las mismas. Si el objeto

principal de la covarianza es ajustar las $\overline{Y}$ de los tratamientos, es como consecuencia del reconocimiento de una relación funcional (regresión) que exige el correspondiente ajuste del error. En todo caso, es necesario medir la regresión independientemente de las otras fuentes de variación que estemos usando en el modelo.

La idea general es evidente viendo la siguiente figura para dos tratamientos



Se observa que para cada tratamiento la variación de *X* contribuye a la variación de *Y*. Así pues, se ve la necesidad de controlar la varianza del error mediante el uso de la covariable. Al mismo tiempo, la distancia entre $\overline{X}_1$ y $\overline{X}_2$ puede contribuir grandemente a la diferencia entre $\overline{Y}_1$ y $\overline{Y}_2$. Si las *Y* de los tratamientos se hubieran observado a partir de una $\overline{X}$ común, que podríamos llamar X_0, entonces serían comparables. Así pues, es evidente la necesidad de ajustar las medias de los tratamientos. Por ejemplo, se pretende estudiar si la obesidad de los trabajadores es función del tipo de trabajo que se realiza; para ello se mide como variable el peso de diferentes trabajadores de diferentes ocupaciones; pero también se ha medido la edad de los trabajadores y se ha observado que hay diferencia entre la edad media en los diferentes trabajos; si la obesidad está linealmente relacionada con la edad, las diferencias encontradas en la obesidad entre diferentes trabajos pueden deberse, o estar altamente influidas por las diferencias de edad, por lo que es preciso ajustar las medias de los diferentes tratamientos (*trabajos*) con la covariable edad, como si la edad media en los diferentes trabajos fuese la misma.

Interpretación de datos.-

Todos los procedimientos estadísticos tienen como objeto la interpretación de datos; por tanto los otros usos del análisis de covarianza tienen que ver con la interpretación de datos. Sin embargo este punto se refiere a que el análisis de covarianza a menudo ayuda al experimentador a entender mejor la naturaleza de los efectos de los tratamientos en los resultados de una investigación. Por ejemplo, es sabido que ciertos tratamientos pueden producir efectos tanto en la variable dependiente como en la independiente. Es correcto el uso de la covarianza con objeto de controlar el error y ajustar las medias de los tratamientos, si la variable independiente mide efectos ambientales y en ella no influyen los tratamientos. Pero si ocurriera que los tratamientos tienen efectos sobre la variable independiente la interpretación de los datos cambia como consecuencia

de que el ajuste elimina parte de los efectos de los tratamientos cuando las medias de la variable independiente están afectadas por los tratamientos.

Partición de la covarianza.-

De la misma manera que se descompuso, en el análisis de la varianza, la suma de cuadrados total en componentes debidos a diferentes factores y al error, en el análisis de covarianza se descompone una suma de productos total.

La covarianza de un experimento se descompone cuando se quiere determinar la relación entre dos o más variables medidas sin que influya otra fuente de variación en esta relación. Por ejemplo, se desea determinar la relación entre el contenido de grasa y el de proteína de la leche en cierta especie, para ello se realiza un experimento factorial con repetición con las hijas de 25 machos en cuatro ganaderías. La suma de productos total de las, como mínimo, 100 observaciones puede dividirse en componentes de acuerdo con las fuentes de variación, o sea, el componente familia, el componente ganadería y el error. Si difieren significativamente los distintos coeficientes de regresión correspondientes a estas fuentes de variación, entonces la regresión total es heterogénea y no interpretable. En este experimento, el interés radica en la diferencias medias de producción entre familias en relación al error, estas diferencias medias miden la relación funcional promedio (regresión media) entre las dos variables observadas, dentro de las familias, tras eliminar los efectos de las ganaderías. Si la regresión *entre* familias es diferente a la regresión *dentro* de familias (error) hay que realizar ciertas correcciones en el modelo que se verán más adelante.

Comparación de regresiones.-

Ya se estudió un ejemplo representativo de esta problemática en el capítulo 11, éste consistía en el cálculo de la regresión de la concentración sanguínea de colesterol sobre la edad en tres muestra de individuos tomadas en tres regiones caracterizadas, entre otras cosas, por el diferente régimen alimenticio. Se deseaba saber si la regresión del colesterol sobre la edad es la misma en las tres regiones.

El modelo lineal del análisis de covarianza.-

El análisis de covarianza es una combinación del análisis de varianza y del análisis de la regresión, por lo tanto, el modelo lineal aditivo para un experimento dado es la combinación correspondiente del modelo del análisis de la varianza más un término adicional para la relación funcional con la variable independiente o covariable.

Si se tiene una sola variable independiente (X) la relación funcional de ésta con la variable dependiente (Y) es, tal como se vio en capítulo 11

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Por lo que los modelos lineales asociados con algunas de las experiencias más comunes

cuando se tiene una covariable son

Una vía o completamente aleatorio

$$Y_{ij} = \alpha + T_i + \beta X_{ij} + \varepsilon_{ij}$$

Dos vías o factorial sin repetición o de bloques aleatorios

$$Y_{ij} = \alpha + T_i + R_j + \beta X_{ij} + \varepsilon_{ij}$$

Jerárquicos o con subgrupos

$$Y_{ijk} = \alpha + T_i + R_{j(i)} + \beta X_{ijk} + \varepsilon_{ijk}$$

Trifactorial sin repetición o Cuadrado latino

$$Y_{ijk} = \alpha + T_i + F_j + C_k + \beta X_{ijk} + \varepsilon_{ijk}$$

Factorial de dos factores con repetición

$$Y_{ijk} = \alpha + T_i + R_j + TR_{ij} + \beta X_{ijk} + \varepsilon_{ijk}$$

Estos modelos muestran que se están estimando un conjunto de líneas paralelas, cuya pendiente común es β y cuyas ordenadas en el origen son $\alpha+T_i$ en el caso del modelo de una vía; $\alpha+T_i+R_j$ en el caso del modelo de dos vías sin repetición o de bloques aleatorios, etc.

En la práctica es más común expresar estos modelos en términos semejantes a los utilizados en el análisis de varianza. Para ello recuérdese que el parámetro μ de la variable *Y* puede estimarse, por regresión, de la siguiente forma

$$\mu = \alpha + \beta \overline{X}_{..}$$

por lo que

$$\alpha = \mu - \beta \overline{X}_{..}$$

que sustituyéndola en los modelos anteriores, se tiene

Una vía o completamente aleatorio

$$Y_{ij} = \mu + T_i + \beta(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

Dos vías o factorial sin repetición o de bloques aleatorios

$$Y_{ij} = \mu + T_i + R_j + \beta(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

Jerárquicos o con subgrupos

$$Y_{ijk} = \mu + T_i + R_{j(i)} + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

Trifactorial sin repetición o Cuadrado latino

$$Y_{ijk} = \mu + T_i + F_j + C_k + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

Factorial de dos factores con repetición

$$Y_{ijk} = \mu + T_i + R_j + TR_{ij} + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

La razón de escribir los modelos lineales de esta última manera es que muestra la base del análisis de covarianza de una manera más próxima.

Si en lugar de una covariable tuviéramos dos, ambas linealmente relacionadas con *Y*, los modelos serían

Una vía o completamente aleatorio

$$Y_{ij} = \mu + T_i + \beta_1 (X_{1ij} - \overline{X}_{1..}) + \beta_2 (X_{2ij} - \overline{X}_{2..}) + \varepsilon_{ij}$$

Dos vías o factorial sin repetición o de bloques aleatorios

$$Y_{ij} = \mu + T_i + R_j + \beta_1 (X_{1ij} - \overline{X}_{1..}) + \beta_2 (X_{2ij} - \overline{X}_{2..}) + \varepsilon_{ij}$$

Jerárquicos o con subgrupos

$$Y_{ijk} = \mu + T_i + R_{j(i)} + \beta_1 (X_{1ijk} - \overline{X}_{1...}) + \beta_2 (X_{2ijk} - \overline{X}_{2...}) + \varepsilon_{ijk}$$

Trifactorial sin repetición o Cuadrado latino

$$Y_{ijk} = \mu + T_i + F_j + C_k + \beta_1 (X_{1ijk} - \overline{X}_{1...}) + \beta_2 (X_{2ijk} - \overline{X}_{2...}) + \varepsilon_{ijk}$$

Factorial de dos factores con repetición

$$Y_{ijk} = \mu + T_i + R_j + TR_{ij} + \beta_1 (X_{1ijk} - \overline{X}_{1...}) + \beta_2 (X_{2ijk} - \overline{X}_{2...}) + \varepsilon_{ijk}$$

Es interesante rescribir los modelos lineales de covarianza de diferentes maneras. Por ejemplo, el modelo de una vía o del diseño completamente aleatorio se puede rescribir de las siguientes maneras

$$Y_{ij} - \beta (X_{ij} - \overline{X}_{..}) = \mu + T_i + \varepsilon_{ij}$$

$$Y_{ij} - T_i = \mu + \beta (X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

O el modelo de dos vías sin repetición o del diseño de bloques aleatorios se puede plantear de las siguientes maneras

$$Y_{ij} - \beta (X_{ij} - \overline{X}_{..}) = \mu + T_i + R_j + \varepsilon_{ij}$$

$$Y_{ij} - T_i - R_j = \mu + \beta (X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

La primera manera de escribir el modelo lineal, hace hincapié en los aspectos del análisis de varianza de valores que han sido ajustados para la regresión con respecto a una covariable. Se está destacando el segundo uso, que se ha descrito al principio del capítulo, aunque obviamente se tiene en mente los usos primero y tercero.

La segunda manera de plantear el modelo lineal, es con objeto de acentuar el enfoque del análisis de la regresión. Se desea medir la regresión de *Y* con respecto a *X* sin interferencias de los efectos de los tratamientos. Ahora interesa más el uso cuarto y quinto. Si *X* no se midiera, entonces $\beta(X_{ij} - X_{..})$ no se podría determinar y quedaría incluida en el error.

El mismo razonamiento sirve para dos o más variables independientes.

Veamos un ejemplo en el que desarrollemos las diferentes maneras de plantear el modelo.

Ejemplo.-

En un estudio de la ganancia de peso (Y) de lechones se probaron tres dietas. Como el peso al inicio de la experiencia influye en la ganancia de peso, se tomó este peso inicial como covariable (X)

Dietas					
A		*B*		*C*	
X	*Y*	*X*	*Y*	*X*	*Y*
14.8	4.22	15.8	3.78	19.8	4.07
18.9	3.84	19.9	4.65	19.9	4.02
19.2	4.43	21.2	4.04	19.2	4.28
19.7	4.46	15.7	3.71	16.7	4.02
21.4	4.74	19.4	4.20	19.4	4.02
19.8	4.41	19.8	4.46	19.8	4.36
17.9	4.42	17.9	4.33	15.9	4.09
16.2	4.43	19.2	4.46	15.2	3.70
18.7	4.61	18.7	4.08	20.7	4.41
16.4	4.01	18.4	4.42	15.4	3.94
183.0	43.57	186.0	42.13	182.0	40.91

Si se realizar el análisis de la varianza, tal como se estudió en el capítulo 5, de las dos variables utilizando los modelos, respectivamente

$$X_{ijk} = \mu + D_i + e_{ijk}$$
$$Y_{ijk} = \mu + D_i + e_{ijk}$$

siendo D_i las tres dietas, se tiene, para la variable X:

FV	*gl*	*SC*	*CM*	*F*
Dietas	2	0.8667	0.4333	0.11*ns*
Error	27	104.3400	3.8644	
Total	29	105.2067		

Como se esperaba, las dietas son *no significativas*, pues los tratamientos se han aplicado después de realizadas las medidas de los *pesos iniciales* y estos se han asignado aleatoriamente a cada tratamiento. Como se verá en el siguiente epígrafe, este es uno de los supuestos del análisis de covarianza.

Y el análisis de la varianza de Y es

FV	*gl*	*SC*	*CM*	*F*
Dietas	2	0.3546	0.1773	2.51*ns*
Error	27	1.9041	0.0705	
Total	29	2.2587		

También da *no significativas* las tres dieta, aunque aquí no se esperaba este resultado, pues las dietas deben influir en el incremento de peso. Por tanto, este resultado puede ser debido a falta de precisión en el error, como consecuencia de no haber utilizado el modelo lineal adecuado. El modelo que se tendría que haber utilizado es el del análisis de covarianza, es decir

$$Y_{ijk} = \mu + D_i + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

que puede utilizarse de las dos maneras vistas más arriba.

Comencemos haciendo uso de la segunda manera de presentar el modelo lineal, esta es

$$Y_{ijk} - D_i = \mu + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

que teniendo en cuenta que

$$\mu = \alpha + \beta\,\overline{X}_{...}$$

el modelo anterior queda de la forma común que se ha utilizado en el capítulo 5, esta es

$$Y_{ijk} - D_i = \alpha + \beta\, X_{ijk} + \varepsilon_{ijk}$$

Se puede, por tanto, estimar la regresión con arreglo a este modelo. Para ello se necesita obtener primero las variables corregidas, restándole a las variables originales los efectos de los tratamientos en los que se encuentran.

Los efectos de las dietas en la X son,

$$\begin{aligned} \tau_{D_1} &= 18.30 - 18.37 = -0.07 \\ \tau_{D_2} &= 18.60 - 18.37 = 0.23 \\ \tau_{D_3} &= 18.20 - 18.37 = -0.17 \end{aligned}$$

Los efectos de las dietas en la Y son

$$\tau_{D_1} = 4.36 - 4.22 = 0.14$$
$$\tau_{D_2} = 4.21 - 4.22 = -0.01$$
$$\tau_{D_3} = 4.09 - 4.22 = -0.13$$

Si ahora le restamos a cada dato el efecto del tratamiento en el que esta, tendremos los datos corregidos para estos efectos, estos datos son

Dietas					
A		*B*		*C*	
X	*Y*	*X*	*Y*	*X*	*Y*
14.87	4.08	15.57	3.79	19.97	4.20
18.97	3.70	19.67	4.66	20.07	4.15
19.27	4.29	20.97	4.05	19.37	4.31
19.77	4.32	15.47	3.72	16.87	4.15
21.47	4.60	19.17	4.21	19.57	4.15
19.87	4.27	19.57	4.47	19.97	4.49
17.97	4.28	17.67	4.34	16.07	4.22
16.27	4.29	18.97	4.47	15.37	3.83
18.77	4.47	18.47	4.09	20.87	4.54
16.47	3.87	18.17	4.43	15.57	4.07
183.7	42.17	183.7	42.23	183.7	42.11

El primer valor de la X y de la Y se han calculado así

$$X_{11} = X_{11} - \tau_{D_1}$$
$$X_{11} = 14.8 - (-0.07) = 14.87$$
$$Y_{11} = Y_{11} - \tau_{D_1}$$
$$Y_{11} = 4.22 - (0.14) = 4.08$$

Como se ve, las sumas de las tres dietas son las mismas, por lo tanto también son iguales las medias (para la Y no salen exactamente iguales como consecuencia de errores de redondeo). Esto es como consecuencia de que se han corregido los valores para los efectos de los tratamientos, eliminando estos efectos.

Si se estima la regresión con éstos datos, corregidos para los efectos de los tratamientos, se obtendría el siguiente resultado

$$b = 0.0776$$

(La regresión de los datos originales, sin corregir, vale 0.07893)

Si, ahora, se hace uso de la primera manera de presentar el modelo lineal, es decir, de

$$Y_{ijk} - \beta(X_{ijk} - \overline{X}_{...}) = \mu + D_i + \varepsilon_{ijk}$$

se esta realizando un análisis de la varianza de la parte de la variable *Y* que queda después de quitarle (ajustada) la parte de valor que es debido a la regresión con la *X*. Estos datos son

Dietas		
A	*B*	*C*
4.497	3.979	3.959
3.798	4.531	3.901
4.365	3.820	4.215
4.356	3.917	4.149
4.504	4.120	3.940
4.299	4.349	4.249
4.456	4.366	4.281
4.598	4.395	3.945
4.584	4.054	4.229
4.162	4.417	4.170
43.619	41.948	41.038

El primer dato, por ejemplo, de la dieta *A*, se calcula así

$$Y_{11} = Y_{11} - b(X_{11} - \overline{X})$$
$$Y_{11} = 4.22 - 0.0776(14.8 - 18.37) = 4.497$$

Como se ve, la suma total de estos datos ajustados es la misma que la de los datos originales, pero no son las mismas las sumas de las dietas porque se le ha quitando la variabilidad debida a la variable *X*, por lo que las estimas de los efectos de los tratamientos serán diferentes.

Si hacemos el análisis de varianza de estos datos ajustados obtenemos las siguientes sumas de cuadrados

FV	*SC*
Dietas	0.3427
Error	1.2606

la suma de cuadrados del error vale 1.2606, mientras que la suma de cuadrados del error de los datos originales, calculada anteriormente, vale 1.9041, por lo que la suma de cuadrados del error a disminuido más del 50%, aumentando, en la misma proporción, la precisión del análisis.

La diferencia entre estas dos sumas de cuadrados del error (1.9041-1.2606=0.6435) es la suma de cuadrados debida a la regresión, que tiene un solo grado de libertad que hay que quitárselo a los grados de libertad del error.

Lógicamente, también ha cambiado la suma de cuadrados debida a las dietas quedando el análisis completo de la siguiente manera.

FV	*gl*	*SC*	*CM*	*F*
Dietas	2	0.3427	0.1715	3.54*
Regresión	1	0.6435	0.6435	13.27**
Error	26	1.2606	0.0485	
Total	29	2.2468		

Como se observa, es significativo el efecto de la *dieta,* que no lo era con la variable original (sin tener en cuenta el peso inicial). Este resultado si tiene sentido. Lo que ocurría con los datos originales es que la variabilidad debida a la regresión con la covariable solapaba la variabilidad debida a la dietas lo suficiente como para que el análisis de varianza no lo detectase.

Insistimos que los grados de libertad del error son uno menos de los que tendría que ser pues se ha calculando haciendo uso de la estima de un parámetro, la *b.*

Este ejemplo se ha realizado de esta manera para comprobar empíricamente el desarrollo de los cálculo con arreglo a las dos presentaciones del mismo modelo lineal, pero no es la manera de realizar los cálculos. El desarrollo de éstos, con el mismo ejemplo, se verá en el epígrafe *Análisis de covarianza de un modelo factorial con repetición.*

Supuestos paramétricos del análisis de covarianza.-

Los supuestos del análisis de covarianza son una combinación de los supuestos del análisis de varianza y los supuestos del análisis de la regresión. Estos supuestos necesarios para el uso correcto del análisis de covarianza son:

1) Los *X* son fijos, medidos sin error, e independientes de los tratamientos.
2) La regresión de *Y* respecto a *X*, después de eliminar diferencias de tratamientos, es lineal e independiente de los tratamientos.
3) Los residuos se distribuyen normal e independientemente con media cero y varianza común.

El supuesto 1 establece que los valores de *X* son fijos. Esto quiere decir que, si se repite el experimento, se vuelven a repetir los mismos valores de *X*. O lo que es lo mismo, las inferencias sólo se aplican al conjunto de los valores de *X* realmente observados. Mientras que los *X* no se seleccionen exactamente o su lectura sea idéntica en diferentes repeticiones, entonces las inferencias se harán para valores interpolados y no para valores extrapolados. El supuesto 1 también establece que para el uso normal del análisis de covarianza, los tratamientos no afectarán a los valores de *X*.

El supuesto 2 establece que el efecto de *X* sobre *Y* es aumentar o disminuir todos los valores de *Y*, en promedio, en un múltiplo constante de la desviación del

correspondiente X respecto de la $\overline{X}$ para todo el experimento, es decir, en $b(X_{ij} - \overline{X})$. Se supone que la regresión es homogénea. Así, no se requiere subíndice en b para relacionarla con bloques o tratamientos, aunque más adelante, y tal como se avanzó en el Capítulo 11, se verá un caso en que sí se relacionará con bloques o tratamientos con objeto de realizar el contraste de homogeneidad de regresión.

El supuesto 3 es aquél del cual depende la validez de las pruebas, t y F. Un análisis, tal como lo determina el modelo, da una estimación válida de la varianza si se han aleatorizado los tratamientos. El supuesto de normalidad no es necesario para estimar los componentes de la varianza de Y, pero sí es necesaria la aleatorización.

Análisis de covarianza del modelo de una vía o del diseño completamente aleatorio.-

Cuando se tiene un solo factor los datos forman una clasificación de un sentido o una vía, siendo los tratamientos las diferentes clases o niveles del factor. Es, por tanto, la combinación del análisis de varianza de una vía con el análisis de regresión a la variable X. El modelo es

$$Y_{ij} = \mu + T_i + \beta(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

La estima de los parámetros del modelo se realiza, como siempre, mediante el método de los mínimos cuadrados de manera que debe cumplirse que la siguiente expresión sea mínima

$$\Sigma_{ij}[Y_{ij} - \mu - T_i - \beta(X_{ij} - \overline{X}_{..})]^2$$

Ajustando el modelo completo para el caso de que todas las hipótesis sean verdaderas, se tiene

$$\Sigma_{ij}\left[\hat{Y}_{ij} - \hat{\mu} - \hat{T}_i - \hat{\beta}(X_{ij} - \overline{X}_{..})\right]^2 = 0$$

Es decir, que la suma de todas las desviaciones es cero.

Por cálculo diferencial se llega al sistema de ecuaciones normales

$$\left.\begin{aligned} N\mu + \beta X_{..} &= Y_{..} \\ n_i\mu + N_i T_i + \beta X_{i.} &= Y_{i.} \\ X_{..}\mu + \Sigma_i T_i X_{i.} + \beta \Sigma_{ij} X_{ij}^2 &= \Sigma_{ij} X_{ij} Y_{ij} \end{aligned}\right\}$$

Despejando y operando se obtienen que las estimas de estos parámetros, son

$$\hat{\mu} = \overline{Y}$$

$$\hat{T}_i = T_i = \overline{Y}_{i.} - Y_{..} - b(\overline{X}_{i.} - X_{..})$$

$$\hat{\beta} = b = \frac{\sum_{ij} X_{ij} Y_{ij} - \sum_i \frac{X_{i.} Y_{i.}}{n_i}}{\sum_{ij} X_{ij}^2 - \sum_i \frac{X_{i.}^2}{n_i}}$$

La descomposición de la variación total y de los grados de libertad totales, así como los parámetros estimados y las pruebas de hipótesis, se realiza tal como se estudió en el Capítulo 5 para el *Análisis de varianza de una vía o del diseño completamente aleatorio*, con la única diferencia de que, al haber dos variables, habrá que realizar tres descomposiciones dos para cada una de las sumas de cuadrados, la de X y la de Y, y una para la suma de productos de ambas variables. Como consecuencia de ello, las anotaciones del capítulo 5 se harían muy largas, por lo que se va a utilizar otra anotación más corta, aunque con el mismo contenido del capítulo 5. Esta nueva anotación es la misma que utilizara Cochran en su trabajo de 1957, esta es

$$E_{XX} = SC_{Error,X} = \sum_{ij} X_{ij}^2 - \sum_i \frac{X_{i.}^2}{n_i}$$

$$E_{YY} = SC_{Error,Y} = \sum_{ij} Y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i}$$

$$E_{XY} = SP_{Error} = \sum_{ij} X_{ij} Y_{ij} - \sum_i \frac{X_{i.} Y_{i.}}{n_i}$$

$$gl = N - t$$

$$T_{XX} = SC_{Trata,X} = \sum_i \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$$

$$T_{YY} = SC_{Trata,Y} = \sum_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{N}$$

$$T_{XY} = SP_{Trata} = \sum_i \frac{X_{i.} Y_{i.}}{n_i} - \frac{X_{..} Y_{..}}{N}$$

$$gl = t - 1$$

$$S_{XX} = SC_{Trata,X} + SC_{Error,X} = \sum_{ij} X_{ij}^2 - \frac{X_{..}^2}{N}$$

$$S_{YY} = SC_{Trata,X} + SC_{Error,Y} = \sum_{ij} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

$$S_{XY} = SP_{Trata} + SP_{Error} = \sum_{ij} X_{ij} Y_{ij} - \frac{X_{..} Y_{..}}{N}$$

$$gl = (N - t) + (t - 1) = N - 1$$

Como se puede ver, en este caso, *S* coincide con los totales.

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con *X*.

Ahora, mediante un proceso secuencial, se le va a extraer a cada suma de cuadrados la componente debida a la regresión quedando un residuo más pequeño para el análisis de covarianza.

En el Capítulo 11 se vio que la estima de *b* es

$$b = \frac{SP}{SC_{(X)}} = \frac{S_{XY}}{S_{XX}}$$

Pero, como se ha visto en el desarrollo del primer ejemplo, la estima de *b* en el análisis de covarianza en lugar de usar, en el numerador, la *SP* total, se utiliza la *SP* que queda después de restarle a la *SP* total la *SP* debida a los tratamientos, es decir, la *SP* del error, E_{XY} según la anotación de este capítulo. Y en lugar de dividir por la *SC* total de la variable *X* se divide por la *SC* que queda después de restarle a la *SC* total de la variable *X* la *SC* debida a los tratamientos, es decir, la *SC* del error, E_{XX} según la anotación de este capítulo.

Por lo que la estima de β del modelo del análisis de covarianza es,

$$b = \frac{E_{XY}}{E_{XX}}$$

Y la fracción de la suma de cuadrados total atribuible a la regresión, corregida para tratamientos es

$$SC_{\text{Regresión}} = b\,E_{XY} = \frac{E_{XY}^2}{E_{XX}}$$

que es la equivalente a la estudiada en el Capítulo 11.

La suma de cuadrados del error de la variable bajo estudio (la *Y*) ajustada para la regresión con la covariable (la *X*) es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = E_{YY} - \frac{E_{XY}^2}{E_{XX}}$$

Si la regresión de *Y* sobre *X*, sin eliminar el efecto de los tratamientos, es

$$b_S = \frac{SP}{SC_{(X)}} = \frac{S_{XY}}{S_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_S S_{XY} = \frac{S_{XY}^2}{S_{XX}}$$

La suma de cuadrados, ajustada para la regresión con la covariable, debida al tratamiento más el error es

$$SC_{\text{Trata+Error}} = S' = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (S'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión con la covariable, atribuible a los tratamientos.

El análisis se puede resumir en la siguiente tabla

					reducción debida a ***b***		ajuste debido a ***b***	
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*
Trata	*t*-1	T_{XX}	T_{XY}	T_{YY}				
Error	*N*-t	E_{XX}	E_{XY}	E_{YY}	1	$\frac{E_{XY}^2}{E_{XY}}$	*N*-*t*-1	*E'*
Tra+Err	(*t*-1)+(*N*-*t*)	S_{XX}	S_{XY}	S_{YY}	1	$\frac{S_{XY}^2}{S_{XY}}$	*N*-2	*S'*
total	*N*-1	*SC*	*SP*	*SC*				
Tratamientos ajustados							*t*-1	*S'*-*E'*

Obsérvese que las fuentes de variación a las que se le ha restado la desviación debida a su regresión, se le resta, asimismo, a sus grados de libertad, los debidos a la regresión, es decir, uno. Las demás fuentes de variación conservan los mismos grados de libertad.

Ajuste de las medias de tratamientos.-

La fórmula para el ajuste de las medias de los tratamientos se ha dado implícitamente cuando se hallaron los parámetros al resolver las ecuaciones normales del epígrafe anterior, y es una aplicación del procedimiento estudiado en el capítulo XI. La idea básica se ilustra en la figura presentada al principio del presente capítulo. El cálculo de una media ajustada (o media minimocuadrática) de un tratamiento es

$$\hat{Y}_{i.} = \overline{Y}_{i.} - b(\overline{X}_{i.} - \overline{X}_{..})$$

donde *b* es el coeficiente de regresión del error.

Las medias ajustadas de los tratamientos son estimaciones de lo que serían las medias de los tratamientos si todas las $\overline{X}_{i.}$ coincidieran con $\overline{X}_{..}$

El *error típico de la media ajustada* de un tratamiento es una simple modificación de la ecuación estudiada en el epígrafe *Fuentes de variación en la regresión* del Capítulo XI, esta es

$$S_{\hat{Y}_{i.}} = S_{Y.X} \sqrt{\frac{1}{n} + \frac{(\overline{X}_{i.} - \overline{X}_{..})^2}{E_{XX}}}$$

siendo $S_{Y.X}$ la raíz cuadrada del cuadrado medio del error ajustado.

La diferencia entre las medias ajustadas del *i-ésimo* y *j-ésimo* tratamiento sería

$$\hat{Y}_{i.} - \hat{Y}_{j.} = \overline{Y}_{i.} - \overline{Y}_{j.} - b(\overline{X}_{i.} - \overline{X}_{j.})$$

Y el *error típico de la diferencia de las medias ajustadas* de dos tratamientos es

$$S_{\hat{Y}_{i.}-\hat{Y}_{j.}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n} + \frac{(\overline{X}_{i.} - \overline{X}_{j.})^2}{E_{XX}}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_i} + \frac{1}{n_j} + \frac{(\overline{X}_{i.} - \overline{X}_{j.})^2}{E_{XX}}\right]}$$

La expresión de la derecha es para el caso en que los tamaños de submuestras para los diferentes tratamientos sean diferentes.

Esta ecuación exige un cálculo separado para cada comparación de dos en dos tratamientos. En la práctica existen poca diferencia para los errores típicos de todos los pares de tratamientos si se cumple que hay un mínimo de *gl*=20 para el *error* del ANOVA y el cuadrado medio de los *tratamientos* de la variable *X* es no significativo, como tiene que serlo, puesto que la *X* es independiente de los tratamientos. En el caso de que se cumplan estas condiciones, se sugiere una aproximación a $S_{\hat{Y}_{i.}-\hat{Y}_{j.}}$ que utiliza un promedio de todas las medias en vez de las $(\overline{X}_{i.} - \overline{X}_{j.})$ por separado. En este caso la expresión es

$$S_{\hat{Y}_{i.}-\hat{Y}_{j.}} = \sqrt{\frac{2S^2_{Y.X}}{n}\left[1 + \frac{T_{XX}}{(t-1)E_{XX}}\right]}$$

aunque siempre será mas correcta la utilización de la fórmula exacta.

Con estas expresiones se pueden hacer comparaciones múltiples de medias tal como se estudió en el Capítulo X, teniendo en cuenta que hay que usar el cuadrado medio del error del análisis de covarianza y los grados de libertad de dicho error. Las pruebas

más correctas en este caso son la *t* y la de *Scheffe*, si bien la más comúnmente utilizada es la prueba *t*.

Pruebas de hipótesis.-

Si queremos probar la hipótesis de igualdad de efectos de los diferentes tratamientos sin ajustar para la regresión, es decir, como si, desconociendo la variable *X*, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{T_{YY}}{t-1}}{\frac{E_{YY}}{N-t}}$$

que se contrastaría con la $F_{(t-1,N-t;\ \alpha)}$

Si la prueba que se desea hacer es la de igualdad de efectos de los tratamientos pero ajustada para la regresión con la covariable, es decir, la prueba *F* de las medias ajustadas, a semejanza de la anterior, la F_o sería

$$F_o = \frac{\frac{S'-E'}{t-1}}{\frac{E'}{N-t-1}}$$

que se contrastaría con la $F_{(t-1,N-t-1;\ \alpha)}$

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, a semejanza de la prueba *F* del Capítulo XI, la prueba *F* en este caso sería

$$F_o = \frac{\frac{E_{XY}^2}{E_{XX}}}{\frac{E'}{(N-t-1)}}$$

que se contrastaría con la $F_{(1,N-t-1;\ \alpha)}$

Si se quiere probar si los tratamientos afectan a la covariable, la prueba sería

$$F_o = \frac{\frac{T_{XX}}{t-1}}{\frac{E_{XX}}{N-t}}$$

que se contrastaría con la $F_{(t-1,N-t;\ \alpha)}$

Aumento de precisión debido a la covarianza.-

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de tratamientos antes y después del ajuste.

El cuadrado medio del error antes del ajuste es

$$\frac{E_{YY}}{N-t}$$

El cuadrado medio del error efectivo después del ajuste para X es

$$S^2_{Y.X} = S^2_{Y.X}\left[1+\frac{T_{XX}}{(t-1)E_{XX}}\right]$$

Un estimador de la precisión relativa es la razón del cuadrado medio del error sin ajustar por el cuadrado medio del error ajustado, multiplicado por 100 para expresarlo en porcentajes.

$$\frac{\frac{E_{YY}}{N-t}}{S^2_{Y.X}}100$$

Ejemplo.-

Resuélvase el mismo ejemplo utilizado para explicar los fundamentos del análisis de covarianza, este era:

En un estudio de la ganancia de peso (Y) de lechones se probaron tres dietas. Como el peso al inicio de la experiencia influye en la ganancia de peso, se tomó este peso inicial como covariable (X)

	Dietas							
	A		*B*		*C*		*Global*	
	X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
	14.8	4.22	15.8	3.78	19.8	4.07		
	18.9	3.84	19.9	4.65	19.9	4.02		
	19.2	4.43	21.2	4.04	19.2	4.28		
	19.7	4.46	15.7	3.71	16.7	4.02		
	21.4	4.74	19.4	4.20	19.4	4.02		
	19.8	4.41	19.8	4.46	19.8	4.36		
	17.9	4.42	17.9	4.33	15.9	4.09		
	16.2	4.43	19.2	4.46	15.2	3.70		
	18.7	4.61	18.7	4.08	20.7	4.41		
	16.4	4.01	18.4	4.42	15.4	3.94		
Σ	183.0	43.57	186.0	42.13	182.0	40.91	551.00	126.61
Σ^2	3384.48	190.48	3487.28	178.35	3353.48	167.76	10225.24	536.59
ΣXY	799.559		786.705		747.444		2333.708	
$\overline{m}$	18.30	4.357	18.6	4.213	18.2	4.091	18.366	4.22

El primer paso es calcular las sumas de cuadrados y de productos

$$E_{XX} = \Sigma_{ij} X_{ij}^2 - \Sigma_i \frac{X_{i.}^2}{n_i} = 10225.24 - \frac{183^2+186^2+182^2}{10} = 104.34$$

$$E_{YY} = \Sigma_{ij} Y_{ij}^2 - \Sigma_i \frac{Y_{i.}^2}{n_i} = 536.59 - \frac{43.57^2+42.13^2+40.91^2}{10} = 1.8990$$

$$E_{XY} = \Sigma_{ij} X_{ij} Y_{ij} - \Sigma_i \frac{X_{i.} Y_{i.}}{n_i} = 2333.708 - \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} =$$

$$= 8.197$$

$$gl = N - t = 30 - 3 = 27$$

$$T_{XX} = \Sigma_i \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} = \frac{183^2+186^2+182^2}{10} - \frac{551^2}{30} = 0.8667$$

$$T_{YY} = \Sigma_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{N} = \frac{43.57^2+42.13^2+40.91^2}{10} - \frac{126.61^2}{30} = 0.3546$$

$$T_{XY} = \Sigma_i \frac{X_{i.} Y_{i.}}{n_i} - \frac{X_{..} Y_{..}}{N} = \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} - \frac{551\times 126.61}{30} =$$

$$= 0.1073$$

$$gl = t - 1 = 3 - 1 = 2$$

$$S_{XX} = 104.34+0.8667 = 103.4733$$

$$S_{YY} = 1.899+0.3546 = 2.2536$$

$$S_{XY} = 8.197+ 0.1073 = 8.3043$$

$$gl = (N-t)+(t-1) = 27+2 = 29$$

El valor de b es

$$b = \frac{E_{XY}}{E_{XX}} = \frac{8.197}{104.34} = 0.0786$$

Y la contribución a la suma de cuadrados atribuible a la regresión es

$$SC_{\text{Regresión}} = b\,E_{XY} = \frac{E_{XY}^2}{E_{XX}} = \frac{8.197^2}{104.34} = 0.6439$$

La suma de cuadrados del error ajustada para la regresión es

$$E' = E_{YY} - \frac{E_{XY}^2}{E_{XX}} = 1.899 - 0.6439 = 1.2551$$

La regresión de Y sobre X, sin eliminar el efecto de los tratamientos, es

$$b = \frac{S_{XY}}{S_{XX}} = \frac{8.3043}{103.4733} = 0.8025$$

por lo que la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b\, S_{XY} = \frac{S_{XY}^2}{S_{XX}} = \frac{8.3043^2}{103.4733} = 0.6665$$

La suma de cuadrados, ajustada para la regresión, debida al tratamiento más el error es

$$S' = S_{YY} - \frac{S_{XY}^2}{S_{XX}} = 2.2536 - 0.6665 = 1.5871$$

Por lo que la suma de cuadrados de los tratamientos ajustados es la diferencia de esta última con la del error

$$S' - E' = 1.5871 - 1.2551 = 0.3320$$

Estos cálculos se pueden resumir en la siguiente tabla

					reducción debida a b		*ajuste debido a b*		
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*	*CM*
Trata	2	0.867	0.107	0.355					
Error	27	104.34	8.197	1.899	1	0.439	26	1.255	0.048
Tra+Err	29	103.47	8.304	2.254	1	0.678	28	1.587	
total	29	103.47	8.304	2.254					
Tratamientos ajustados							2	0.3320	0.166

Estas sumas de cuadrados y de productos contienen todos los elementos esenciales para el análisis de la varianza de X y de Y y para el análisis de covarianza.

Así, para probar el supuesto de que los tratamientos no afectan a la variable X, puesto que los tratamientos se han aplicado después de medir los valores de X, se aría con el ANOVA de esta variable cuya F es la siguiente

$$F_o = \frac{\frac{0.8667}{2}}{\frac{104.34}{27}} = 0.11ns$$

$$F_{(2,27;\,0.05)} = 3.35$$

es decir, como es de esperar, los tratamientos no influyen en la variable *X*. Al cumplirse este supuesto es correcto realizar el análisis de covarianza.

Para probar la hipótesis de que no hay diferencias entre las medias no ajustadas de los tratamientos para la cantidad de bacilos (*Y*), como si se hubiera planteado la experiencia sin tener en cuenta la variable independiente (*X*), hubiera sido un ANOVA cuya *F* sería

$$F_o = \frac{\frac{0.3546}{2}}{\frac{1.899}{27}} = 2.52ns$$

$$F_{(2,27;\,0.05)} = 3.35$$

Por lo que se concluye que existen diferencias entre las ganancias de peso de las tres dietas (sin tener en cuenta el peso inicial).

Pero como la ganancia de peso al final de los tratamientos puede ser función, además de los tratamientos, del peso al comienzo del tratamiento (*covariable*) es por lo que lo correcto es contrastar las sumas de cuadrados ajustadas para la regresión de *Y* con la variable *X*.

$$F_o = \frac{0.166}{0.048} = 3.46\,*$$

$$F_{(2,26;\,0.05)} = 3.37$$

el valor de *F* ha pasado de 2.52*ns* a 3.46*, por lo que se concluye que existen diferencias entre las medias de la ganancia de peso final ajustadas para el peso inicial. O, en este caso concreto, son efectivas las dietas.

Ahora se puede probar la significación del coeficiente de regresión (H_o:β=0)

$$F_o = \frac{\frac{E^2_{XY}}{E_{XX}}}{\frac{E'}{(N\text{-}t\text{-}1)}} = \frac{0.6439}{\frac{1.255}{26}} = 13.339$$

$$F_{(1,26;\,0.01)} = 7.72$$

Existe una regresión significativa entre ambas variables.

La regresión de *Y* con respecto a *X* sin tener en cuenta los tratamientos es

$$b = \frac{S_{XY}}{S_{XX}} = \frac{8.304}{103.4733} = 0.803$$

Y su prueba de significación

$$F_o = \frac{\frac{S^2_{XY}}{S_{XX}}}{\frac{S'}{(N\text{-}1)}} = \frac{0.6664}{\frac{1.5871}{28}} = 11.75^{**}$$

$$F_{(1,28;\,0.01)} = 7.64$$

Las medias ajustadas de los tratamientos o medias minimocuadráticas son

$$A: \hat{Y}_{1.} = \overline{Y}_{1.} - b(\overline{X}_{1.} - \overline{X}_{..}) = 4.357 - 0.0786(18.3 - 18.37) = 4.36$$

$$B: \hat{Y}_{2.} = \overline{Y}_{2.} - b(\overline{X}_{2.} - \overline{X}_{..}) = 4.21 - 0.0786(18.6 - 18.37) = 4.19$$

$$C: \hat{Y}_{3.} = \overline{Y}_{3.} - b(\overline{X}_{3.} - \overline{X}_{..}) = 4.09 - 0.0786(18.2 - 18.37) = 4.10$$

Y sus errores típicos son

$$S_{\hat{Y}_{1.}} = \sqrt{0.048\left[\frac{1}{10} + \frac{(18.3 - 18.37)^2}{8.197}\right]} = 0.0695$$

$$S_{\hat{Y}_{2.}} = \sqrt{0.048\left[\frac{1}{10} + \frac{(18.6 - 18.37)^2}{8.197}\right]} = 0.0715$$

$$S_{\hat{Y}_{3.}} = \sqrt{0.048\left[\frac{1}{10} + \frac{(18.2 - 18.37)^2}{8.197}\right]} = 0.0705$$

Para comparar el tratamiento *A* con el *B*, por ejemplo, se tiene que la diferencia de estas dos medias ajustadas es

$$\hat{Y}_A - \hat{Y}_B = 4.357 - 4.213 - 0.0786(18.3 - 18.6) = 0.1676$$

Y el error típico de la diferencias de estas medias ajustadas es

$$S_{\hat{Y}_{1.} - \hat{Y}_{2.}} = \sqrt{0.048\left[\frac{2}{10} + \frac{(18.3 - 18.6)^2}{8.197}\right]} = 0.1006$$

Con este error típico se puede realizar la prueba *t*

$$t = \frac{4.36 - 4.19}{0.1006} = 1.689 ns$$

$$t_{(26;\,0.05/2)} = 2.0555$$

Es evidente que la diferencia entre las medias ajustadas del tratamiento *A* y del

tratamiento *B*, (4.36 y 4.19) no es significativa.

La diferencia mayor se encuentra entre las medias ajustadas de los tratamientos *A* y *C*. La prueba para esta diferencia sería

$$t = \frac{4.36 - 4.10}{\sqrt{0.048\left[\frac{2}{10} + \frac{(18.3 - 18.2)^2}{8.197}\right]}} = \frac{0.26}{0.09828} = 2.646^*$$

$$t_{(26;\,0.05/2)} = 2.0555$$

por lo que existe diferencia significativa entre ambas dietas.

La cuestión que se puede plantear ahora es de cuál ha sido el aumento de precisión del análisis de covarianza con respecto al ANOVA. El cuadrado medio del error efectivo después del ajuste por *X* es

$$S^2_{Y.X} = S^2_{Y.X}\left[1 + \frac{T_{XX}}{(t-1)E_{XX}}\right] = 0.048\left[1 + \frac{0.8667}{2 \times 8.197}\right] = 0.0505$$

por lo que el aumento de precisión relativa es

$$\frac{\frac{E_{YY}}{N-t}}{S^2_{Y.X}} \times 100 = \frac{\frac{1.899}{30-3}}{0.0505} 100 = 139.3\%$$

Una aumento bastante aceptable.

Archivo del programa *SAS* (C17-1.SAS).-

```
title 'Análisis de covarianza de una vía';
options ls=75 ps=60;
data ancova;
infile 'c17-1.dat';
input dieta $ x y @@;
proc anova;
  class dieta;
  model x = dieta;
proc anova;
  class dieta;
  model y = dieta;
proc glm;
  class dieta;
   model y=dieta x / solution;
   lsmeans dieta / stderr tdiff;
run;
```

Archivo de datos (C17-1.DAT).-

```
A 14.8  4.22  B 15.8  3.78  C 19.8  4.07
A 18.9  3.84  B 19.9  4.65  C 19.9  4.02
A 19.2  4.43  B 21.2  4.04  C 19.2  4.28
A 19.7  4.46  B 15.7  3.71  C 16.7  4.02
```

```
A 21.4   4.74   B 19.4   4.20   C 19.4   4.02
A 19.8   4.41   B 19.8   4.46   C 19.8   4.36
A 17.9   4.42   B 17.9   4.33   C 15.9   4.09
A 16.2   4.43   B 19.2   4.46   C 15.2   3.70
A 18.7   4.61   B 18.7   4.08   C 20.7   4.41
A 16.4   4.01   B 18.4   4.42   C 15.4   3.94
```

Archivo de resultados (C17-1.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: X

                                   Sum of          Mean
Source                   DF       Squares        Square  F Value     Pr > F
Model                     2    0.86666667    0.43333333     0.11     0.8943
Error                    27   104.34000000    3.86444444
Corrected Total          29   105.20666667

                  R-Square          C.V.      Root MSE              X Mean
                  0.008238      10.70319       1.96582             18.3667

Source                   DF      Anova SS   Mean Square  F Value     Pr > F
DIETA                     2    0.86666667    0.43333333     0.11     0.8943

                    Analysis of Variance Procedure
Dependent Variable: Y

                                   Sum of          Mean
Source                   DF       Squares        Square  F Value     Pr > F
Model                     2    0.35458667    0.17729333     2.51     0.0997
Error                    27    1.90411000    0.07052259
Corrected Total          29    2.25869667

                  R-Square          C.V.      Root MSE              Y Mean
                  0.156987      6.292415       0.26556             4.22033

Source                   DF      Anova SS   Mean Square  F Value     Pr > F
DIETA                     2    0.35458667    0.17729333     2.51     0.0997

                    General Linear Models Procedure
Dependent Variable: Y

                                   Sum of          Mean
Source                   DF       Squares        Square  F Value     Pr > F
Model                     3    0.99854688    0.33284896     6.87     0.0015
Error                    26    1.26014978    0.04846730
Corrected Total          29    2.25869667

                  R-Square          C.V.      Root MSE              Y Mean
                  0.442090      5.216481       0.22015             4.22033

Source                   DF     Type I SS   Mean Square  F Value     Pr > F
DIETA                     2    0.35458667    0.17729333     3.66     0.0398
X                         1    0.64396022    0.64396022    13.29     0.0012

Source                   DF    Type III SS   Mean Square  F Value     Pr > F
DIETA                     2    0.34305656    0.17152828     3.54     0.0437
X                         1    0.64396022    0.64396022    13.29     0.0012

                                  T for H0:     Pr > |T|    Std Error of
Parameter          Estimate     Parameter=0                    Estimate
INTERCEPT      2.661199348 B          6.68        0.0001      0.39838682
DIETA     A    0.258143952 B          2.62        0.0144      0.09847896
          B    0.090575810 B          0.92        0.3678      0.09883209
          C    0.000000000 B             .             .               .
X              0.078560475          3.65        0.0012      0.02155257
NOTE: The X'X matrix has been found to be singular and a generalized
      inverse was used to solve the normal equations.   Estimates followed
      by the letter 'B' are biased, and are not unique estimators of the
      parameters.
```

```
                          Least Squares Means
DIETA                      Y        Std Err      Pr > |T|    LSMEAN
                        LSMEAN       LSMEAN     H0:LSMEAN=0   Number
A                   4.36223737    0.06963329          0.0001      1
B                   4.19466922    0.06979986          0.0001      2
C                   4.10409341    0.06971107          0.0001      3

                T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|
                      i/j            1             2             3
                      1              .         1.698312      2.621311
                                               0.1014        0.0144
                      2        -1.69831            .       0.916462
                                0.1014                       0.3678
                      3        -2.62131      -0.91646            .
                                0.0144        0.3678
NOTE: To ensure overall protection level, only probabilities associated
      with pre-planned comparisons should be used.
```

Estos son los mismos valores que los obtenidos manualmente, sin tanto error de redondeo.

La suma de cuadrados que hay que mirar es la **tipo III**. La suma de cuadrados **tipo I** debida al factor (dieta) es la del ANOVA para la variable *Y*. Si, en el modelo del programa SAS, se pone primero la covariable y después el factor, esto es, **model y = x dieta / solution;** se hubiera obtenido la misma suma de cuadrados, debida al factor, en el tipo I y en el tipo III, pero lo suma de cuadrados tipo I debida a la regresión, sería la de la regresión sin considerar las correcciones del factor, esto es, la de la regresión de los treinta pares de valores, como si las dietas no existieran.

Análisis de covarianza del modelo factorial sin repetición o del diseño de bloques aleatorios.-

No existe ninguna novedad con respecto a lo estudiado en el modelo de una vía. La descomposición de la variación total y de los grados de libertad totales, así como los parámetros estimados y las pruebas de hipótesis, se realiza tal como se estudió en el Capítulo 5 para el *Análisis Factorial sin repetición*, con la única diferencia de que, al haber dos variables, habrá que realizar tres descomposiciones dos para cada una de las sumas de cuadrados, la de *X* y la de *Y*, y una para la suma de productos de ambas variables. Como consecuencia de ello, las anotaciones del capítulo 7 se hacen muy largas, por lo que se va a utilizar el tipo de anotación visto en el anterior modelo.

La regresión se calcula en la fila del error y la suma de cuadrados ajustadas de ambas fuentes de variación se calcularan restándole a la suma de cuadrados ajustada del factor más el error, la suma de cuadrados ajustada del error.

A los diferentes niveles del primer factor vamos a seguir denominándoles *tratamientos* y la los diferentes niveles del segundo factor, muchas veces, denominaremos como *bloques*, pues estos modelos casi siempre responden a una experiencia en bloques aleatorios o de datos relacionados (datos emparejados si hubiera dos tratamientos). En el caso de que los diferentes niveles del segundo factor no sean bloques, sino otros tratamientos o un factor de clasificación, todo se realiza exactamente igual a como se va ha describir a continuación.

Seguimos utilizando la simbología de *T* para el primer factor y de *R* para el segundo factor, proveniente de los ejemplos de los capítulos del análisis de varianza. Naturalmente, estas anotaciones pueden cambiar en cada ejemplo con arreglo a cuales sean las denominaciones de los factores.

El modelo lineal que se va ha utilizar es

$$Y_{ij} = \mu + T_i + R_j + \beta(X_{ij} - \overline{X}_{..}) + \varepsilon_{ij}$$

A pesar de haber dos factores cruzados, no se incluye la interacción en el modelo porque no se puede estimar, ya que al haber un solo valor por casilla, se confunde la interacción con el error (ver capítulo 5). En el caso de que exista duda del supuesto de aditividad, se puede realizar la *Prueba de no aditividad* de Tukey que se describió en el capítulo 6.

La descomposición de la variación total y de los grados de libertad total es

$$E_{XX} = \Sigma_{ij} X_{ij}^2 - \Sigma_i \frac{X_{i.}^2}{r} - \Sigma_j \frac{X_{.j}^2}{t} + \frac{X_{..}^2}{tr}$$

$$E_{YY} = \Sigma_{ij} Y_{ij}^2 - \Sigma_i \frac{Y_{i.}^2}{r} - \Sigma_j \frac{Y_{.j}^2}{t} + \frac{Y_{..}^2}{tr}$$

$$E_{XY} = \Sigma_{ij} X_{ij} Y_{ij} - \Sigma_i \frac{X_{i.} Y_{i.}}{r} - \Sigma_j \frac{X_{.j} Y_{.j}}{t} + \frac{X_{..} Y_{..}}{tr}$$

$$gl = (t-1)(r-1)$$

$$T_{XX} = \Sigma_i \frac{X_{i.}^2}{r} - \frac{X_{..}^2}{tr}$$

$$T_{YY} = \Sigma_i \frac{Y_{i.}^2}{r} - \frac{Y_{..}^2}{tr}$$

$$T_{XY} = \Sigma_i \frac{X_{i.} Y_{i.}}{r} - \frac{X_{..} Y_{..}}{tr}$$

$$gl = t-1$$

$$R_{XX} = \Sigma_i \frac{X_{.j}^2}{t} - \frac{X_{..}^2}{tr}$$

$$R_{YY} = \Sigma_i \frac{Y_{.j}^2}{t} - \frac{Y_{..}^2}{tr}$$

$$R_{XY} = \Sigma_i \frac{X_{.j} Y_{.j}}{t} - \frac{X_{..} Y_{..}}{tr}$$

$$gl = r-1$$

La suma de cuadrados y la suma de productos de los tratamientos más el error (*S*)

es

$$S_{XX} = E_{XX} + T_{XX} = \Sigma_{ij} X_{ij}^2 - \frac{X_{.j}^2}{t}$$

$$S_{YY} = E_{YY} + T_{YY} = \Sigma_{ij} Y_{ij}^2 - \frac{Y_{.j}^2}{t}$$

$$S_{XY} = E_{XY} + T_{XY} = \Sigma_{ij} X_{ij} Y_{ij} - \frac{X_{.j} Y_{.j}}{t}$$

$$gl = r(t-1)$$

Como se puede ver, en este caso S no coincide con los totales que son

$$\text{total}_{XX} = \Sigma_{ij} X_{ij}^2 - \frac{X_{..}^2}{tr}$$

$$\text{total}_{YY} = \Sigma_{ij} Y_{ij}^2 - \frac{Y_{..}^2}{tr}$$

$$\text{total}_{XY} = \Sigma_{ij} X_{ij} Y_{ij} - \frac{X_{..} Y_{..}}{tr}$$

$$gl = tr - 1$$

En el caso de que se este interesado también en el análisis de covarianza del segundo factor, se puede hallar, así mismo, las sumas de cuadrados y la suma de productos de dicho factor más el error (a esta suma se le puede simbolizar como U)

$$U_{XX} = E_{XX} + R_{XX} = \Sigma_{ij} X_{ij}^2 - \frac{X_{i.}^2}{r}$$

$$U_{YY} = E_{YY} + R_{YY} = \Sigma_{ij} Y_{ij}^2 - \frac{Y_{i.}^2}{r}$$

$$U_{XY} = E_{XY} + R_{XY} = \Sigma_{ij} X_{ij} Y_{ij} - \frac{X_{i.} Y_{i.}}{r}$$

$$gl = r(t-1)$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con X.

Ahora, y tal como se hizo en el análisis de una vía, mediante un proceso secuencial, se le va a extraer a cada suma de cuadrados la componente debida a la regresión quedando un residuo más pequeño.

La estima de β del modelo del análisis de covarianza es, como ya se ha visto antes

$$b = \frac{E_{XY}}{E_{XX}}$$

es decir, que la estima de b en el análisis de covarianza se realiza como en el capítulo 11, pero en lugar de usar, en el numerador, la *SP* total, se utiliza la *SP* que queda después de restarle a la *SP* total la *SP* debida a los dos factores, es decir, la *SP* del error, E_{XY} según la anotación de este capítulo. Y en lugar de dividir por la *SC* total de la variable *X* se divide por la *SC* que queda después de restarle a la *SC* total de la variable *X* la *SC* debida a los dos factores, es decir, la *SC* del error, E_{XX} según la anotación de este capítulo.

Por lo que la contribución a la suma de cuadrados total, atribuible a la regresión con los datos de *Y* corregidos para los efectos de los dos factores es

$$SC_{\text{Regresión}} = b\,E_{XY} = \frac{E_{XY}^2}{E_{XX}}$$

La suma de cuadrados del error de la variable bajo estudio (la *Y*) ajustada para la regresión con la covariable (la *X*) es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = E_{YY} - \frac{E_{XY}^2}{E_{XX}}$$

La regresión de *Y* sobre *X*, pero corrigiendo los valores de *Y* solo para los efectos de los *tratamientos* e ignorando el segundo factor es

$$b_S = \frac{S_{XY}}{S_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_S\,S_{XY} = \frac{S_{XY}^2}{S_{XX}}$$

La fracción de la suma de cuadrados total, ajustada para la regresión, debida al tratamiento más el error es

$$SC_{\text{Trata+Error}} = S' = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (S'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a los tratamientos.

Si la regresión de *Y* sobre *X*, corrigiendo los valores de *Y* para los efecto de los

diferentes niveles del segundo factor, ignorando los efectos de los *tratamientos* es

$$b_U = \frac{U_{XY}}{U_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_U U_{XY} = \frac{U_{XY}^2}{U_{XX}}$$

La fracción de la suma de cuadrados, ajustada para la regresión, debida a los bloques más el error es

$$SC_{\text{Bloques+Error}} = U' = U_{YY} - \frac{U_{XY}^2}{U_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (U'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a los bloques.

El análisis se puede resumir en la siguiente tabla

					reducción debida a ***b***		*ajuste debido a* ***b***	
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*
Trata	t-1	T_{XX}	T_{XY}	T_{YY}				
Bloques	r-1	R_{XX}	R_{XY}	R_{YY}				
Error	(r-1)(t-1)	E_{XX}	E_{XY}	E_{YY}	1	$\frac{E_{XY}^2}{E_{XY}}$	tr-t-r	E'
Tra+Err	$r(t$-1)	S_{XX}	S_{XY}	S_{YY}	1	$\frac{S_{XY}^2}{S_{XY}}$	rt-r-1	S'
Blo+Err	$t(r$-1)	U_{XX}	U_{XY}	U_{YY}	1	$\frac{U_{XY}^2}{U_{XY}}$	tr-t-1	U'
total	rt-1	*SC*	*SP*	*SC*				
Tratamientos ajustados							t-1	S'-E'
Bloques ajustados							r-1	U'-E'

Ajustes de las medias de los tratamientos y de los bloques.-

Las fórmulas para el ajuste de las medias de los tratamientos y de los bloques son las mismas de las dadas para el modelo de una vía. El cálculo de una media ajustada (o media minimocuadrática) de un tratamiento y de un bloque son, respectivamente

Medias ajustadas de tratamientos

$$\hat{Y}_{i.} = \overline{Y}_{i.} - b(\overline{X}_{i.} - \overline{X}_{..})$$

Medias ajustadas de bloques

$$\hat{Y}_{.j} = \overline{Y}_{.j} - b(\overline{X}_{.j} - \overline{X}_{..})$$

donde *b* es el coeficiente de regresión del error.

Las medias ajustadas de los tratamientos y de los bloques son estimaciones de lo que serían las medias de los tratamientos o de los bloques si todas las $\overline{X}_{i.}$ o la $\overline{X}_{.i}$ coincidieran con $\overline{X}_{..}$, respectivamente.

Los *errores típico de las medias ajustadas* de los tratamientos y de los bloques son, respectivamente

Tratamientos

$$S_{\hat{Y}_{i.}} = S_{Y.X}\sqrt{\frac{1}{n} + \frac{(\overline{X}_{i.} - \overline{X}_{..})^2}{E_{XX}}}$$

Bloques

$$S_{\hat{Y}_{.j}} = S_{Y.X}\sqrt{\frac{1}{n} + \frac{(\overline{X}_{.j} - \overline{X}_{..})^2}{E_{XX}}}$$

siendo $S_{Y.X}$ la raíz cuadrada del cuadrado medio del error ajustado.

La diferencia entre las medias ajustadas del *i-ésimo* y *j-ésimo* tratamiento o bloque sería, respectivamente

Diferencias entre tratamientos

$$\hat{Y}_{i.} - \hat{Y}_{j.} = \overline{Y}_{i.} - \overline{Y}_{j.} - b(\overline{X}_{i.} - \overline{X}_{j.})$$

Diferencias entre bloques

$$\hat{Y}_{.i} - \hat{Y}_{.j} = \overline{Y}_{.i} - \overline{Y}_{.j} - b(\overline{X}_{.i} - \overline{X}_{.j})$$

Y el *error típico de la diferencia de las medias ajustadas* de dos tratamientos o dos bloques es, respectivamente

$$S_{\hat{Y}_{i.}-\hat{Y}_{j.}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n_t}+\frac{(\overline{X}_{i.}-\overline{X}_{j.})^2}{E_{XX}}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{t_{i.}}}+\frac{1}{n_{t_{j.}}}+\frac{(\overline{X}_{i.}-\overline{X}_{j.})^2}{E_{XX}}\right]}$$

$$S_{\hat{Y}_{.i}-\hat{Y}_{.j}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n_b}+\frac{(\overline{X}_{.i}-\overline{X}_{.j})^2}{E_{XX}}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{b.i}}+\frac{1}{n_{b.j}}+\frac{(\overline{X}_{.i}-\overline{X}_{.j})^2}{E_{XX}}\right]}$$

La expresión de la derecha es para el caso en que los tamaños de submuestras para los diferentes tratamientos (n_t) o bloques (n_b) sean diferentes.

En el caso de que se cumpla que el error tenga un mínimo de gl=20 y el cuadrado medio de los tratamientos o los bloques (según las medias que se deseen comparar) de la variable X sea no significativa, se sugiere una aproximación a $S_{\hat{Y}_{i.}-\hat{Y}_{j.}}$ que utiliza un promedio de todas las medias en vez de las ($\overline{X}_{i.}-\overline{X}_{j.}$) o las ($\overline{X}_{.i}-\overline{X}_{.j}$) por separado. En este caso la expresión es, respectivamente

$$S_{\hat{Y}_{i.}-\hat{Y}_{j.}} = \sqrt{\frac{2S^2_{Y.X}}{n_t}\left[1+\frac{T_{XX}}{(t-1)E_{XX}}\right]}$$

$$S_{\hat{Y}_{.i}-\hat{Y}_{.j}} = \sqrt{\frac{2S^2_{Y.X}}{n_b}\left[1+\frac{T_{XX}}{(r-1)E_{XX}}\right]}$$

aunque siempre será mas correcta la utilización de la fórmula exacta.

Con estas expresiones se pueden hacer comparaciones múltiples de medias tal como se estudió en el Capítulo 10, teniendo en cuenta que hay que usar el cuadrado medio del error del análisis de covarianza y los grados de libertad de dicho error. Las pruebas más correctas en este caso son la *t* y la de *Scheffe*, si bien la más comúnmente utilizada es la prueba *t*.

Pruebas de hipótesis.-

Si se quiere probar el supuesto de que la covariable no esta influida por los efectos de los tratamientos y de los bloques sería realizar un ANOVA, que con los resultados de la tabla anterior, se realizarían con las siguientes F_o. Para los tratamientos

$$F_o = \frac{\frac{T_{XX}}{t-1}}{\frac{E_{XX}}{tr-t-r+1}}$$

que se contrastaría con la $F_{(t-1,tr-t-r+1;\,\alpha)}$

Y para los bloques la F_o sería

$$F_o = \frac{\frac{R_{XX}}{r-1}}{\frac{E_{XX}}{tr-t-r+1}}$$

que se contrastaría con la $F_{(r-1,tr-t-r+1;\,\alpha)}$

Si se quiere probar la hipótesis de igualdad de efectos de los diferentes tratamientos sin ajustar para la regresión, es decir, como si, desconociendo la variable X, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{T_{YY}}{t-1}}{\frac{E_{YY}}{tr-t-r+1}}$$

que se contrastaría con la $F_{(t-1,tr-t-r+1;\,\alpha)}$

Si se quieren probar los bloques en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{R_{YY}}{r-1}}{\frac{E_{YY}}{tr-t-r+1}}$$

que se contrastaría con la $F_{(r-1,tr-t-r+1;\,\alpha)}$

Si la prueba que se desea hacer es la de igualdad de efectos de los tratamientos pero ajustada para la regresión con la covariable, es decir, la prueba F de las medias ajustadas, a semejanza de la anterior, la F_o sería

$$F_o = \frac{\frac{S'-E'}{t-1}}{\frac{E'}{tr-t-r}}$$

que se contrastaría con la $F_{(t-1,tr-t-r;\,\alpha)}$

Si se quieren probar los bloques ajustados, la pruebas sería

$$F_o = \frac{\frac{U'-E'}{r-1}}{\frac{E'}{tr-t-r}}$$

que se contrastaría con la $F_{(r-1,tr-t-r;\,\alpha)}$

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, a semejanza de la prueba F del Capítulo XI, la prueba F en este caso sería

$$F_o = \frac{\frac{E_{XY}^2}{E_{XX}}}{\frac{E'}{(tr-t-r)}}$$

que se contrastaría con la $F_{(1,tr-t-r;\,\alpha)}$

Aumento de precisión debido a la covarianza.-

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de tratamientos antes y después del ajuste.

El cuadrado medio del error antes del ajuste es

$$\frac{E_{YY}}{tr-t-r+1}$$

El cuadrado medio del error efectivo después del ajuste para X es para los tratamientos y los bloques, respectivamente

Para los tratamientos

$$S^2_{Y.X(Tra)} = S^2_{Y.X}\left[1+\frac{T_{XX}}{(t-1)E_{XX}}\right]$$

Para los bloques

$$S^2_{Y.X(Blo)} = S^2_{Y.X}\left[1+\frac{R_{XX}}{(r-1)E_{XX}}\right]$$

Un estimador de la precisión relativa es la razón del cuadrado medio del error sin ajustar por el cuadrado medio del error ajustado, tanto para tratamientos como para bloques, multiplicado por 100 para expresarlo en porcentajes.

Para los tratamientos

$$\frac{\frac{E_{YY}}{tr-t-r+1}}{S^2_{Y.X(Tra)}}100$$

Para los bloques

$$\frac{\frac{E_{YY}}{tr-t-r+1}}{S^2_{Y.X(Blo)}}100$$

Ejemplo.-

Se estudian los efectos de tres *dietas* en el *incremento de peso* en cerdos. Como el *peso inicial* influye también en la aptitud al engorde, es por lo que se mide, también, esta variable como covariable. El diseño es en bloque aleatorios pues la experiencia se realiza situando las unidades experimentales en 10 jaulas separadas. Siendo los datos los siguientes

Jaula	*Dietas*

	A		B		C		Global	
	X	Y	X	Y	X	Y	ΣX	ΣY
1	14.8	4.22	15.8	3.78	19.8	4.07	50.40	12.07
2	18.9	3.84	19.9	4.65	19.9	4.02	58.70	12.51
3	19.2	4.43	21.2	4.04	19.2	4.28	59.60	12.75
4	19.7	4.46	15.7	3.71	16.7	4.02	52.10	12.19
5	21.4	4.74	19.4	4.20	19.4	4.02	60.20	12.96
6	19.8	4.41	19.8	4.46	19.8	4.36	59.40	13.23
7	17.9	4.42	17.9	4.33	15.9	4.09	51.70	12.84
8	16.2	4.43	19.2	4.46	15.2	3.70	50.60	12.59
9	18.7	4.61	18.7	4.08	20.7	4.41	58.10	13.10
10	16.4	4.01	18.4	4.42	15.4	3.94	50.20	12.37
Σ	183.0	43.57	186.0	42.13	182.0	40.91	551.00	126.61
Σ^2	3384.48	190.48	3487.28	178.35	3353.48	167.76	10225.24	536.59
ΣXY	799.559		786.705		747.444		2333.708	

Calculemos primero las sumas de cuadrados y las sumas de productos

El error

$$EXX = 10225.24 - \frac{183^2+186^2+182^2}{10} - \frac{50.4^2+58.7^2+59.6^2+52.1^2+60.2^2+59.4^2+51.7^2+50.6^2+58.1^2+50.2^2}{3} + \frac{551^2}{30} = 46.4667$$

$$EYY = 536.59 - \frac{43.57^2+42.13^2+40.91^2}{10} - \frac{\left(12.07^2+12.51^2+12.74^2+12.19^2+12.96^2+13.23^2+12.84^2+12.59^2+13.10^2+12.37^2\right)}{3} + \frac{126.61^2}{30} = 1.4554$$

$$EXY = 2333.708 - \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} - \frac{\left(\begin{array}{l}50.4\times 12.07+58.7\times 12.51+59.6\times 12.75+52.1\times 12.19+60.2\times 12.96+\\+59.4\times 13.23+51.7\times 12.84+50.6\times 12.59+58.1\times 13.1+50.2\times 12.37\end{array}\right)}{3} + \frac{551\times 126.61}{30} = 4.706$$

$$gl = (3-1)\times(10-1) = 18$$

Las dietas

$$T_{XX} = \frac{183^2+186^2+182^2}{10} - \frac{551^2}{30} = 0.8667$$

$$T_{YY} = \frac{43.57^2+42.13^2+40.91^2}{10} - \frac{126.61^2}{30} = 0.3546$$

$$T_{XY} = \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} - \frac{551\times 126.61}{30} = 0.1073$$

$$gl = (3-1) = 2$$

Las jaulas

$$R_{XX} = \frac{50.4^2+58.7^2+59.6^2+52.1^2+60.2^2+59.4^2+51.7^2+50.6^2+58.1^2+50.2^2}{3} - \frac{551^2}{30} =$$

$$= 57.8733$$

$$R_{YY} = \frac{\left(\begin{array}{l}12.07^2+12.51^2+12.74^2+12.19^2+12.96^2+\\ +13.23^2+12.84^2+12.59^2+13.10^2+12.37^2\end{array}\right)}{3} - \frac{126.61^2}{30} = 0.4436$$

$$R_{XY} = \frac{\left(\begin{array}{l}50.4\times 12.07+58.7\times 12.51+59.6\times 12.75+52.1\times 12.19+60.2\times 12.96+\\ +59.4\times 13.23+51.7\times 12.84+50.6\times 12.59+58.1\times 13.1+50.2\times 12.37\end{array}\right)}{3} -$$

$$- \frac{551\times 126.61}{30} = 3.491$$

$$gl = (10-1) = 9$$

La suma de las sumas de cuadrados y suma de productos del error más la *dieta* es

$$S_{XX} = E_{XX}+T_{XX} = 46.4667+0.8667 = 47.3334$$
$$S_{YY} = E_{YY}+T_{YY} = 1.4554+0.3546 = 1.81$$
$$S_{XY} = E_{XY}+T_{XY} = 4.706+0.1073 = 4.8133$$
$$gl = 18+2 = 20$$

La suma de las sumas de cuadrados y suma de productos del error más la *jaula* es

$$U_{XX} = E_{XX}+R_{XX} = 46.4667+57.8733 = 104.34$$
$$U_{YY} = E_{YY}+R_{YY} = 1.4554+0.4436 = 1.899$$
$$U_{XY} = E_{XY}+R_{XY} = 4.706+3.491 = 8.197$$
$$gl = 18+9 = 27$$

Los totales son

$$\text{total}_{XX} = 10225.24 - \frac{551^2}{30} = 105.2067$$

$$\text{total}_{YY} = 536.59 - \frac{126.61^2}{30} = 2.2536$$

$$\text{total}_{XY} = 2333.708 - \frac{551 \times 126.61}{30} = 8.3043$$

$$gl = 10 \times 3 - 1 = 29$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con *X*.

La estima de β del modelo del análisis de covarianza se realiza, como ya se ha visto antes, utilizando los residuos

$$b = \frac{4.706}{46.4667} = 0.1013$$

Por lo que la contribución a la suma de cuadrados atribuible a la regresión ajustada para *dietas* y *jaulas* es

$$SC_{\text{Regresión}} = \frac{4.706^2}{46.4667} = 0.4767$$

La suma de cuadrados del error ajustada para la regresión es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = 1.4554 - 0.4767 = 0.9787$$

La regresión de *Y* sobre *X*, sin eliminar el efecto de las *dietas* y dejando los efectos de *jaulas* es

$$b_S = \frac{4.8133}{47.3334} = 0.1017$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{4.8133^2}{47.3334} = 0.4895$$

La suma de cuadrados, ajustada para la regresión, debida a la *dieta* más el error es

$$SC_{\text{Trata+Error}} = S' = 1.81 - 0.4895 = 1.3205$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$S' - E' = 1.3205 - 0.9787 = 0.3418$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a las *dietas*.

Si la regresión de Y sobre X, sin eliminar el efecto de las *jaulas* y dejando los efectos de las *dietas* es

$$b_U = \frac{8.197}{104.34} = 0.0786$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{8.197^2}{104.34} = 0.6440$$

La suma de cuadrados, ajustada para la regresión, debida a las *jaulas* más el error es

$$SC_{\text{Bloques+Error}} = U' = 1.899 - 0.644 = 1.255$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$U' - E' = 1.255 - 0.9787 = 0.2763$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a las *jaulas*.

Los resultados se pueden resumir en la siguiente tabla

					reducción debida a b		*ajuste debido a b*		
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*	*CM*
Dietas	2	0.8666	0.1073	0.3546					
Jaulas	9	57.8733	3.491	0.4436					
Error	18	46.4667	4.706	1.4554	1	0.4767	17	0.9787	0.0576
Die+Err	20	47.3334	4.8133	1.81	1	0.4895	19	1.3205	
Jau+Err	27	104.34	8.197	1.899	1	0.644	26	1.255	
total	29	105.207	8.3043	2.2536					
Dietas ajustados							2	0.3418	0.1709
Jaulas ajustados							9	0.2763	0.0307

Para probar el supuesto de que las fuentes de variación no influyen en la covariable, probemos primero las dietas, la F_o de esta ANOVA es

$$F_o = \frac{\frac{0.8667}{2}}{\frac{46.4667}{18}} = 0.168ns$$

$$F_{(2,18;\,0.05)} = 3.55$$

como es de esperar, la aleatorización ha sido correcta y no existe el efecto dieta en el peso inicial.

Y para las jaulas, la F_o sería

$$F_o = \frac{\frac{57.8333}{9}}{\frac{46.4667}{18}} = 2.489\,*$$

$$F_{(9,18;\,0.05)} = 2.46$$

aquí si se ha fallado la aleatorización y las jaulas son significativas en el peso inicial.

Si se quiere probar la hipótesis de igualdad de efectos de las diferentes dietas para la variable Y sin ajustar para la regresión, es decir, como si, desconociendo la variable X, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{0.3546}{2}}{\frac{1.4554}{18}} = 2.193ns$$

$$F_{(2,18;\,0.05)} = 3.55$$

la conclusión sería que, ignorando el peso inicial, las dietas no influyen significativamente en el incremento de peso.

Si se quieren probar las jaulas en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{0.4436}{9}}{\frac{1.4554}{18}} = 0.61ns$$

$$F_{(9,18;\,0.05)} = 2.46$$

las jaulas no han influido en el incremento de peso, ignorando el peso inicial.

Si la prueba que se desea hacer es la de igualdad de efectos de las dietas pero ajustada para la regresión con la covariable, es decir, la prueba F de las medias ajustadas, la F_o sería

$$F_o = \frac{\frac{0.3418}{2}}{\frac{0.9787}{17}} = 2.968ns$$

$$F_{(2,17;\,0.05)} = 3.59$$

la F ha pasado de 2.19 a 2.96, un incremento insuficiente como para que sean significativas las dietas. Teniendo en cuenta que, en la siguiente, la F de las jaulas es menor de 1 y que la jaulas dieron significativas para el peso inicial, esto esta restando precisión al experimento, lo correcto en este caso es realizar el modelo sin el factor jaula, en cuyo caso tendríamos los resultados del primer ejemplo de este capítulo, en el que la F de los efectos de las dietas ajustadas para la regresión si da significativa.

Si se quieren probar las jaulas ajustados, la pruebas sería

$$F_o = \frac{\frac{0.2763}{9}}{\frac{0.9787}{17}} = 0.533ns$$

$$F_{(9,17;\,0.05)} = 2.49$$

esta prueba es no significativa, pero una F inferior a 1 hace que se pierda precisión (ver el epígrafe *Estima conjunta (englobe) de la significación de varios factores en modelos compuestos* del Capítulo 9), por lo que lo recomendable es eliminarla del modelo, para que pase al error.

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, la prueba F sería

$$F_o = \frac{\frac{4.706^2}{46.4667}}{\frac{0.9787}{18}} = 8.276^{**}$$

$$F_{(1,18;\,0.01)} = 8.28$$

efectivamente existe una regresión significativa del peso inicial en el peso final, después de haber corregido para los efectos de los dos factores.

El cálculo de una media ajustada (o media minimocuadrática) de las dietas son

$$\hat{Y}_{D_1} = 4.357 - 0.1013 \times (18.6 - 18.37) = 4.333$$

$$\hat{Y}_{D_2} = 4.213 - 0.1013 \times (18.6 - 18.37) = 4.189$$

$$\hat{Y}_{D_3} = 4.091 - 0.1013 \times (18.2 - 18.37) = 4.108$$

Y de las jaulas

$$\hat{Y}_{J_1} = 4.02 - 0.1013 \times (16.80 - 18.37) = 4.179$$
$$\hat{Y}_{J_2} = 4.17 - 0.1013 \times (19.57 - 18.37) = 4.048$$
$$\hat{Y}_{J_3} = 4.25 - 0.1013 \times (19.87 - 18.37) = 4.098$$
$$\hat{Y}_{J_4} = 4.06 - 0.1013 \times (17.37 - 18.37) = 4.161$$
$$\hat{Y}_{J_5} = 4.32 - 0.1013 \times (20.07 - 18.37) = 4.148$$
$$\hat{Y}_{J_6} = 4.41 - 0.1013 \times (19.80 - 18.37) = 4.265$$
$$\hat{Y}_{J_7} = 4.28 - 0.1013 \times (17.23 - 18.37) = 4.395$$
$$\hat{Y}_{J_8} = 4.20 - 0.1013 \times (16.87 - 18.37) = 4.352$$
$$\hat{Y}_{J_9} = 4.37 - 0.1013 \times (19.37 - 18.37) = 4.269$$
$$\hat{Y}_{J_{10}} = 4.12 - 0.1013 \times (16.73 - 18.37) = 4.286$$

Los *errores típico de las medias ajustadas* de la primera dieta y de la primera jaula, por ejemplo, son, respectivamente

Primera dieta

$$S_{\hat{Y}_{D1}} = \sqrt{0.0576 \left[\frac{1}{10} + \frac{(18.6 - 18.37)^2}{46.4667} \right]} = 0.0763$$

Primera jaula

$$S_{\hat{Y}_{J1}} = \sqrt{0.0576 \left[\frac{1}{3} + \frac{(16.8 - 18.37)^2}{46.4667} \right]} = 0.1492$$

La diferencia entre las medias ajustadas de la primera y segunda dieta, y la primera y segunda jaula es, respectivamente

Diferencia entre la 1° y 2° dieta

$$\hat{Y}_{D_1} - \hat{Y}_{D_2} = 4.357 - 4.213 - 0.1013 \times (18.3 - 18.6) = 0.1744$$

Diferencia entre la 1° y 2° jaula

$$\hat{Y}_{J_1} - \hat{Y}_{J_2} = 4.02 - 4.17 - 0.1013 \times (16.8 - 19.57) = 0.1306$$

El *error típico de la diferencia de las medias ajustadas* de la primera y segunda dieta, y de la primera y segunda jaula es, respectivamente

$$S_{\hat{Y}_{D1}-\hat{Y}_{D2}} = \sqrt{0.0576\left[\frac{2}{10}+\frac{(18.3-18.6)^2}{46.4667}\right]} = 0.1078$$

$$S_{\hat{Y}_{J1}-\hat{Y}_{J2}} = \sqrt{0.0576\left[\frac{2}{3}+\frac{(16.8-19.57)^2}{46.4667}\right]} = 0.2189$$

La prueba *t* para contrastar las media ajustada de la primera y segunda dieta y de la primera y segunda jaula, por ejemplo, son

Primera y segunda dieta

$$t = \frac{0.1744}{0.1078} = 1.6178ns$$

$$t_{(17;\,0.05/2)} = 2.098$$

Primera y segunda jaula

$$t = \frac{0.1306}{0.2189} = 0.5966ns$$

$$t_{(17;\,0.05/2)} = 2.1098$$

Resumiendo, se tiene que si se hubiese analizado esta experiencia con un ANOVA, se hubiera concluido que las tres dietas son iguales, mientras que teniendo en cuenta el peso con que comenzó cada unidad experimental el experimento, es decir, analizado con el análisis de covarianza, se concluye que las dietas son diferentes. Aquella dieta que tenga una mayor media ajusta y la prueba *t* de diferente de la siguiente, será la que propicie un mayor incremento de peso. Esto es como consecuencia de la mayor precisión de la covarianza.

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de dietas antes y después del ajuste.

El cuadrado medio del error antes del ajuste es

$$\frac{1.4554}{18} = 0.0808$$

El cuadrado medio del error efectivo después del ajuste para *X* es para las dietas y las jaulas, respectivamente

Para los tratamientos

$$S^2_{Y.X(Tra)} = 0.0576\left[1+\frac{0.8667}{2\times 46.4667}\right] = 0.0581$$

Para los bloques

$$S^2_{Y.X(Blo)} = 0.0576\left[1+\frac{57.8733}{9\times 46.4667}\right] = 0.0656$$

Un estimador de la precisión relativa es la razón del cuadrado medio del error sin ajustar por el cuadrado medio del error ajustado, tanto para dietas como para jaulas, multiplicado por 100 para expresarlo en porcentajes.

Para los tratamientos

$$\frac{0.0808}{0.0581}100 = 139.1\%$$

Para los bloques

$$\frac{0.0808}{0.0656}100 = 123.2\%$$

En el epígrafe *Partición de la covarianza* hay otro ejemplo de modelo de dos vías sin repetición o del diseño de bloques aleatorios.

Archivo del programa *SAS* (C17-2.SAS).-

```
title 'Análisis de covarianza de dos factores sin repetición';
options ls=75 ps=60;
data ancova;
infile 'c17-2.dat';
input dieta $ jaula x y @@;
proc anova;
  class dieta jaula;
  model x = dieta jaula ;
proc anova;
  class dieta jaula;
  model y = dieta jaula ;
proc glm;
  class dieta jaula;
   model y=dieta jaula x / solution;
   lsmeans dieta jaula / stderr tdiff;
run;
```

Archivo de datos (C17-2.DAT).-

```
A  1 14.8  4.22  B  1 15.8  3.78  C  1 19.8  4.07
A  2 18.9  3.84  B  2 19.9  4.65  C  2 19.9  4.02
A  3 19.2  4.43  B  3 21.2  4.04  C  3 19.2  4.28
A  4 19.7  4.46  B  4 15.7  3.71  C  4 16.7  4.02
A  5 21.4  4.74  B  5 19.4  4.20  C  5 19.4  4.02
A  6 19.8  4.41  B  6 19.8  4.46  C  6 19.8  4.36
A  7 17.9  4.42  B  7 17.9  4.33  C  7 15.9  4.09
A  8 16.2  4.43  B  8 19.2  4.46  C  8 15.2  3.70
A  9 18.7  4.61  B  9 18.7  4.08  C  9 20.7  4.41
A 10 16.4  4.01  B 10 18.4  4.42  C 10 15.4  3.94
```

Archivo de resultados (C17-2.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: X

                                  Sum of          Mean
Source                    DF        Squares        Square  F Value     Pr > F
Model                     11     58.7400000     5.3400000     2.07     0.0825
Error                     18     46.4666667     2.5814815
Corrected Total           29    105.2066667

                  R-Square          C.V.       Root MSE              X Mean
                  0.558330      8.747907        1.60670             18.3667

Source                    DF       Anova SS   Mean Square  F Value     Pr > F
DIETA                      2      0.8666667     0.4333333     0.17     0.8468
JAULA                      9     57.8733333     6.4303704     2.49     0.0475

                    Analysis of Variance Procedure
Dependent Variable: Y

                                  Sum of          Mean
Source                    DF        Squares        Square  F Value     Pr > F
Model                     11     0.79775000     0.07252273     0.89     0.5633
Error                     18     1.46094667     0.08116370
Corrected Total           29     2.25869667

                  R-Square          C.V.       Root MSE              Y Mean
                  0.353190      6.750473        0.28489             4.22033

Source                    DF       Anova SS   Mean Square  F Value     Pr > F
DIETA                      2     0.35458667     0.17729333     2.18     0.1415
JAULA                      9     0.44316333     0.04924037     0.61     0.7758

                  General Linear Models Procedure
Dependent Variable: Y

                                  Sum of          Mean
Source                    DF        Squares        Square  F Value     Pr > F
Model                     12     1.27435910     0.10619659     1.83     0.1230
Error                     17     0.98433757     0.05790221
Corrected Total           29     2.25869667

                  R-Square          C.V.       Root MSE              Y Mean
                  0.564201      5.701653        0.24063             4.22033

Source                    DF     Type I SS   Mean Square  F Value     Pr > F
DIETA                      2     0.35458667     0.17729333     3.06     0.0732
JAULA                      9     0.44316333     0.04924037     0.85     0.5830
X                          1     0.47660910     0.47660910     8.23     0.0106

Source                    DF    Type III SS   Mean Square  F Value     Pr > F
DIETA                      2     0.34172722     0.17086361     2.95     0.0794
JAULA                      9     0.27581221     0.03064580     0.53     0.8337
X                          1     0.47660910     0.47660910     8.23     0.0106

                                    T for H0:     Pr > |T|    Std Error of
Parameter              Estimate     Parameter=0                  Estimate
INTERCEPT        2.316179340 B            3.83        0.0013     0.60428399
DIETA       A    0.255872310 B            2.38        0.0295     0.10767034
            B    0.081489240 B            0.75        0.4630     0.10853487
            C    0.000000000 B               .             .              .
JAULA       1   -0.106751793 B           -0.54        0.5940     0.19648667
            2   -0.240284553 B           -1.09        0.2910     0.22046520
            3   -0.190667623 B           -0.85        0.4095     0.22546713
            4   -0.124142037 B           -0.63        0.5385     0.19774048
            5   -0.140923003 B           -0.62        0.5465     0.22901321
            6   -0.023915830 B           -0.11        0.9163     0.22432203
            7    0.106028216 B            0.54        0.5979     0.19726378
            8    0.059829747 B            0.30        0.7645     0.19652895
```

```
          9     -0.023362506 B          -0.11       0.9157       0.21735340
          10     0.000000000 B            .           .            .
X                0.101276901            2.87       0.0106       0.03530017
NOTE: The X'X matrix has been found to be singular and a generalized
      inverse was used to solve the normal equations.   Estimates followed
      by the letter 'B' are biased, and are not unique estimators of the
      parameters.

                                  Least Squares Means
          DIETA                     Y       Std Err      Pr > |T|    LSMEAN
                                 LSMEAN        LSMEAN    H0:LSMEAN=0    Number
          A                 4.36375179    0.07612988          0.0001      1
          B                 4.18936872    0.07653799          0.0001      2
          C                 4.10787948    0.07632061          0.0001      3

                 T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|
                      i/j          1           2           3
                      1            .     1.612683    2.376442
                                         0.1252      0.0295
                      2     -1.61268           .    0.750812
                                0.1252                0.4630
                      3     -2.37644    -0.75081           .
                                0.0295      0.4630
NOTE: To ensure overall protection level, only probabilities associated
      with pre-planned comparisons should be used.

          JAULA                     Y       Std Err      Pr > |T|    LSMEAN
                                 LSMEAN        LSMEAN    H0:LSMEAN=0    Number
          1                 4.18200048    0.14953001          0.0001      1
          2                 4.04846772    0.14524160          0.0001      2
          3                 4.09808465    0.14867571          0.0001      3
          4                 4.16461023    0.14334169          0.0001      4
          5                 4.14782927    0.15133397          0.0001      5
          6                 4.26483644    0.14785393          0.0001      6
          7                 4.39478049    0.14457277          0.0001      7
          8                 4.34858202    0.14867571          0.0001      8
          9                 4.26538977    0.14334169          0.0001      9
          10                4.28875227    0.15041629          0.0001     10

                 T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|
   i/j          1           2           3           4           5           6           7
   1            .    0.608606    0.374086    0.088057    0.149997    -0.37114    -1.07973
                       0.5508      0.7130      0.9309      0.8825      0.7151      0.2953
   2    -0.60861           .    -0.25217    -0.54975     -0.5037     -1.1003    -1.62558
          0.5508                  0.8039      0.5896      0.6209      0.2865      0.1224
   3    -0.37409    0.252173           .    -0.30887    -0.25303    -0.84867    -1.36504
          0.7130      0.8039                  0.7612      0.8033      0.4079      0.1900
   4    -0.08806     0.54975    0.308872           .    0.076846    -0.46741    -1.17118
          0.9309      0.5896      0.7612                  0.9396      0.6461      0.2577
   5      -0.15    0.503699    0.253025    -0.07685           .    -0.59486    -1.12014
          0.8825      0.6209      0.8033      0.9396                  0.5598      0.2782
   6    0.371136      1.1003    0.848667     0.46741    0.594857           .     -0.6006
          0.7151      0.2865      0.4079      0.6461      0.5598                  0.5560
   7    1.079734     1.62558    1.365039    1.171177    1.120137    0.600599           .
          0.2953      0.1224      0.1900      0.2577      0.2782      0.5560
   8    0.847801    1.374337    1.122321    0.932618    0.885814    0.377081    -0.23463
          0.4083      0.1872      0.2773      0.3641      0.3881      0.7108      0.8173
   9    0.385424    1.103371    0.848129    0.482725    0.593679    0.002808    -0.61494
          0.7047      0.2852      0.4081      0.6354      0.5605      0.9978      0.5467
  10    0.543303    1.089898    0.845656    0.627803    0.615349    0.106614    -0.53749
          0.5940      0.2910      0.4095      0.5385      0.5465      0.9163      0.5979

   i/j          8           9          10
   1      -0.8478    -0.38542     -0.5433
            0.4083      0.7047      0.5940
   2    -1.37434    -1.10337     -1.0899
          0.1872      0.2852      0.2910
   3    -1.12232    -0.84813    -0.84566
          0.2773      0.4081      0.4095
```

```
   4   -0.93262   -0.48272    -0.6278
         0.3641     0.6354     0.5385
   5   -0.88581   -0.59368   -0.61535
         0.3881     0.5605     0.5465
   6   -0.37708   -0.00281   -0.10661
         0.7108     0.9978     0.9163
   7   0.234631   0.614944   0.537495
         0.8173     0.5467     0.5979
   8          .   0.386254   0.304432
                    0.7041     0.7645
   9   -0.38625          .   -0.10749
         0.7041                0.9157
  10   -0.30443   0.107486          .
         0.7645     0.9157
NOTE: To ensure overall protection level, only probabilities associated
      with pre-planned comparisons should be used.
```

Estos son los mismos resultados que los obtenidos anualmente, pero con un menor error de redondeo.

La suma de cuadrados que hay que mirar es la **tipo III**. La suma de cuadrados **tipo I** debida a los factores (dieta y jaula) es la del ANOVA para la variable *Y*. Si, en el modelo del programa SAS, se pone primero la covariable y después los factores, esto es, **model y = x dieta jaula/ solution;** se hubiera obtenido la misma suma de cuadrados debida al último factor del modelo (en este caso *jaula*), en el tipo I y en el tipo III, pero lo suma de cuadrados tipo I debida a la regresión, sería la de la regresión sin considerar las correcciones del factor, esto es, la de la regresión de los treinta pares de valores, como si las dietas no existieran, y la suma de cuadrados tipo I del primer factor (dieta) sería la suma de cuadrados del ANCOVA para este solo factor, como si el siguiente factor no existiera.

Análisis de covarianza de un modelo factorial con repetición.-

No existe ninguna novedad con respecto a lo estudiado en los modelos anteriores. La descomposición de la variación total y de los grados de libertad totales, así como los parámetros estimados y las pruebas de hipótesis, se realiza tal como se estudió en el Capítulo 7 para el *Análisis de varianza Factorial con repetición*, con la única diferencia de que, al haber dos variables, habrá que realizar tres descomposiciones dos para cada una de las sumas de cuadrados, la de *X* y la de *Y*, y una para la suma de productos de ambas variables. Como consecuencia de ello, las anotaciones del capítulo 7 se hacen muy largas, por lo que se va a utilizar el tipo de anotación propia de éste capítulo.

La regresión se calcula en la fila del error y la suma de cuadrados ajustadas para los dos factores y la interacción, si la hubiera, se calcularan restando de la suma de cuadrados ajustada de dichas fuentes de variación más el error, la suma de cuadrados ajustada del error.

Al primer factor vamos a seguir simbolizandolo con la *T* y al segundo con la *R*, anotación que procede de los ejemplos utilizados en los capítulo 6 y sucesivos.

El modelo lineal que se aplicará es

$$Y_{ijk} = \mu + T_i + R_j + TR_{ij} + \beta(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

Al haber repeticiones se puede estimar la interacción, por lo que se incluye esta en el modelo. Recuérdese que en el anterior modelo no se introducía la interacción por no poderse estimar, al haber una sola medida por casilla (ver capítulo 5).

La descomposición de la variación total y de los grados de libertad total es

$$\begin{aligned} E_{XX} &= \Sigma_{ijk} X_{ijk}^2 - \Sigma_{ij} \frac{X_{ij.}^2}{n} \\ E_{YY} &= \Sigma_{ijk} Y_{ijk}^2 - \Sigma_{ij} \frac{Y_{ij.}^2}{n} \\ E_{XY} &= \Sigma_{ijk} X_{ijk} Y_{ijk} - \Sigma_{ij} \frac{X_{ij.} Y_{ij.}}{n} \\ gl &= N - tr \end{aligned}$$

$$\begin{aligned} T_{XX} &= \Sigma_i \frac{X_{i..}^2}{rn} - \frac{X_{...}^2}{N} \\ T_{YY} &= \Sigma_i \frac{Y_{i..}^2}{rn} - \frac{Y_{...}^2}{N} \\ T_{XY} &= \Sigma_i \frac{X_{i..} Y_{i..}}{rn} - \frac{X_{...} Y_{...}}{N} \\ gl &= t - 1 \end{aligned}$$

$$\begin{aligned} R_{XX} &= \Sigma_j \frac{X_{.j.}^2}{tn} - \frac{X_{...}^2}{N} \\ R_{YY} &= \Sigma_j \frac{Y_{.j.}^2}{tn} - \frac{Y_{...}^2}{N} \\ R_{XY} &= \Sigma_j \frac{X_{.j.} Y_{.j.}}{tn} - \frac{X_{...} Y_{...}}{N} \\ gl &= r - 1 \end{aligned}$$

$$\begin{aligned} TR_{XX} &= \Sigma_{ij} \frac{X_{ij.}^2}{n} - \Sigma_i \frac{X_{i..}^2}{rn} - \Sigma_j \frac{X_{.j.}^2}{tn} + \frac{X_{...}^2}{N} \\ TR_{YY} &= \Sigma_{ij} \frac{Y_{ij.}^2}{n} - \Sigma_i \frac{Y_{i..}^2}{rn} - \Sigma_j \frac{Y_{.j.}^2}{tn} + \frac{Y_{...}^2}{N} \\ TR_{XY} &= \Sigma_{ij} \frac{X_{ij.} Y_{ij.}}{n} - \Sigma_i \frac{X_{i..} Y_{i..}}{rn} - \Sigma_j \frac{X_{.j.} Y_{.j.}}{tn} + \frac{X_{...} Y_{...}}{N} \\ gl &= (t-1)(r-1) \end{aligned}$$

La suma de cuadrados y la suma de productos del factor T más el error (S) es

$$S_{XX} = E_{XX} + T_{XX} = \sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n} + \sum_i \frac{X_{i..}^2}{rn} - \frac{X_{...}^2}{N}$$

$$S_{YY} = E_{YY} + T_{YY} = \sum_{ijk} Y_{ijk}^2 - \sum_{ij} \frac{Y_{ij.}^2}{n} + \sum_i \frac{Y_{i..}^2}{rn} - \frac{Y_{...}^2}{N}$$

$$S_{XY} = E_{XY} + T_{XY} = \sum_{ijk} X_{ijk} Y_{ijk} - \sum_{ij} \frac{X_{ij.} Y_{ij.}}{n} + \sum_i \frac{X_{i..} Y_{i..}}{rn} - \frac{X_{...} Y_{...}}{N}$$

$$gl = tr - r$$

La suma de cuadrados y la suma de productos del factor R más el error (U) es

$$U_{XX} = E_{XX} + R_{XX} = \sum_{ijk} X_{ijk}^2 - \sum_{ij} \frac{X_{ij.}^2}{n} + \sum_i \frac{X_{.j.}^2}{tn} - \frac{X_{...}^2}{N}$$

$$U_{YY} = E_{YY} + R_{YY} = \sum_{ijk} Y_{ijk}^2 - \sum_{ij} \frac{Y_{ij.}^2}{n} + \sum_i \frac{Y_{.j.}^2}{tn} - \frac{Y_{...}^2}{N}$$

$$U_{XY} = E_{XY} + R_{XY} = \sum_{ijk} X_{ijk} Y_{ijk} - \sum_{ij} \frac{X_{ij.} Y_{ij.}}{n} + \sum_i \frac{X_{.j.} Y_{.j.}}{tn} - \frac{X_{...} Y_{...}}{N}$$

$$gl = tr - t$$

La suma de cuadrados y la suma de productos de la *interacción* más el error (V) es

$$V_{XX} = E_{XX} + TR_{XX} = \sum_{ijk} X_{ijk}^2 - \sum_i \frac{X_{i..}^2}{rn} - \sum_j \frac{X_{.j.}^2}{tn} + \frac{X_{...}^2}{N}$$

$$V_{YY} = E_{YY} + TR_{YY} = \sum_{ijk} Y_{ijk}^2 - \sum_i \frac{Y_{i..}^2}{rn} - \sum_j \frac{Y_{.j.}^2}{tn} + \frac{Y_{...}^2}{N}$$

$$V_{XY} = E_{XY} + TR_{XY} = \sum_{ijk} X_{ijk} Y_{ijk} - \sum_i \frac{X_{i..} Y_{i..}}{rn} - \sum_j \frac{X_{.j.} Y_{.j.}}{tn} + \frac{X_{...} Y_{...}}{N}$$

$$gl = N - t - r + 1$$

Como se puede ver, ni S ni U ni V coincide con los totales que son

$$\text{total}_{XX} = \sum_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$$

$$\text{total}_{YY} = \sum_{ijk} Y_{ijk}^2 - \frac{Y_{...}^2}{N}$$

$$\text{total}_{XY} = \sum_{ijk} X_{ijk} Y_{ijk} - \frac{X_{...} Y_{...}}{N}$$

$$gl = trn - 1 = N - 1$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con X.

Ahora, y tal como se hizo en los modelos anteriores, mediante un proceso secuencial, se le va a extraer a cada suma de cuadrados la componente debida a la regresión quedando un residuo más pequeño para el análisis de covarianza.

La estima de β del modelo del análisis de covarianza es, como ya se ha visto antes

$$b = \frac{E_{XY}}{E_{XX}}$$

es decir, que la estima de b en el análisis de covarianza se realiza como en el capítulo 11, pero en lugar de usar, en el numerador, la SP total, se utiliza la SP que queda después de restarle a la SP total la SP debida a las tres fuentes de variación, es decir, la SP del error, E_{XY} según la anotación de este capítulo. Y en lugar de dividir por la SC total de la variable X se divide por la SC que queda después de restarle a la SC total de la variable X la SC debida a las tres fuentes de variación, es decir, la SC del error, E_{XX} según la anotación de este capítulo.

Por lo que la contribución a la suma de cuadrados total, atribuible a la regresión corregida para tratamientos y bloques es

$$SC_{\text{Regresión}} = b\,E_{XY} = \frac{E_{XY}^2}{E_{XX}}$$

La suma de cuadrados del error de la variable bajo estudio (la Y) ajustada para la regresión con la covariable (la X) es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = E_{YY} - \frac{E_{XY}^2}{E_{XX}}$$

La regresión de Y sobre X, corrigiendo los valores de Y para los efectos de los diferentes niveles del factor T e ignorando los efectos del factor R y de la *interacción,* es

$$b_S = \frac{S_{XY}}{S_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_S\,S_{XY} = \frac{S_{XY}^2}{S_{XX}}$$

La fracción de la suma de cuadrados total, ajustada para la regresión, debida al

factor *T* más el error es

$$SC_{T+Error} = S' = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (*S*-*E*) es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible al factor *T*.

La regresión de *Y* sobre *X*, corrigiendo los valores de *Y* para los efectos de los diferentes niveles del factor *R* e ignorando los efectos del factor *T* y de la *interacción* es

$$b_U = \frac{U_{XY}}{U_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{Regresión} = b_U \; U_{XY} = \frac{U_{XY}^2}{U_{XX}}$$

La fracción de la suma de cuadrados, ajustada para la regresión, debida al factor *R* más el error es

$$SC_{R+Error} = U' = U_{YY} - \frac{U_{XY}^2}{U_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (*U*-*E*) es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible al factor *R*.

La regresión de *Y* sobre *X*, corrigiendo los valores de *Y* para los efectos de la *interacción* e ignorando los efectos de los factores *R* y *T*, es

$$b_V = \frac{V_{XY}}{V_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{Regresión} = b_V \, V_{XY} = \frac{V_{XY}^2}{V_{XX}}$$

La fracción de la suma de cuadrados, ajustada para la regresión, debida a la *interacción* más el error es

$$SC_{T\times R+Error} = V' = V_{YY} - \frac{V_{XY}^2}{V_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (V'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a la *interacción*.

El análisis se puede resumir en la siguiente tabla

					reducción debida a ***b***		*ajuste debido a* ***b***	
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*
T	t-1	T_{XX}	T_{XY}	T_{YY}				
R	r-1	R_{XX}	R_{XY}	R_{YY}				
TR	(t-1)(r-1)	TR_{XX}	TR_{XY}	TR_{YY}				
Error	N-tr	E_{XX}	E_{XY}	E_{YY}	1	$\frac{E_{XY}^2}{E_{XY}}$	N-tr-1	E'
T+Err	tr-r	S_{XX}	S_{XY}	S_{YY}	1	$\frac{S_{XY}^2}{S_{XY}}$	tr-r-1	S'
R+Err	tr-t	U_{XX}	U_{XY}	U_{YY}	1	$\frac{U_{XY}^2}{U_{XY}}$	tr-t-1	U'
TR+Err	N-t-r+1	V_{XX}	V_{XX}	V_{XX}	1	$\frac{V_{XY}^2}{V_{XY}}$	N-t-r	V'
total	N-1	*SC*	*SP*	*SC*				
T ajustado							t-1	S'-E'
R ajustado							r-1	U'-E'
Interacción ajustada							N-tr	V'-E'

Ajustes de las medias.-

Las fórmulas para el ajuste de las medias son las mismas de las dadas para los anteriores modelos. El cálculo de una media ajustada (o media minimocuadrática) de los diferentes niveles de los factores principales y de las diferentes casillas (interacción) son, respectivamente

Medias ajustadas del factor T

$$\hat{Y}_{i..} = \overline{Y}_{i..} - b(\overline{X}_{i..} - \overline{X}_{...})$$

Medias ajustadas del factor R

$$\hat{Y}_{.j.} = \overline{Y}_{.j.} - b(\overline{X}_{.j.} - \overline{X}_{...})$$

Medias ajustadas de la interacción

$$\hat{Y}_{ij.} = \overline{Y}_{ij.} - b(\overline{X}_{ij.} - \overline{X}_{...})$$

donde b es el coeficiente de regresión del error.

Los *errores típico de las medias ajustadas* de los diferentes niveles de los factores principales o de las diferentes casillas (interacción) son, respectivamente

Factor T

$$S_{\hat{Y}_{i..}} = S_{Y.X}\sqrt{\frac{1}{rn} + \frac{(\overline{X}_{i..} - \overline{X}_{...})^2}{E_{XX}}}$$

Factor R

$$S_{\hat{Y}_{.j.}} = S_{Y.X}\sqrt{\frac{1}{tn} + \frac{(\overline{X}_{.j.} - \overline{X}_{...})^2}{E_{XX}}}$$

Interacción

$$S_{\hat{Y}_{ij.}} = S_{Y.X}\sqrt{\frac{1}{n} + \frac{(\overline{X}_{ij.} - \overline{X}_{...})^2}{E_{XX}}}$$

siendo $S_{Y.X}$ la raíz cuadrada del cuadrado medio del error ajustado.

La diferencia entre las medias ajustadas de dos niveles de los factores principales o de dos casillas (interacción) sería, respectivamente

Diferencia entre dos niveles de T

$$\hat{Y}_{i.} - \hat{Y}_{j.} = \overline{Y}_{i.} - \overline{Y}_{j.} - b(\overline{X}_{i.} - \overline{X}_{j.})$$

Diferencia entre dos niveles de R

$$\hat{Y}_{.i} - \hat{Y}_{.j} = \overline{Y}_{.i} - \overline{Y}_{.j} - b(\overline{X}_{.i} - \overline{X}_{.j})$$

Diferencia entre dos casillas

$$\hat{Y}_{ij.} - \hat{Y}_{kl.} = \overline{Y}_{ij.} - \overline{Y}_{kl.} - b(\overline{X}_{ij.} - \overline{X}_{kl.})$$

Y los *errores típicos de la diferencia de medias ajustadas* sería, respectivamente

$$S_{\hat{Y}_{i..}-\hat{Y}_{j..}} = \sqrt{S^2_{X.Y}\left[\frac{2}{rn}+\frac{(\overline{X}_{i..}-\overline{X}_{j..})^2}{Exx}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{t_i}}+\frac{1}{n_{t_j}}+\frac{(\overline{X}_{i..}-\overline{X}_{j..})^2}{Exx}\right]}$$

$$S_{\hat{Y}_{.i.}-\hat{Y}_{.j.}} = \sqrt{S^2_{X.Y}\left[\frac{2}{nt}+\frac{(\overline{X}_{.i.}-\overline{X}_{.j.})^2}{Exx}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{r_i}}+\frac{1}{n_{r_j}}+\frac{(\overline{X}_{.i.}-\overline{X}_{.j.})^2}{Exx}\right]}$$

$$S_{\hat{Y}_{ij.}-\hat{Y}_{kl.}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n}+\frac{(\overline{X}_{ij.}-\overline{X}_{kl.})^2}{Exx}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{ij}}+\frac{1}{n_{kl}}+\frac{(\overline{X}_{ij.}-\overline{X}_{kl.})^2}{Exx}\right]}$$

La expresión de la derecha es para el caso en que los tamaños de submuestras sean diferentes.

Con estas expresiones se pueden hacer comparaciones múltiples de medias tal como se estudió en el Capítulo 9, teniendo en cuenta que hay que usar el cuadrado medio del error del análisis de covarianza y los grados de libertad de dicho error. Las pruebas más correctas en este caso son la *t* y la de *Scheffe*, si bien la más comúnmente utilizada es la prueba *t*.

Pruebas de hipótesis.-

Consúltese los epígrafes del capítulo 7, *Parámetros estimados con factores de efectos fijos y con factores de efectos aleatorios* y *Otras pruebas de significación* del *factorial con repetición* en los que se explica que el término de contraste para la prueba de los factores principales puede ser el cuadrado medio del *error* o de la *interacción*. Supongamos, para las explicaciones que siguen, que este es un modelo *fijo*, en ese caso todas las posibles pruebas de hipótesis son

Si se quiere probar el supuesto de que la covariable no esta influida por los efectos de los tratamientos sería realizar un ANOVA, que con los resultados de la tabla anterior, se realizarían con las siguientes F_o. Para el factor *T*

$$F_o = \frac{\frac{Txx}{t-1}}{\frac{Exx}{N-tr}}$$

que se contrastaría con la $F_{(t-1,N-tr;\,\alpha)}$

Para el factor *R* la F_o sería

$$F_o = \frac{\frac{Rxx}{r-1}}{\frac{Exx}{N-tr}}$$

que se contrastaría con la $F_{(r-1,N-tr;\,\alpha)}$

Y para la interacción, la F_o sería

$$F_o = \frac{\frac{TR_{XX}}{(t-1)(r-1)}}{\frac{E_{XX}}{N-tr}}$$

que se contrastaría con la $F_{((t-1)(r-1),N-tr;\,\alpha)}$

Si se quiere probar la hipótesis de igualdad de efectos de los diferentes tratamientos sin ajustar para la regresión, es decir, como si, desconociendo la variable X, se tratara de un ANOVA, está claro que, para el factor T la prueba F_o sería,

$$F_o = \frac{\frac{T_{YY}}{t-1}}{\frac{E_{YY}}{N-tr}}$$

que se contrastaría con la $F_{(t-1,N-tr;\,\alpha)}$

Si se quieren probar el factor R en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{R_{YY}}{r-1}}{\frac{E_{YY}}{N-tr}}$$

que se contrastaría con la $F_{(r-1,N-tr;\,\alpha)}$

Y para la interacción, la F_o sería

$$F_o = \frac{\frac{TR_{YY}}{(t-1)(r-1)}}{\frac{E_{YY}}{N-tr}}$$

que se contrastaría con la $F_{((t-1)(r-1),N-tr;\,\alpha)}$

Si la prueba que se desea hacer es la de igualdad de efectos de los tratamientos en la variable bajo estudio (la Y) pero ajustada para la regresión con la covariable, es decir, la prueba F de las medias ajustadas, a semejanza de la anterior, la F_o para el factor T sería

$$F_o = \frac{\frac{S'-E'}{t-1}}{\frac{E'}{N-tr-1}}$$

que se contrastaría con la $F_{(t-1,N-tr-1;\,\alpha)}$

Si se quieren probar el factor R ajustado, la pruebas sería

$$F_o = \frac{\frac{U'-E'}{r-1}}{\frac{E'}{N-tr-1}}$$

que se contrastaría con la $F_{(r-1,N-tr-1;\,\alpha)}$

Y para la interacción ajustada, la F_o sería

$$F_o = \frac{\frac{V'-E'}{(t-1)(r-1)}}{\frac{E'}{N-tr-1}}$$

que se contrastaría con la $F_{((t-1)(r-t),N-tr-1;\,\alpha)}$

Recuérdese que el término de contraste de los dos factores principales o de alguno de ellos, puede ser el cuadrado medio de la interacción y no el del error. Repásese los epígrafes del Capítulo 7 citados al comienzo del presente epígrafe.

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, a semejanza de la prueba *F* del Capítulo XI, la prueba *F* en este caso sería

$$F_o = \frac{\frac{E^2_{XY}}{E_{XX}}}{\frac{E'}{(N-tr-1)}}$$

que se contrastaría con la $F_{(1,N-tr-1;\,\alpha)}$

Aumento de precisión debido a la covarianza.-

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de las diversas fuentes de variación antes y después del ajuste.

El cuadrado medio del error antes del ajuste es

$$\frac{E_{YY}}{tr-t-r+1}$$

El cuadrado medio del error efectivo después del ajuste para *X* es, para las tres fuentes de variación, respectivamente

Para el factor *T*

$$S'^2_{Y.X(T)} = S^2_{Y.X}\left[1+\frac{T_{XX}}{(t-1)E_{XX}}\right]$$

Para el factor *R*

$$S'^2_{Y.X(R)} = S^2_{Y.X}\left[1+\frac{R_{XX}}{(r-1)E_{XX}}\right]$$

Para la interacción

$$S'^2_{Y.X(TR)} = S^2_{Y.X}\left[1+\frac{TR_{XX}}{(t-1)(r-1)E_{XX}}\right]$$

Un estimador de la precisión relativa es la razón del cuadrado medio del error sin ajustar por el cuadrado medio del error ajustado, para las tres fuentes de variación, multiplicado por 100 para expresarlo en porcentajes.

Para el factor T

$$\frac{\frac{E_{YY}}{tr-t-r+1}}{S^2_{Y.X(T)}}100$$

Para el factor R

$$\frac{\frac{E_{YY}}{tr-t-r+1}}{S^2_{Y.X(R)}}100$$

Para la interacción

$$\frac{\frac{E_{YY}}{tr-t-r+1}}{S^2_{Y.X(TR)}}100$$

Recuérdese que en el caso de un modelo mixto o aleatorio, el término de contraste puede ser la interacción, por lo que hay que utilizar esta como termino de error tanto en las pruebas de hipótesis como para calcular el aumento de precisión con la covarianza.

Ejemplo.-

Se ilustrará con el mismo ejemplo que se viene utilizando desde el principio de este capítulo.

En un estudio de la *ganancia de peso* (Y) de lechones se probaron tres *dietas* en los dos *razas*. Como el *peso al inicio* de la experiencia influye en la ganancia de peso, se tomó este peso inicial como covariable (X)

	Dietas					
	A		*B*		*C*	
	X	*Y*	*X*	*Y*	*X*	*Y*

Raza	R1	14.8	4.22	15.8	3.78	19.8	4.07
		18.9	3.84	19.9	4.65	19.9	4.02
		19.2	4.43	21.2	4.04	19.2	4.28
		19.7	4.46	15.7	3.71	16.7	4.02
		21.4	4.74	19.4	4.20	19.4	4.02
	R2	19.8	4.41	19.8	4.46	19.8	4.36
		17.9	4.42	17.9	4.33	15.9	4.09
		16.2	4.43	19.2	4.46	15.2	3.70
		18.7	4.61	18.7	4.08	20.7	4.41
		16.4	4.01	18.4	4.42	15.4	3.94

En la siguiente tabla se presentan los sumatorios y medias de estos datos.

	X	Y	X	Y	X	Y	X	Y
$\Sigma X \mid \Sigma Y$	94.0	21.69	92.0	20.38	95.0	20.41	281.0	62.48
$\Sigma X^2 \mid \Sigma Y^2$	1790.9	94.54	1717.9	83.64	1811.9	83.36	5320.7	261.54
ΣXY	409.39		377.63		387.88		1174.90	
$\bar{X} \mid \bar{Y}$	18.8	4.34	18.4	4.08	19.0	4.08	18.7	4.16
$\Sigma X \mid \Sigma Y$	89.0	21.88	94.0	21.75	87.0	20.50	270.0	64.13
$\Sigma X^2 \mid \Sigma Y^2$	1593.5	95.94	1769.3	94.71	1541.5	84.40	4904.3	275.05
ΣXY	390.17		409.07		359.56		1158.81	
$\bar{X} \mid \bar{Y}$	17.8	4.38	18.8	4.35	17.4	4.10	18.0	4.27
$\Sigma X \mid \Sigma Y$	183.0	43.57	186.0	42.13	182.0	40.91	551.0	126.61
$\Sigma X^2 \mid \Sigma Y^2$	3384.4	190.5	3487.2	178.4	3353.4	167.8	10225	536.59
ΣXY	799.56		786.7		747.44		2333.7	
$\bar{X} \mid \bar{Y}$	18.3	4.35	18.6	4.21	18.2	4.09	18.37	4.22

Primero se calculan las sumas de cuadrados y las sumas de productos, con sus grados de libertad. Para los factores principales, éstas son

El error

$$E_{XX} = 10225.24 - \frac{94^2+92^2+95^2+89^2+94^2+87^2}{5} = 95.0400$$

$$E_{YY} = 536.5951 - \frac{21.69^2+20.38^2+20.41^2+21.88^2+21.75^2+20.50^2}{5} = 1.7120$$

$$E_{XY} = 2333.7080 -$$

$$- \frac{94\times21.69+92\times20.38+95\times20.41+89\times21.88+94\times2.1.75+87\times20.5}{5} = 8.0901$$

$$gl = 30 - 3\times2 = 24$$

Las dietas

$$T_{XX} = \frac{183^2+186^2+182^2}{10} - \frac{551^2}{30} = 0.8643$$

$$T_{YY} = \frac{43.57^2+42.13^2+40.91^2}{10} - \frac{126.6^2}{30} = 0.3546$$

$$T_{XY} = \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} - \frac{551\times 126.6}{30} = 0.1072$$

$$gl = (3-1) = 2$$

Las razas

$$R_{XX} = \frac{281^2+270^2}{15} - \frac{551^2}{30} = 4.0332$$

$$R_{YY} = \frac{62.48^2+64.13^2}{15} - \frac{126.61^2}{30} = 0.0907$$

$$R_{XY} = \frac{281\times 6'248+270\times 6'413}{15} - \frac{551\times 126.61}{30} = -0.6052$$

$$gl = (2-1) = 1$$

Y la interacción

$$TR_{XX} = \frac{94^2+92^2+95^2+89^2+94^2+87^2}{5} - \frac{183^2+186^2+182^2}{10} -$$

$$- \frac{281^2+270^2}{15} + \frac{551^2}{30} = 5.2695$$

$$TR_{YY} = \frac{21.69^2+20.38^2+20.41^2+21.88^2+21.75^2+20.50^2}{5} -$$

$$- \frac{43.57^2+42.13^2+40.91^2}{10} - \frac{62.48^2+64.13^2}{15} + \frac{126.61^2}{30} = 0.1014$$

$$TR_{XY} = \frac{94\times 21.69+92\times 20.38+95\times 20.41+89\times 21.88+94\times 21.75+87\times 20.5}{5} -$$

$$- \frac{183\times 43.57+186\times 42.13+182\times 40.91}{10} - \frac{281\times 62.48+270\times 64.13}{15} +$$

$$+ \frac{551\times 126.61}{30} = 0.7126$$

$$gl = (3-1)\times(2-1) = 2$$

La suma de cuadrados y suma de productos del error más la *dieta* es

$$S_{XX} = E_{XX}+T_{XX} = 95.0400+0.8642 = 95.9042$$
$$S_{YY} = E_{YY}+T_{YY} = 1.7120+0.3546 = 2.0666$$
$$S_{XY} = E_{XY}+T_{XY} = 8.0901+0.1072 = 8.1973$$
$$gl = 24+2 = 26$$

La suma de cuadrados y suma de productos del error más la *raza* es

$$U_{XX} = E_{XX}+R_{XX} = 95.0400+4.0332 = 99.0732$$
$$U_{YY} = E_{YY}+R_{YY} = 1.7120+0.0907 = 1.8027$$
$$U_{XY} = E_{XY}+R_{XY} = 8.0901+0.6052 = 7.4849$$
$$gl = 24+1 = 25$$

La suma de cuadrados y suma de productos del error más la *interacción* es

$$V_{XX} = E_{XX}+TR_{XX} = 95.0400+5.2695 = 100.3096$$
$$V_{YY} = E_{YY}+TR_{YY} = 1.7120+0.1014 = 1.8134$$
$$V_{XY} = E_{XY}+TR_{XY} = 8.0901+0.7126 = 8.8027$$
$$gl = 24+2 = 26$$

Los totales son

$$\text{total}_{XX} = 10225.24 - \frac{551^2}{30} = 105.207$$
$$\text{total}_{YY} = 536.5951 - \frac{126.61^2}{30} = 2.2587$$
$$\text{total}_{XY} = 2333.708 - \frac{551\times 126.61}{30} = 8.3047$$
$$gl = 30-1 = 29$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con X.

La estima de β del modelo del análisis de covarianza se realiza, como en los modelos anteriormente estudiados, a partir de la fracción residual de la suma de productos y de la suma de cuadrados de X, es decir

$$b = \frac{8.09088}{95.04004} = 0.08512$$

Por lo que la contribución a la suma de cuadrados total atribuible a la regresión corregida para *dietas*, *raza* y la *interacción* es

$$SC_{\text{Regresión}} = \frac{8.090088^2}{95.04004} = 0.68865$$

La suma de cuadrados del error de la variable Y ajustada para la regresión con X es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = 1.71197 - 0.68865 = 1.02332$$

La regresión de Y sobre X, corrigiendo los valores de Y para los efectos de las diferentes *Dietas* e ignorando los efectos de la *Raza* y de la *interacción,* es

$$b_S = \frac{8.1973}{95.9042} = 0.08547$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{8.1973^2}{95.9042} = 0.70065$$

La suma de cuadrados, ajustada para la regresión, debida a la *dieta* más el error es

$$SC_{\text{Dietas+Error}} = S' = 2.0666 - 0.7006 = 1.3659$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$S' - E' = 0.34262$$

es la cantidad de suma de cuadrados, ajustada para la regresión, atribuible a las *dietas.*

La regresión de Y sobre X, corrigiendo los valores de Y para los efectos de las *Razas* e ignorando los efectos de las *Dietas* y de la *interacción,* es

$$b_U = \frac{7.4849}{99.0732} = 0.07555$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{7.4849^2}{99.0732} = 0.56548$$

La suma de cuadrados, ajustada para la regresión, debida a las *Razas* más el error es

$$SC_{\text{Raza+Error}} = U' = 1.8027 - 0.56548 = 1.2373$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$U' - E' = 0.21388$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a las *razas*.

La regresión de *Y* sobre *X*, corrigiendo los valores de *Y* para los efectos de la *Interacción* e ignorando los efectos de las *Dietas* y de las *Razas,* es

$$b_U = \frac{8.8027}{100.3096} = 0.08775$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{8.8027^2}{100.3096} = 0.77248$$

La suma de cuadrados, ajustada para la regresión, debida a la *interacción* más el error es

$$SC_{\text{Interacción+Error}} = V' = 1.8134 - 0.77248 = 1.04092$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$U' - E' = 0.01760$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a la *interacción*.

Los resultados se pueden resumir en la siguiente tabla

					reducción debida a ***b***		*ajuste debido a* ***b***		
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*	*CM*
Dietas	2	0.8643	0.1072	0.3546					
Razas	1	4.0332	-0.6052	0.0907					
DiexRaz	2	5.2695	0.7126	0.1014					
Error	24	95.0400	8.0901	1.7120	1	0.6887	23	1.0223	0.0445
Die+Err	26	95.9042	8.1973	2.0666	1	0.7006	25	1.3659	
Raz+Err	25	99.0732	7.4849	1.8027	1	0.5655	24	1.2372	
DxR+Err	26	100.3096	8.8027	1.8134	1	0.7725	25	1.0409	
total	29	105.207	8.3047	2.2587					

Dietas ajustadas	2	0.3426	0.1713
Razas ajustados	1	0.2139	0.2139
Interacción ajustada	2	0.0176	0.0088

Todas las posibles pruebas de hipótesis que se pueden plantear son las siguientes:

Para probar el supuesto de que las fuentes de variación no influyen en la covariable, probemos primero las dietas, la F_o de este ANOVA es

$$F_o = \frac{\frac{0.8643}{2}}{\frac{95.04}{24}} = 0.109ns$$

$$F_{(2,24;\,0.05)} = 3.405$$

como es de esperar, la aleatorización ha sido correcta y las dietas no tienen efecto en el peso inicial.

Para las razas la F_o sería

$$F_o = \frac{\frac{4.0332}{1}}{\frac{95.04}{24}} = 1.018ns$$

$$F_{(1,24;\,0.05)} = 4.26$$

como, también, es de esperar, la aleatorización ha sido correcta y no existe el efecto raza en el peso inicial.

Y para la interacción la F_o sería

$$F_o = \frac{\frac{5.2695}{2}}{\frac{95.04}{24}} = 0.665ns$$

$$F_{(2,24;\,0.05)} = 3.40$$

afortunadamente no existe interacción.

Si se quiere probar la hipótesis de igualdad de efectos de las diferentes dietas para la variable Y sin ajustar para la regresión, es decir, como si, desconociendo la variable X, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{0.3546}{2}}{\frac{1.7120}{24}} = 2.486ns$$

$$F_{(2,24;\,0.05)} = 3.40$$

la conclusión sería que, ignorando el peso inicial, las dietas no influyen significativamente

en el incremento de peso.

Si se quiere probar las razas en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{0.0907}{1}}{\frac{1.7120}{24}} = 1.271ns$$

$$F_{(1,24;\,0.05)} = 4.26$$

las razas no han influido en el incremento de peso, ignorando el peso inicial.

Y si se quiere probar la interacción sin ajustar para los valores de la X, la F_o sería

$$F_o = \frac{\frac{0.1014}{2}}{\frac{1.7120}{24}} = 0.711ns$$

$$F_{(2,24;\,0.05)} = 3.40$$

No existe interacción, en el peso final, entre las dietas y las razas

Si la prueba que se desea hacer es la de igualdad de efectos de las dietas pero ajustada para la regresión con la covariable, es decir, la prueba F de las medias ajustadas, la F_o sería

$$F_o = \frac{0.1713}{0.0445} = 3.85^*$$

$$F_{(2,23;\,0.05)} = 3.42$$

la F ha pasado de valer $2.49ns$ a 3.85*, un incremento lo suficientemente grande como para que sean significativas (lo que antes no era) las dietas en el incremento de peso ajustado para el peso inicial.

Si se quieren probar las razas ajustados, la pruebas sería

$$F_o = \frac{0.2139}{0.0445} = 4.807^*$$

$$F_{(1,23;\,0.05)} = 4.28$$

la F ha pasado de valer $1.271ns$ a 4.807*, un incremento lo suficientemente grande como para que sean significativas (lo que antes no era) las razas en el incremento de peso ajustado para el peso inicial.

Si se quiere probar la interacción ajustada, la pruebas sería

$$F_o = \frac{0.0080}{0.0445} = 0.18ns$$
$$F_{(2,23;\,0.05)} = 3.42$$

Afortunadamente, sigue sin ser significativa

Es decir, el ANOVA no detecta diferencias significativas, en el *incremento de peso,* ni para las *dietas* ni para las *razas*, pero cuando éste se ajusta para la regresión con el *peso al inicio*, el análisis de covarianza si detecta diferencias para las tres dietas y para las dos razas. No existe interacción en ningún caso.

Recuérdese que en el caso de que hubiera sido un modelo mixto, el factor fijo se habría contrastado con el cuadrado medio de la interacción y no con el del error (véase el capítulo 7). En este ejemplo, si consideramos *razas* como aleatorio, el contraste para las *dietas*, que son de efectos fijos, sería

$$F_o = \frac{0.1713}{0.0088} = 19.466\,*$$
$$F_{(2,2;\,0.05)} = 19.00$$

las dietas son significativas.

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, la prueba F sería

$$F_o = \frac{\frac{8.09009^2}{95.04004}}{\frac{1.0233}{23}} = 15.478\,***$$
$$F_{(1,23;\,0.001)} = 9.47$$

efectivamente existe una regresión significativa del peso inicial en el peso final, después de haber corregido para los efectos de los dos factores y la interacción.

El cálculo de las medias ajustadas (medias minimocuadráticas) de las dietas es

$$\hat{Y}_{D_1} = 4.357 - 0.08512 \times (18.3 - 18.367) = 4.3627$$
$$\hat{Y}_{D_2} = 4.213 - 0.08512 \times (18.6 - 18.367) = 4.1932$$
$$\hat{Y}_{D_3} = 4.091 - 0.08512 \times (18.2 - 18.367) = 4.1052$$

De las razas

$$\hat{Y}_{R_1} = 4.1653 - 0.08512 \times (18.733 - 18.367) = 4.1342$$
$$\hat{Y}_{R_2} = 4.2753 - 0.08512 \times (18 - 18.367) = 4.3066$$

Y de las casillas (interacción)

$$\hat{Y}_{C_{11}} = 4.338 - 0.08512 \times (18.8 - 18.367) = 4.3011$$
$$\hat{Y}_{C_{21}} = 4.076 - 0.08512 \times (18.4 - 18.367) = 4.0732$$
$$\hat{Y}_{C_{31}} = 4.082 - 0.08512 \times (19.0 - 18.367) = 4.0281$$
$$\hat{Y}_{C_{12}} = 4.376 - 0.08512 \times (17.8 - 18.367) = 4.4243$$
$$\hat{Y}_{C_{22}} = 4.350 - 0.08512 \times (18.8 - 18.367) = 4.3131$$
$$\hat{y}_{C_{32}} = 4.100 - 0.08512 \times (17.4 - 18.367) = 4.1823$$

Los *errores típico de las medias ajustadas* de la primera dieta, de la primer raza y de la primera casilla, por ejemplo, son, respectivamente

Primera dieta

$$S_{\hat{Y}_{D_1}} = \sqrt{0.0445\left[\frac{1}{10} + \frac{(18.3 - 18.367)^2}{95.04}\right]} = 0.0667$$

Primera raza

$$S_{\hat{Y}_{R_1}} = \sqrt{0.0445\left[\frac{1}{15} + \frac{(18.733 - 18.367)^2}{95.04}\right]} = 0.0550$$

Primera dieta dentro de la primera raza

$$S_{\hat{Y}_{C_{11}}} = \sqrt{0.0445\left[\frac{1}{5} + \frac{(18.8 - 18.367)^2}{95.04}\right]} = 0.0948$$

Recuérdese que estos errores típicos cambian si el modelo es mixto o aleatorio, puesto que el cuadrado medio que se tomará será el de la interacción y no el del error. Y por lo tanto, cambiaría el valor de la prueba ***t*** de contrastes de medias.

La diferencia entre las medias ajustadas de la primera y segunda dieta, del primer y segunda raza, y de la primera y segunda casilla es, respectivamente

Diferencia entre la 1° y 2° dieta

$$\hat{Y}_{D_1} - \hat{Y}_{D_2} = 4.357 - 4.213 - 0.08512(18.3 - 18.6) = 0.1695$$

Diferencia entre la 1° y 2° raza

$$\hat{Y}_{S_1} - \hat{Y}_{S_2} = 4.165 - 4.275 - 0.08512(18.733 - 18.0) = -0.1724$$

Diferencia entre la 1° y 2° dieta dentro de la 1° raza

$$\hat{Y}_{C_{11}} - \hat{Y}_{C_{21}} = 4.338 - 4.076 - 0.08512(18.8 - 18.4) = 0.2280$$

El *error típico de la diferencia de las medias ajustadas* de la primera y segunda

dieta, del primer y segunda raza, y de la primera y segunda casilla es, respectivamente

$$S_{\hat{Y}_{D1}-\hat{Y}_{D2}} = \sqrt{0.0445\left[\frac{2}{10}+\frac{(18.3-18.6)^2}{95.04}\right]} = 0.0945$$

$$S_{\hat{Y}_{R1}-\hat{Y}_{R}} = \sqrt{0.0445\left[\frac{2}{15}+\frac{(41.65-42.75)^2}{95.04}\right]} = 0.0771$$

$$S_{\hat{Y}_{C11}-\hat{Y}_{C21}} = \sqrt{0.0445\left[\frac{2}{5}+\frac{(18.8-18.4)^2}{95.04}\right]} = 0.1337$$

La prueba *t* para contrastar si son significativas las diferencias entre, por ejemplo, las medias ajustadas de la primera y segunda dieta, del primer y segunda raza, y de la primera y segunda casilla son

Primera y segunda dieta

$$t = \frac{0.1695}{0.0945} = 1.7936ns$$

$$t_{(23;\,0.05/2)} = 2.0687$$

Primera y segunda raza

$$t = \frac{-0.1724}{0.0771} = -2.2361*$$

$$t_{(23;\,0.05/2)} = 2.0697$$

Primera y segunda dieta en la primera raza

$$t = \frac{0.228}{0.1337} = 1.7053ns$$

Las dos raza son significativamente diferentes, mientras que la primera y segunda dieta no son significativamente diferentes y, como no existe interacción, no lo son ni en el total ni en la primera raza.

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias antes y después del ajuste.

El cuadrado medio del error antes del ajuste es

$$CM_E = \frac{1.711975}{24} = 0.07133$$

El cuadrado medio del error efectivo después del ajuste con *X* es para las tres fuentes de variación, respectivamente

Para las dietas

$$S'^2_{Y.X(D)} = 0.04449\left[1+\frac{0.86426}{2\times 95.04004}\right] = 0.04469$$

Para las razas

$$S'^2_{Y.X(R)} = 0.04449\left[1+\frac{4.0332}{1\times 95.04004}\right] = 0.04638$$

Para la interacción

$$S'^2_{Y.X(DR)} = 0.04449\left[1+\frac{5.2695}{2\times 95.04004}\right] = 0.04572$$

Un estimador de la precisión relativa es la razón del cuadrado medio del error sin ajustar por el cuadrado medio del error ajustado, multiplicado por 100 para expresarlo en porcentajes.

Para los tratamientos

$$\frac{0.071333}{004469}100 = 159.6\%$$

Para las razas

$$\frac{0.071333}{0.046378}100 = 153.8\%$$

Para la interacción

$$\frac{0.071333}{0.045723}100 = 156.0\%$$

Se constata que ha habido un considerable aumento de la precisión en la prueba de las tres fuentes de variación.

Archivo del programa *SAS* (C17-3.SAS).-

```
title 'Análisis de Covarianza de dos factores con repetición';
options ls=75 ps=60;
data cov_fact;
infile 'c17-3.dat';
input dieta $ raza $ x y @@;
title 'ANOVA de la variable X';
proc anova;
 class dieta raza;
   model x  = dieta raza dieta*raza;
run;
title 'ANOVA de la variable Y';
proc anova;
 class dieta raza;
   model y  = dieta raza dieta*raza;
run;
title 'Análisis de Covarianza de dos factores con repetició';
proc glm;
 class dieta raza;
   model y  = dieta raza dieta*raza x / solution;
```

```
    lsmeans dieta raza / stderr tdiff;
run;
title 'Si razas es aleatorio, la prueba de las dietas es';
test h=dieta e=dieta*raza;
lsmeans dieta raza / e=dieta*raza stderr tdiff;
 run;
```

Archivo de datos (C17-3.DAT).-

```
A 1 14.8  4.22  B 1 15.8  3.78  C 1 19.8  4.07
A 1 18.9  3.84  B 1 19.9  4.65  C 1 19.9  4.02
A 1 19.2  4.43  B 1 21.2  4.04  C 1 19.2  4.28
A 1 19.7  4.46  B 1 15.7  3.71  C 1 16.7  4.02
A 1 21.4  4.74  B 1 19.4  4.20  C 1 19.4  4.02
A 2 19.8  4.41  B 2 19.8  4.46  C 2 19.8  4.36
A 2 17.9  4.42  B 2 17.9  4.33  C 2 15.9  4.09
A 2 16.2  4.43  B 2 19.2  4.46  C 2 15.2  3.70
A 2 18.7  4.61  B 2 18.7  4.08  C 2 20.7  4.41
A 2 16.4  4.01  B 2 18.4  4.42  C 2 15.4  3.94
```

Archivo de resultados (C17-3.LST).-

ANOVA de la variable X

Dependent Variable: X

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10.1666667	2.0333333	0.51	0.7634
Error	24	95.0400000	3.9600000		
Corrected Total	29	105.2066667			

R-Square	C.V.	Root MSE	X Mean
0.096635	10.83471	1.98997	18.3667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DIETA	2	0.86666667	0.43333333	0.11	0.8968
RAZA	1	4.03333333	4.03333333	1.02	0.3229
DIETA*RAZA	2	5.26666667	2.63333333	0.66	0.5235

ANOVA de la variable Y

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.54669667	0.10933933	1.53	0.2171
Error	24	1.71200000	0.07133333		
Corrected Total	29	2.25869667			

R-Square	C.V.	Root MSE	Y Mean
0.242041	6.328481	0.26708	4.22033

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DIETA	2	0.35458667	0.17729333	2.49	0.1045
RAZA	1	0.09075000	0.09075000	1.27	0.2705
DIETA*RAZA	2	0.10136000	0.05068000	0.71	0.5015

Análisis de Covarianza de dos factores con repetició

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.23533408	0.20588901	4.63	0.0032
Error	23	1.02336258	0.04449403		
Corrected Total	29	2.25869667			

R-Square	C.V.	Root MSE	Y Mean
0.546923	4.998090	0.21094	4.22033

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DIETA	2	0.35458667	0.17729333	3.98	0.0327
RAZA	1	0.09075000	0.09075000	2.04	0.1667
DIETA*RAZA	2	0.10136000	0.05068000	1.14	0.3375
X	1	0.68863742	0.68863742	15.48	0.0007

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DIETA	2	0.34258172	0.17129086	3.85	0.0361
RAZA	1	0.21389494	0.21389494	4.81	0.0387
DIETA*RAZA	2	0.01761402	0.00880701	0.20	0.8218
X	1	0.68863742	0.68863742	15.48	0.0007

Parameter		Estimate		T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		2.618876263	B	6.75	0.0001	0.38812272
DIETA	A	0.241951178	B	1.81	0.0834	0.13368813
	B	0.130829125	B	0.96	0.3489	0.13680353
	C	0.000000000	B	.	.	.
RAZA	1	-0.154195286	B	-1.12	0.2748	0.13782635
	2	0.000000000	B	.	.	.
DIETA*RAZA	A 1	0.031073232	B	0.16	0.8709	0.18911308
	A 2	0.000000000	B	.	.	.
	B 1	-0.085755892	B	-0.44	0.6619	0.19356618
	B 2	0.000000000	B	.	.	.
	C 1	0.000000000	B	.	.	.
	C 2	0.000000000	B	.	.	.
X		0.085122054		3.93	0.0007	0.02163703

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

Least Squares Means

DIETA	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number
A	4.36267480	0.06671944	0.0001	1
B	4.19313819	0.06689463	0.0001	2
C	4.10518701	0.06680125	0.0001	3

T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|

i/j	1	2	3
1	.	1.792965	2.728831
		0.0861	0.0120
2	-1.79297	.	0.928444
	0.0861		0.3628
3	-2.72883	-0.92844	.
	0.0120	0.3628	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

RAZA	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	T / Pr > |T| H0: LSMEAN1=LSMEAN2
1	4.13412191	0.05503826	0.0001	-2.19255
				0.0387
2	4.30654475	0.05503826	0.0001	

Standard Errors and Probabilities calculated using the Type III MS for DIETA*RAZA as an Error term

DIETA	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number
A	4.36267480	0.02968354	0.0001	1
B	4.19313819	0.02976149	0.0001	2
C	4.10518701	0.02971994	0.0001	3

T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|

i/j	1	2	3
1	.	4.030033	6.133568
		0.0564	0.0256
2	-4.03003	.	2.086855
	0.0564		0.1722
3	-6.13357	-2.08685	.
	0.0256	0.1722	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

```
  Standard Errors and Probabilities calculated using the Type III MS for
                        DIETA*RAZA as an Error term
     RAZA                Y       Std Err      Pr > |T|     T / Pr > |T| H0:
                    LSMEAN        LSMEAN    H0:LSMEAN=0    LSMEAN1=LSMEAN2
     1          4.13412191    0.02448658          0.0001        -4.92817
                                                                 0.0388
     2          4.30654475    0.02448658          0.0001

Tests of Hypotheses using the Type III MS for DIETA*RAZA as an error term
Source                     DF     Type III SS    Mean Square  F Value     Pr > F
DIETA                       2      0.34258172     0.17129086    19.45     0.0489
```

Estos son los mismos resultados que los obtenidos manualmente, pero con un menor error de redondeo.

La suma de cuadrados que hay que mirar es la **tipo III**. La suma de cuadrados **tipo I** debida a los factores (dieta y raza) e interacción es la del ANOVA para la variable *Y*. Si, en el modelo del programa SAS, se pone primero la covariable y después los factores, esto es, **model y = x dieta raza dieta*raza/ solution;** se hubiera obtenido la misma suma de cuadrados debida al último factor del modelo (en este caso *la interacción*), en el tipo I y en el tipo III, pero lo suma de cuadrados tipo I debida a la regresión, sería la de la regresión sin considerar las correcciones de los factores, esto es, la de la regresión de los treinta pares de valores, como si las dietas y las razas no existieran, y la suma de cuadrados tipo I de los factores (dieta y raza) sería la suma de cuadrados del ANCOVA para solo estos factores, como si el último factor no existiera.

Análisis de covarianza de un modelo jerárquico o con subgrupos.-

Siguiendo el mismo orden del presentado en los capítulos del análisis de la varianza, se va ha estudiar el análisis de covarianza de diseños jerárquicos (anidados o encajados), si bien no existe ninguna novedad con respecto a lo estudiado en los anteriores diseños. La descomposición de las componentes de la varianza y de los grados de libertad se realiza tal como se estudió en el capítulo 8, tanto para la suma de cuadrados de X como de Y así como para la suma de productos, utilizando el tipo de anotación específica de este capítulo. La regresión se calcula en la fila del error y las sumas de cuadrados ajustadas para ambas fuentes de variación se calculan restándole a la suma de cuadrados ajustada del factor más el error, la suma de cuadrados ajusta del error.

El modelo lineal que se va ha utilizar es

$$Y_{ijk} = \mu + T_i + R_{j(i)} + \beta(X_{ijk} - \overline{X}...) + \varepsilon_{ijk}$$

Como en el modelo anterior, también hay dos fuentes de variación y medidas repetidas, pero al estar un factor jerárquico al otro (factor principal) no se puede incluir la interacción en el modelo, pues ésta no se puede estimar, ver el epígrafe *Interacciones* del Capítulo 9.

La descomposición de la variación total y de los grados de libertad total es:

La componente *error*

$$E_{XX} = \Sigma_{ijk} X_{ijk}^2 - \Sigma_{ij} \frac{X_{ij.}^2}{n_{ij.}}$$

$$E_{YY} = \Sigma_{ijk} Y_{ijk}^2 - \Sigma_{ij} \frac{Y_{ij.}^2}{n_{ij.}}$$

$$E_{XY} = \Sigma_{ijk} X_{ijk} Y_{ijk} - \Sigma_{ij} \frac{X_{ij.} Y_{ij.}}{n_{ij.}}$$

$$gl = N - r$$

La debida al factor principal, T

$$T_{XX} = \Sigma_i \frac{X_{i..}^2}{n_{i..}} - \frac{X_{...}^2}{N}$$

$$T_{YY} = \Sigma_i \frac{Y_{i..}^2}{n_{i..}} - \frac{Y_{...}^2}{N}$$

$$T_{XY} = \Sigma_i \frac{X_{i..} Y_{i..}}{n_{i..}} - \frac{X_{...} Y_{...}}{N}$$

$$gl = t - 1$$

La debida al factor jerárquico, R

$$R_{XX} = \Sigma_{ij}\frac{X_{ij.}^2}{n_{ij.}} - \Sigma_i \frac{X_{i..}^2}{n_{i..}}$$

$$R_{YY} = \Sigma_{ij}\frac{Y_{ij.}^2}{n_{ij.}} - \Sigma_i \frac{Y_{i..}^2}{n_{i..}}$$

$$R_{XY} = \Sigma_{ij}\frac{X_{ij.}Y_{ij.}}{n_{ij.}} - \Sigma_i \frac{X_{i..}Y_{i..}}{n_{i..}}$$

$$gl = r - t$$

La suma de cuadrados del factor principal, T, más el error es

$$S_{XX} = E_{XX} + T_{XX} = \Sigma_{ijk} X_{ijk}^2 + \Sigma_i \frac{X_{i..}^2}{n_{i..}} - \Sigma_{ij}\frac{X_{ij.}^2}{n_{ij.}} - \frac{X_{...}^2}{N}$$

$$S_{YY} = E_{YY} + T_{YY} = \Sigma_{ijk} Y_{ijk}^2 + \Sigma_i \frac{Y_{i..}^2}{n_{i..}} - \Sigma_{ij}\frac{Y_{ij.}^2}{n_{ij.}} - \frac{X_{...}^2}{N}$$

$$S_{XY} = E_{XY} + T_{XY} = \Sigma_{ijk} X_{ijk} Y_{ijk} + \Sigma_i \frac{X_{i..}Y_{i..}}{n_{i..}} - \Sigma_{ij}\frac{X_{ij.}Y_{ij.}}{n_{ij.}} - \frac{X_{...}Y_{...}}{N}$$

$$gl = N - r + t - 1$$

Y la suma de cuadrados del factor jerárquico, R, más el error es

$$U_{XX} = E_{XX} + R_{XX} = \Sigma_{ijk} X_{ijk}^2 - \Sigma_i \frac{X_{i..}^2}{n_{i..}}$$

$$U_{YY} = E_{YY} + R_{YY} = \Sigma_{ijk} Y_{ijk}^2 - \Sigma_i \frac{Y_{i..}^2}{n_{i..}}$$

$$U_{XY} = E_{XY} + R_{XY} = \Sigma_{ijk} X_{ijk} Y_{ijk} - \Sigma_i \frac{X_{i..}Y_{i..}}{n_{i..}}$$

$$gl = N - t$$

Como se puede ver, ni S ni U coincide con los totales que son

$$\text{total}_{XX} = \Sigma_{ijk} X_{ijk}^2 - \frac{X_{...}^2}{N}$$

$$\text{total}_{YY} = \Sigma_{ijk} Y_{ijk}^2 - \frac{Y_{...}^2}{N}$$

$$\text{total}_{XY} = \Sigma_{ijk} X_{ijk} Y_{ijk} - \frac{X_{...} Y_{...}}{N}$$

$$gl = N - 1$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con *X*.

Ahora, y tal como se hizo en el análisis de los modelos anteriores, mediante un proceso secuencial, se le va a extraer a cada suma de cuadrados la componente debida a la regresión quedando un residuo más pequeño para el análisis de covarianza.

La estima de β del modelo del análisis de covarianza es, como ya se ha visto en los anteriores modelos

$$b = \frac{E_{XY}}{E_{XX}}$$

es decir, es la fracción residual de la suma de los productos que queda después de quitarle a la suma de productos total las debidas a los dos factores, dividida por la suma de cuadrados residual de la variable *X*, es decir, la suma de cuadrados total de *X* menos la suma de cuadrados debida a los dos factores. Por lo que la contribución a la suma de cuadrados atribuible a la regresión ajustada para ambos factores es

$$SC_{\text{Regresión}} = b\, E_{XY} = \frac{E_{XY}^2}{E_{XX}}$$

La suma de cuadrados del error de la variable *Y* ajustada para la regresión con la covariable es, pues, la diferencia entre la suma de cuadrados del error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = E_{YY} - \frac{E_{XY}^2}{E_{XX}}$$

Si la regresión de *Y* sobre *X*, sin eliminar los efectos del factor principal, *T* y dejando los efectos del factor jerárquico es

$$b_S = \frac{S_{XY}}{S_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_S S_{XY} = \frac{S_{XY}^2}{S_{XX}}$$

La suma de cuadrados, ajustada para la regresión, debida al factor principal más el error es

$$SC_{\text{T+Error}} = S' = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (S'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible al factor principal.

Si la regresión de Y sobre X, sin eliminar los efectos del factor jerárquico y dejando los efectos del factor principal es

$$b_U = \frac{U_{XY}}{U_{XX}}$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = b_U U_{XY} = \frac{U_{XY}^2}{U_{XX}}$$

La suma de cuadrados, ajustada para la regresión, debida al factor jerárquico más el error es

$$SC_{\text{R(T)+Error}} = U' = U_{YY} - \frac{U_{XY}^2}{U_{XX}}$$

Por lo que la diferencia entre las sumas de cuadrados residuales (U'-E') es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible al factor jerárquico.

El análisis se puede resumir en la siguiente tabla

					reducción debida a **b**		ajuste debido a **b**	
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*
T	*t*-1	T_{XX}	T_{XY}	T_{YY}				
R(T)	*r*-*t*	R_{XX}	R_{XY}	R_{YY}				
Error	*N*-*r*	E_{XX}	E_{XY}	E_{YY}	1	$\frac{E_{XY}^2}{E_{XY}}$	*N*-*r*-1	*E'*
T+*Err*	*N*+*t*-*r*-1	S_{XX}	S_{XY}	S_{YY}	1	$\frac{S_{XY}^2}{S_{XY}}$	*N*+*t*-*r*-2	*S'*
R(T)+*Err*	*N*-*t*	U_{XX}	U_{XY}	U_{YY}	1	$\frac{U_{XY}^2}{U_{XY}}$	*N*-*t*-1	*U'*
total	*N*-1	*SC*	*SP*	*SC*				
Factor principal ajustado							*t*-1	*S'*-*E'*
Factor jerárquico ajustado							*r*-*t*	*U'*-*E'*

Ajustes de las medias.-

Las fórmulas para el ajuste de las medias son las mismas de las dadas para los anteriores modelos. El cálculo de una media ajustada (o media minimocuadrática) de los diferentes niveles del factor principal y del factor jerárquico son, respectivamente

Medias ajustadas del factor *T*

$$\hat{Y}_{i..} = \overline{Y}_{i..} - b(\overline{X}_{i..} - \overline{X}_{...})$$

Medias ajustadas del factor *R(T)*

$$\hat{Y}_{ij.} = \overline{Y}_{ij.} - b(\overline{X}_{ij.} - \overline{X}_{i..})$$

donde *b* es el coeficiente de regresión del error.

Los *errores típico de las medias ajustadas* se calculan, respectivamente

Factor T

$$S_{\hat{Y}_{i..}} = S_{Y.X} \sqrt{\frac{1}{n_{i..}} + \frac{(\overline{X}_{i..} - \overline{X}_{...})^2}{E_{XX}}}$$

Factor $R(T)$

$$S_{\hat{Y}_{ij.}} = S_{Y.X} \sqrt{\frac{1}{n_{ij.}} + \frac{(\overline{X}_{ij.} - \overline{X}_{i..})^2}{E_{XX}}}$$

siendo $S_{Y.X}$ la raíz cuadrada del cuadrado medio del error ajustado.

La diferencia entre las medias ajustadas del *i-ésimo* y *j-ésimo* nivel del factor T es

$$\hat{Y}_{i..} - \hat{Y}_{j..} = \overline{Y}_{i..} - \overline{Y}_{j..} - b(\overline{X}_{i..} - \overline{X}_{j..})$$

Y la diferencia entre las medias ajustadas del *i-ésimo* y *j-ésimo* nivel del factor R dentro del nivel *k-ésimo* del factor T es

$$\hat{Y}_{ki.} - \hat{Y}_{kj.} = \overline{Y}_{ki.} - \overline{Y}_{kj.} - b(\overline{X}_{ki.} - \overline{X}_{kj.})$$

El *error típico de la diferencia de dos medias ajustadas* del factor principal es

$$S_{\hat{Y}_{i..} - \hat{Y}_{j..}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n_t} + \frac{(\overline{X}_{i..} - \overline{X}_{j..})^2}{E_{XX}}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{t_i.}} + \frac{1}{n_{t_j.}} + \frac{(\overline{X}_{i..} - \overline{X}_{j..})^2}{E_{XX}}\right]}$$

La expresión de la derecha es para el caso en que los tamaños de submuestras para los diferentes tratamientos (n_t) sean diferentes.

Y el *error típico de la diferencia de dos medias ajustadas* del factor jerárquico dentro del *k-ésimo* nivel del factor principal es

$$S_{\hat{Y}_{ki.} - \hat{Y}_{kj.}} = \sqrt{S^2_{X.Y}\left[\frac{2}{n_{t_k}} + \frac{(\overline{X}_{ki.} - \overline{X}_{kj.})^2}{E_{XX}}\right]} = \sqrt{S^2_{X.Y}\left[\frac{1}{n_{t_{ki}}} + \frac{1}{n_{t_{kj}}} + \frac{(\overline{X}_{ki.} - \overline{X}_{kj.})^2}{E_{XX}}\right]}$$

Con estas expresiones se pueden hacer comparaciones múltiples de medias tal como se estudió en el Capítulo 9, teniendo en cuenta que hay que usar los cuadrados medios del análisis de covarianza con los grados de libertad correspondientes. Las pruebas más correctas en este caso son la *t* y la de *Scheffe*, si bien la más comúnmente utilizada es la prueba *t*.

Pruebas de hipótesis.-

Consúltese los epígrafes del capítulo 8, *Parámetros estimados en un modelo jerárquico* y *Pruebas de hipótesis de un modelo jerárquico* en los que se explica que el término de contraste para la prueba del factor principal suele ser el cuadrado medio del factor jerárquico, mientras que el término de contraste del factor jerárquico es el cuadrado medio del *error*, en ese caso todas las posibles pruebas de hipótesis son:

Si se quiere probar el supuesto de que la covariable no esta influida por los efectos de los factores, sería realizar un ANOVA, que con los resultados de la tabla anterior, se realizarían con las siguientes F_o.

Para el factor principal

$$F_o = \frac{\frac{T_{XX}}{t-1}}{\frac{R_{XX}}{r-t}}$$

que se contrastaría con la $F_{(t-1,r-t;\,\alpha)}$

Y para el factor jerárquico la F_o sería

$$F_o = \frac{\frac{R_{XX}}{r-t}}{\frac{E_{XX}}{N-r}}$$

que se contrastaría con la $F_{(r-t,N-r;\,\alpha)}$

Si se quiere probar la hipótesis de igualdad de efectos, en la variable *Y*, de los diferentes niveles del factor principal, *T*, sin ajustar para la regresión, es decir, como si, desconociendo la variable *X*, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{T_{YY}}{t-1}}{\frac{E_{YY}}{r-t}}$$

que se contrastaría con la $F_{(t-1,r-t;\,\alpha)}$

Si se quieren probar el factor jerárquico en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{R_{YY}}{r-t}}{\frac{E_{YY}}{N-r}}$$

que se contrastaría con la $F_{(r-t,N-r;\,\alpha)}$

Si la prueba que se desea hacer es la de igualdad de efectos de los niveles del factor principal en la variable *Y* ajustada para la regresión con la covariable, es decir, la

prueba F de las medias ajustadas, a semejanza de la anterior, la F_o sería

$$F_o = \frac{\frac{S'-E'}{t-1}}{\frac{U'-E'}{r-t}}$$

que se contrastaría con la $F_{(t-1,r-t;\,\alpha)}$.

Si se quieren probar el factor jerárquico ajustado, la pruebas sería

$$F_o = \frac{\frac{U'-E'}{r-t}}{\frac{E'}{N-r-1}}$$

que se contrastaría con la $F_{(r-t,N-r-1;\,\alpha)}$.

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, a semejanza de la prueba F del Capítulo 11, la prueba F en este caso sería

$$F_o = \frac{\frac{E_{XY}^2}{E_{XX}}}{\frac{E'}{(N-r-1)}}$$

que se contrastaría con la $F_{(1,N-r-1;\,\alpha)}$

Aumento de precisión debido a la covarianza.-

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de los diferentes niveles de los factores antes y después del ajuste.

Los cuadrados medios de los dos términos de contraste, es decir, el del factor $R(T)$ y el del *error*, antes del ajuste son

$$CM_{\text{error}} = \frac{E_{YY}}{N-r}$$

$$CM_{R(T)} = \frac{R_{YY}}{r-t}$$

El cuadrado medio efectivo de los término de contraste después del ajuste para X, para el factor principal T y para el factor jerárquico $R(T)$ son, respectivamente

Para el factor T

$$S'^2_{Y.X(T)} = S^2_{Y.X}\left[1+\frac{T_{XX}}{(t-1)R_{XX}}\right]$$

Para el factor $R(T)$

$$S'^2_{Y.X(R_{(T)})} = S^2_{Y.X}\left[1+\frac{R_{XX}}{(r-t)E_{XX}}\right]$$

Un estimador de la precisión relativa es la razón del cuadrado medio del término de contraste sin ajustar por el cuadrado medio ajustado, multiplicado por 100 para expresarlo en porcentajes.

Para T

$$\frac{\frac{R_{YY}}{rt}}{S'^2_{Y.X(T)}}100$$

Para $R(T)$

$$\frac{\frac{E_{YY}}{N-r}}{S'^2_{Y.X(R_{(T)})}}100$$

Ejemplo.-

Se quiere probar la eficacia de cuatro *insecticidas* para el ganado, para ello se toman doce *animales* y aleatoriamente se separan en cuatro grupos a cada uno de los cuales se le aplicará uno de los insecticidas. La aplicación del insecticida se realiza rociando cada individuo después de cuatro conteos de parásitos (X) en cuatro zonas del cuerpo, transcurrido el tiempo adecuado, se vuelve a medir el número de parásitos (Y) en las mismas cuatro zonas.

		Zona										
		1		2		3		4				
Insecticida	*Animal*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	ΣX	ΣY	ΣXY
1	1	52	10	66	21	54	13	52	11	224	55	3180
	2	64	20	53	11	56	19	65	22	238	72	4357
	3	51	8	64	18	48	8	52	10	215	44	2464
2	4	58	13	42	5	59	14	48	7	207	39	2126
	5	54	11	63	17	55	13	66	19	238	60	3634
	6	53	9	38	1	46	3	43	1	180	14	696
3	7	49	6	43	4	60	14	56	12	208	36	1978
	8	56	12	57	13	44	2	47	4	204	31	1689
	9	42	4	43	3	62	15	59	14	206	36	2053
4	10	55	8	41	1	52	7	44	2	192	18	933
	11	41	1	55	9	45	2	56	10	197	22	1186
	12	52	6	37	1	51	7	49	5	189	19	951

Los sumatorios para los insecticidas y totales son

Insecticida	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY
1	677	171	38651	2749	10001
2	625	113	33397	1471	6456
3	618	103	32454	1171	5720
4	578	59	28288	415	3070
total	2498	446	132790	5806	25247

Primeramente se calculan las sumas de cuadrados y las sumas de productos tal como se estudió en el capítulo 8

$$E_{XX} = 132790 - \frac{\left(\begin{array}{l}224^2+238^2+215^2+207^2+238^2+180^2+\\+208^2+204^2+206^2+192^2+197^2+189^2\end{array}\right)}{4} = 1878.00$$

$$E_{YY} = 5806 - \frac{\left(\begin{array}{l}55^2+72^2+44^2+39^2+60^2+14^2+\\+36^2+31^2+36^2+18^2+22^2+19^2\end{array}\right)}{4} = 760.00$$

$$E_{XY} = 25247 -$$

$$- \frac{\left(\begin{array}{l}224\times55+238\times72+215\times44+207\times39+238\times60+180\times14+\\208\times36+204\times31+206\times36+192\times18+197\times22+189x19\end{array}\right)}{4} =$$

$$= 1147.50$$

$$gl = 48 - 12 = 36$$

$$T_{XX} = \frac{677^2+625^2+618^2+578^2}{12} - \frac{2498^2}{48} = 413.4141$$

$$T_{YY} = \frac{171^2+113^2+103^2+59^2}{12} - \frac{446^2}{48} = 530.9165$$

$$T_{XY} = \frac{677\times171+625\times113+618\times103+578\times59}{12} - \frac{2498\times446}{48} = 468.416$$

$$gl = (4-1) = 3$$

$$R_{XX} = \frac{\left(\begin{array}{l} 224^2+238^2+215^2+207^2+238^2+180^2+ \\ +208^2+204^2+206^2+192^2+197^2+189^2 \end{array}\right)}{4} -$$

$$-\frac{677^2+625^2+618^2+578^2}{12} = 498.50$$

$$R_{YY} = \frac{\left(\begin{array}{l} 55^2+72^2+44^2+39^2+60^2+14^2+ \\ +36^2+31^2+36^2+18^2+22^2+19^2 \end{array}\right)}{4} -$$

$$-\frac{171^2+113^2+103^2+59^2}{12} = 371.00$$

$$R_{XY} = \frac{\left(\begin{array}{l} 224\times 55+238\times 72+215\times 44+207\times 39+238\times 60+180\times 14+ \\ 208\times 36+204\times 31+206\times 36+192\times 18+197\times 22+189x19 \end{array}\right)}{4} -$$

$$-\frac{677\times 171+625\times 113+618\times 103+578\times 59}{12} =$$

$$= 420.50$$

$$gl = 12-4 = 8$$

La suma de cuadrados y suma de productos del error más los *insecticidas* es

$$S_{XX} = E_{XX}+T_{XX} = 1878.00+413.4141 = 2291.41411$$
$$S_{YY} = E_{YY}+T_{YY} = 760.00+530.9165 = 1290.9165$$
$$S_{XY} = E_{XY}+T_{XY} = 1147.50+468.416 = 1615.916$$
$$gl = 36+3 = 39$$

La suma de cuadrados y suma de productos del error más los *animales* es

$$U_{XX} = E_{XX}+R_{XX} = 1878.00+498.50 = 2376.50$$
$$U_{YY} = E_{YY}+R_{YY} = 760.00+371.00 = 1131.00$$
$$U_{XY} = E_{XY}+R_{XY} = 1147.50+420.50 = 1568.00$$
$$gl = 36+8 = 44$$

Los totales son

$$\text{total}_{XX} = 132790 - \frac{2498^2}{48} = 2789.9141$$

$$\text{total}_{YY} = 5806 - \frac{446^2}{48} = 1661.9165$$

$$\text{total}_{XY} = 25247 - \frac{2498 \times 446}{48} = 2036.416$$

$$gl = 48 - 1 = 47$$

Estas son las sumas de cuadrados y las sumas de productos aún no ajustadas para las desviaciones de la regresión con *X*.

La estima de β del modelo del análisis de covarianza es

$$b = \frac{1147.50}{1878.00} = 0.6110$$

Por lo que la contribución a la suma de cuadrados atribuible a la regresión ajustada para *insecticida* y *animal* es

$$SC_{\text{Regresión}} = \frac{1147.50^2}{1878.00} = 701.1482$$

La suma de cuadrados de la variable *Y* debida al error, ajustada para la regresión con la covariable es, pues, la diferencia entre la suma de cuadrados debida al error y la suma de cuadrados debida a la regresión

$$SC_{\text{Error}} = E' = 760.00 - 701.1482 = 58.8518$$

La regresión de *Y* sobre *X*, sin eliminar el efecto de los *insecticidas* y dejando los efectos de *animal* es

$$b_S = \frac{1615.916}{2291.1141} = 0.7053$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{1615.916^2}{2291.4141} = 1139.552$$

La suma de cuadrados de la variable *Y*, ajustada para la regresión con la covariable, debida a los *insecticidas* más el error es

$$SC_{\text{Insecticidas+Error}} = S' = 1290.9165 - 1139.552 = 151.364$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$S' - E' = 92.5127$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a los *insecticidas*.

Si la regresión de Y sobre X, sin eliminar el efecto de los *animales* y dejando los efectos de los *insecticidas* es

$$b_U = \frac{1568.00}{2376.5} = 0.65979$$

la suma de cuadrados atribuible a esta regresión es

$$SC_{\text{Regresión}} = \frac{1568^2}{2376.5} = 1034.557$$

La suma de cuadrados, ajustada para la regresión, debida a los *animales* más el error es

$$SC_{\text{Animales+Error}} = U' = 1131.00 - 1034.557 = 96.443$$

Por lo que la diferencia entre las sumas de cuadrados residuales

$$U' - E' = 37.5915$$

es la cantidad de suma de cuadrados, ajustada para la desviación de la regresión, atribuible a los *animales*.

Los resultados se pueden resumir en la siguiente tabla

					reducción debida a b		*ajuste debido a b*		
FV	*gl*	SC_X	SP_{XY}	SC_Y	*gl*	*SC*	*gl*	*SC*	*CM*
Insecti	3	413.41	469.4	530.92					
Ani(in)	8	498.50	420.5	371.00					
Error	36	1878.91	1147.5	760.00	1	701.85	35	58.85	1.687
Ins+Err	39	2291.41	1615.9	1290.92	1	1139.55	38	151.36	
Ani+Err	44	2376.50	1568.0	1131.00	1	1034.56	43	96.44	
total	47	2789.91	2036.4	1661.92					
Insecticidas ajustados							3	92.51	30.84
Animales ajustados							8	37.59	4.70

Pasemos a realizar todas las posibles pruebas de hipótesis.

Para probar el supuesto de que las fuentes de variación no influyen en la covariable, probemos primero los insecticidas, la F_o de esta ANOVA es

$$F_o = \frac{\frac{413.4141}{3}}{\frac{498.50}{8}} = 2.211ns$$

$$F_{(3,8;\,0.05)} = 4.07$$

Y para los animales la F_o sería

$$F_o = \frac{\frac{498.50}{8}}{\frac{1878.00}{36}} = 1.194ns$$

$$F_{(8,36;\,0.05)} = 2.21$$

Como era de esperar, se cumple el supuesto de no efectos de los factores sobre la medida inicial.

Si se quiere probar la hipótesis de igualdad de efectos de los diferentes insecticidas sobre la cantidad final de parásitos, sin ajustar para la regresión, es decir, como si, desconociendo la cantidad inicial de parásitos, se tratara de un ANOVA, está claro que la prueba F_o sería,

$$F_o = \frac{\frac{530.9165}{3}}{\frac{371.00}{8}} = 3.816ns$$

$$F_{(3,8;\,0.05)} = 4.07$$

Se concluiría que los insecticidas no son efectivos.

Si se quieren probar los animales, en las mismas condiciones, la F_o sería

$$F_o = \frac{\frac{371.00}{8}}{\frac{760.00}{36}} = 2.197ns$$

$$F_{(8,36;\,0'05)} = 2.21$$

se concluiría que no existe variabilidad entre los animales.

Si la prueba que se desea hacer es la de igualdad de efectos de los insecticidas pero ajustándolos para la regresión con la covariable, es decir, la prueba F de las medias ajustadas, a semejanza de la anterior, la F_o sería

$$F_o = \frac{\frac{92.5132}{3}}{\frac{37.5915}{8}} = 6.563^*$$

$$F_{(3,8;\,0.05)} = 4.07$$

La F ha pasado de vales 3.816*ns* a valer 6.563*, ahora si se concluye que los efectos de los insecticidas son significativos, esto es, que hay uno/s mejor que los otros. En la prueba anterior no se detectaba esta variabilidad como consecuencia de que la variabilidad de la medida inicial, y su regresión con la medida final, solapaba la variabilidad de la medida final.

Si se quieren probar los animales ajustados, la pruebas sería

$$F_o = \frac{\frac{37.5915}{8}}{\frac{58.8518}{35}} = 2.794^*$$

$$F_{(8,35;\,0.05)} = 2.22$$

Si existe variabilidad entre los animales. En la prueba anterior no se detectaba esta variabilidad como consecuencia de que la variabilidad de la medida inicial, y su regresión con la medida final, solapaba la variabilidad de la medida final.

Si la prueba que se desea hacer es la de $\beta=0$, siendo β la del modelo del análisis de covarianza, la prueba F sería

$$F_o = \frac{\frac{1147.50^2}{1878.00}}{\frac{58.8518}{35}} = 416.98^{***}$$

$$F_{(1,35;\,0.001)} = 12.90$$

Lógicamente, un gran porcentaje de la variabilidad de la cantidad de insectos al final de la experiencia es función lineal de la cantidad de insectos al principio de la experiencia.

El cálculo de las medias ajustada (o medias minimocuadráticas) de los insecticidas es

$$\hat{Y}_{I_1} = 14.25 - 0.611 \times (5.642 - 52.04) = 11.57$$
$$\hat{Y}_{I_2} = 9.42 - 0.611 \times (52.08 - 52.04) = 9.39$$
$$\hat{Y}_{I_3} = 8.58 - 0.611 \times (51.5 - 52.04) = 8.91$$
$$\hat{Y}_{I_4} = 4.92 - 0.611 \times (48.17 - 52.04) = 7.28$$

El *error típico de la media ajustada* del primer insecticida, por ejemplo, es

$$S_{\hat{Y}_{I_1}} = \sqrt{1.6815\left[\frac{1}{12}+\frac{(56.4167-52.0417)^2}{1878.0}\right]} = 0.3966$$

La diferencia entre las medias ajustadas del primer y segundo insecticida es

Diferencia entre el 1° y 2° insecticida

$$\hat{Y}_{I_1} - \hat{Y}_{I_2} = 14.25 - 9.4167 - 0.611 \times (56.4167 - 52.083) = 2.1855$$

El *error típico de la diferencia de las medias ajustadas* del primero y segundo insecticida es

$$S_{\hat{Y}_{I_1}\text{-}\hat{Y}_{I_2}} = \sqrt{1.6815\left[\frac{2}{12}+\frac{(56.4167-52.0833)^2}{1878.0}\right]} = 0.5450$$

El cálculo de las medias ajustada (o medias minimocuadráticas) de los animales del primer insecticida, por ejemplo, es

$$\hat{Y}_{11.} = 13.75 - 0.6110 \times (56.0 - 56.416) = 14.005$$
$$\hat{Y}_{12.} = 18.00 - 0.6110 \times (59.5 - 56.416) = 16.116$$
$$\hat{Y}_{13.} = 11.00 - 0.6110 \times (53.75 - 56.416) = 12.629$$

El *error típico de la media ajustada* del primer animal del primer insecticida, por ejemplo, es

$$S_{\hat{Y}_{11.}} = \sqrt{1.6815\left[\frac{1}{4}+\frac{(56-56.417)^2}{1878.0}\right]} = 0.4205$$

La diferencia entre las medias ajustadas del primero y segundo animal del primer insecticida, por ejemplo, es

$$\hat{Y}_{11.} - \hat{Y}_{12.} = 13.75 - 18 - 0.6110 \times (56 - 59.5) = -2.111$$

Y el *error típico de la diferencia de medias ajustadas* del primer y segundo animal dentro del primer insecticida, es

$$S_{\hat{Y}_{11.}-\hat{Y}_{12.}} = \sqrt{1.6815\left[\frac{2}{4}+\frac{(56-59.5)^2}{1878}\right]} = 0.9229$$

La prueba *t* para contrastar las media ajustada del primer y segundo insecticida, por ejemplo, es

$$t = \frac{2.1855}{0.5450} = 4.010^{***}$$

$$t_{(35;\ 0.001/2)} = 3.5915$$

Y la prueba t para contrastar las media ajustada del primer y segundo animal dentro del primer insecticida, por ejemplo, es

$$t = \frac{-2.1114}{0.9229} = 2.288^{*}$$

$$t_{(35;\ 0.05/2)} = 2.0301$$

Para probar la efectividad de la covarianza como medio de controlar el error, se hace comparando la varianza de las medias de insecticidas y de animales antes y después del ajuste.

Los cuadrados medios de los dos términos de contraste, es decir, el del factor *animal dentro de insecticida* y el del *error*, antes del ajuste son

$$CM_{\text{Error}} = \frac{760.0}{36} = 21.1111$$

$$CM_{A(I)} = \frac{371.0}{8} = 46.375$$

El cuadrado medio efectivo de los término de contraste después del ajuste para X, para el factor principal *Insecticida* y para el factor jerárquico *Animal* son, respectivamente

Para Insecticida

$$S^{2}_{Y.X(I)} = 1.6815\left[1+\frac{413.4141}{3\times 498.5}\right] = 2.1463$$

Para Animal

$$S^{2}_{Y.X(A_{(I)})} = 1.6815\left[1+\frac{498.5}{8\times 1878}\right] = 1.7373$$

Un estimador de la precisión relativa es la razón del cuadrado medio del término de contraste sin ajustar por el cuadrado medio ajustado, multiplicado por 100 para expresarlo en porcentajes.

Para Insecticida

$$\frac{46.375}{2.1463}100 = 260.1\%$$

Para Animal

$$\frac{21.1111}{1.7373}100 = 121.5\%$$

Se comprueba un gran aumento en la precisión en el contraste de ambos factores.

Archivo del programa *SAS* (C17-4.SAS).-

```
title 'ancova de Jerarquico';
option ls=75 ps=60;
data cov_jera;
infile 'c17-4.dat';
  do insecti = 1 to 4;
   do animal = 1 to 3;
    do zona = 1 to 4;
     input x y @@;output;
    end;
   end;
  end;
title 'ANOVA de X';
proc glm;
 class insecti animal;
 model x = insecti animal(insecti) ;
   random animal(insecti) / test;
run;
title 'ANOVA de Y';
proc glm;
 class insecti animal;
 model y = insecti animal(insecti) ;
   random animal(insecti) / test;
run;
title 'ANCOVA';
proc glm;
 class insecti animal;
 model y = insecti animal(insecti) x / solution;
   random animal(insecti) / test;
   lsmeans insecti animal(insecti) / stderr tdiff;
run;
```

Archivo de datos (C17-4.DAT).-

```
 52   10    66   21    54   13    52   11
 64   20    53   11    56   19    65   22
 51    8    64   18    48    8    52   10
 58   13    42    5    59   14    48    7
 54   11    63   17    55   13    66   19
 53    9    38    1    46    3    43    1
 49    6    43    4    60   14    56   12
 56   12    57   13    44    2    47    4
 42    4    43    3    62   15    59   14
 55    8    41    1    52    7    44    2
 41    1    55    9    45    2    56   10
 52    6    37    1    51    7    49    5
```

Archivo de resultados (C17-4.LST).-

```
                          ANOVA de X
                  General Linear Models Procedure
Dependent Variable: X
                                  Sum of          Mean
Source                     DF       Squares        Square  F Value     Pr > F
Model                      11    911.916667     82.901515     1.59     0.1440
Error                      36   1878.000000     52.166667
Corrected Total            47   2789.916667
```

R-Square	C.V.	Root MSE	X Mean
0.326862	13.87859	7.22265	52.0417

Source	DF	Type I SS	Mean Square	F Value	Pr > F
INSECTI	3	413.416667	137.805556	2.64	0.0641
ANIMAL(INSECTI)	8	498.500000	62.312500	1.19	0.3296

Source	DF	Type III SS	Mean Square	F Value	Pr > F
INSECTI	3	413.416667	137.805556	2.64	0.0641
ANIMAL(INSECTI)	8	498.500000	62.312500	1.19	0.3296

Source	Type III Expected Mean Square
INSECTI	Var(Error) + 4 Var(ANIMAL(INSECTI)) + Q(INSECTI)
ANIMAL(INSECTI)	Var(Error) + 4 Var(ANIMAL(INSECTI))

Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: X

Source: INSECTI
Error: MS(ANIMAL(INSECTI))

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
3	137.80555556	8	62.3125	2.2115	0.1644

Source: ANIMAL(INSECTI)
Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
8	62.3125	36	52.166666667	1.1945	0.3296

ANOVA de Y
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	901.916667	81.992424	3.88	0.0010
Error	36	760.000000	21.111111		
Corrected Total	47	1661.916667			

R-Square	C.V.	Root MSE	Y Mean
0.542697	49.44950	4.59468	9.29167

Source	DF	Type I SS	Mean Square	F Value	Pr > F
INSECTI	3	530.916667	176.972222	8.38	0.0002
ANIMAL(INSECTI)	8	371.000000	46.375000	2.20	0.0511

Source	DF	Type III SS	Mean Square	F Value	Pr > F
INSECTI	3	530.916667	176.972222	8.38	0.0002
ANIMAL(INSECTI)	8	371.000000	46.375000	2.20	0.0511

Source	Type III Expected Mean Square
INSECTI	Var(Error) + 4 Var(ANIMAL(INSECTI)) + Q(INSECTI)
ANIMAL(INSECTI)	Var(Error) + 4 Var(ANIMAL(INSECTI))

Tests of Hypotheses for Mixed Model Analysis of Variance

Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: Y

Source: INSECTI
Error: MS(ANIMAL(INSECTI))

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
3	176.97222222	8	46.375	3.8161	0.0576

Source: ANIMAL(INSECTI)
Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
8	46.375	36	21.111111111	2.1967	0.0511

ANCOVA

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	1603.06483	133.58874	79.45	0.0001
Error	35	58.85184	1.68148		
Corrected Total	47	1661.91667			

R-Square	C.V.	Root MSE	Y Mean
0.964588	13.95572	1.29672	9.29167

Source	DF	Type I SS	Mean Square	F Value	Pr > F
INSECTI	3	530.916667	176.972222	105.25	0.0001
ANIMAL(INSECTI)	8	371.000000	46.375000	27.58	0.0001
X	1	701.148163	701.148163	416.98	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
INSECTI	3	92.513603	30.837868	18.34	0.0001
ANIMAL(INSECTI)	8	37.591462	4.698933	2.79	0.0167
X	1	701.148163	701.148163	416.98	0.0001

Parameter			Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT			-24.12080671 B	-15.51	0.0001	1.55541360
INSECTI	1		2.27835463 B	2.43	0.0203	0.93732032
	2		0.12480032 B	0.14	0.8928	0.91938745
	3		1.65315495 B	1.79	0.0828	0.92569591
	4		0.00000000 B	.	.	.
ANIMAL(INSECTI)	1	1	1.37519968 B	1.50	0.1437	0.91938745
	2	1	3.48662141 B	3.74	0.0007	0.93292191
	3	1	0.00000000 B	.	.	.
	1	2	2.12559904 B	2.26	0.0299	0.93890109
	2	2	2.64017572 B	2.60	0.0135	1.01439115
	3	2	0.00000000 B	.	.	.
	1	3	-0.30551118 B	-0.33	0.7410	0.91704109
	2	3	-0.94448882 B	-1.03	0.3101	0.91704109
	3	3	0.00000000 B	.	.	.
	1	4	-0.70826677 B	-0.77	0.4452	0.91719364
	2	4	-0.47204473 B	-0.51	0.6107	0.91886994
	3	4	0.00000000 B	.	.	.
X			0.61102236	20.42	0.0001	0.02992252

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

Source	Type III Expected Mean Square
INSECTI	Var(Error) + 3.7594 Var(ANIMAL(INSECTI)) + Q(INSECTI)
ANIMAL(INSECTI)	Var(Error) + 3.8951 Var(ANIMAL(INSECTI))
X	Var(Error) + Q(X)

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

Source: INSECTI

Error: 0.9652*MS(ANIMAL(INSECTI)) + 0.0348*MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
3	30.837867602	8.21	4.5938260202	6.7129	0.0135

Source: ANIMAL(INSECTI)

Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
8	4.6989327385	35	1.6814810589	2.7945	0.0167

Source: X

Error: MS(Error)

DF	Type III MS	Denominator DF	Denominator MS	F Value	Pr > F
1	701.14816294	35	1.6814810589	416.9825	0.0001

Least Squares Means

INSECTI	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number
1	11.5767772	0.3965616	0.0001	1
2	9.3912074	0.3743327	0.0001	2
3	8.9143038	0.3746814	0.0001	3
4	7.2843783	0.3918772	0.0001	4

T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|

i/j	1	2	3	4
1	.	4.009986	4.845742	7.348586
		0.0003	0.0001	0.0001
2	-4.00999	.	0.900377	3.885698
	0.0003		0.3741	0.0004
3	-4.84574	-0.90038	.	3.025677
	0.0001	0.3741		0.0046
4	-7.34859	-3.8857	-3.02568	.
	0.0001	0.0004	0.0046	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

ANIMAL	INSECTI	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number
1	1	11.3313698	0.6590896	0.0001	1
2	1	13.4427915	0.6856939	0.0001	2
3	1	9.9561701	0.6503716	0.0001	3
1	2	9.9282149	0.6484184	0.0001	4
2	2	10.4427915	0.6856939	0.0001	5
3	2	7.8026158	0.6817379	0.0001	6
1	3	9.0254593	0.6483609	0.0001	7
2	3	8.3864816	0.6491085	0.0001	8
3	3	9.3309704	0.6485622	0.0001	9
1	4	6.9695487	0.6595423	0.0001	10
2	4	7.2057708	0.6537187	0.0001	11
3	4	7.6778155	0.6640239	0.0001	12

T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|

i/j	1	2	3	4	5	6	7
1	.	-2.28786	1.495778	1.515784	0.962831	3.622182	2.493691
		0.0283	0.1437	0.1386	0.3422	0.0009	0.0175
2	2.28786	.	3.737313	3.716024	3.271826	5.560159	4.67946
	0.0283		0.0007	0.0007	0.0024	0.0001	0.0001
3	-1.49578	-3.73731	.	0.030424	-0.52161	2.258418	1.01339
	0.1437	0.0007		0.9759	0.6052	0.0303	0.3178
4	-1.51578	-3.71602	-0.03042	.	-0.54407	2.263922	0.98452
	0.1386	0.0007	0.9759		0.5898	0.0299	0.3316
5	-0.96283	-3.27183	0.52161	0.544071	.	2.60272	1.501438
	0.3422	0.0024	0.6052	0.5898		0.0135	0.1422
6	-3.62218	-5.56016	-2.25842	-2.26392	-2.60272	.	-1.30015
	0.0009	0.0001	0.0303	0.0299	0.0135		0.2020
7	-2.49369	-4.67946	-1.01339	-0.98452	-1.50144	1.300152	.
	0.0175	0.0001	0.3178	0.3316	0.1422	0.2020	
8	-3.1698	-5.31381	-1.70506	-1.68092	-2.16103	0.624903	-0.6965
	0.0032	0.0001	0.0970	0.1017	0.0376	0.5361	0.4907
9	-2.1585	-4.33896	-0.68002	-0.65134	-1.17324	1.630557	0.333149
	0.0378	0.0001	0.5010	0.5191	0.2486	0.1120	0.7410
10	-4.60277	-6.60965	-3.20136	-3.20285	-3.54643	-0.90423	-2.22333
	0.0001	0.0001	0.0029	0.0029	0.0011	0.3721	0.0327
11	-4.39407	-6.45083	-2.96778	-2.95929	-3.34799	-0.64475	-1.97662
	0.0001	0.0001	0.0054	0.0055	0.0020	0.5233	0.0560
12	-3.83146	-5.83812	-2.43071	-2.42826	-2.80006	-0.13574	-1.45241
	0.0005	0.0001	0.0203	0.0205	0.0083	0.8928	0.1553

T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|

i/j	8	9	10	11	12

```
   1   3.169802   2.158503   4.602769   4.394072   3.831458
         0.0032     0.0378     0.0001     0.0001     0.0005
   2   5.313812   4.338959   6.609647   6.450832   5.838119
         0.0001     0.0001     0.0001     0.0001     0.0001
   3   1.705064   0.680018   3.201363    2.96778   2.430711
         0.0970     0.5010     0.0029     0.0054     0.0203
   4   1.680924   0.651338   3.202853   2.959289   2.428262
         0.1017     0.5191     0.0029     0.0055     0.0205
   5   2.161032   1.173238   3.546431   3.347989   2.800056
         0.0376     0.2486     0.0011     0.0020     0.0083
   6    -0.6249   -1.63056   0.904227   0.644753   0.135743
         0.5361     0.1120     0.3721     0.5233     0.8928
   7   0.696504   -0.33315   2.223332   1.976625   1.452406
         0.4907     0.7410     0.0327     0.0560     0.1553
   8         .   -1.02993   1.537967   1.285599   0.767154
                    0.3101     0.1330     0.2070     0.4481
   9   1.029931          .   2.558751   2.311539   1.785851
         0.3101               0.0150     0.0268     0.0828
  10   -1.53797   -2.55875          .   -0.25741   -0.77221
         0.1330     0.0150               0.7984     0.4452
  11    -1.2856   -2.31154   0.257412          .   -0.51372
         0.2070     0.0268     0.7984               0.6107
  12   -0.76715   -1.78585   0.772211   0.513723          .
         0.4481     0.0828     0.4452     0.6107
NOTE: To ensure overall protection level, only probabilities associated
      with pre-planned comparisons should be used.
```

Estos son los mismos resultados que los obtenidos manualmente, pero con un menor error de redondeo.

La suma de cuadrados que hay que mirar es la **tipo III**. La suma de cuadrados **tipo I** debida a los factores (insecticida y animal) es la del ANOVA para la variable *Y*. Si, en el modelo del programa SAS, se pone primero la covariable y después los factores, esto es, **model y = x insecti animal(insecti) / solution;** se hubiera obtenido la misma suma de cuadrados debida al último factor del modelo (en este caso *la interacción*), en el tipo I y en el tipo III, pero lo suma de cuadrados tipo I debida a la regresión, sería la de la regresión sin considerar las correcciones de los factores, esto es, la de la regresión de los 48 pares de valores, como si los insecticidas y los animales no existieran

Las pruebas *F* que hay que mirar son las que están bajo el epígrafe ***Tests of Hypotheses for Mixel Model Análisis of Variance*** .

Generalización del método para la realización del análisis de covarianza.-

En el capítulo 9 se estudió la *Generalización del método para el cálculo de las sumas de cuadrados y de los grados de libertad* tanto para cualquier modelo factorial como jerárquico como complejo, así como el *Cálculo de las estimas de los cuadrados medios por el método rápido* con objeto de saber cuales son los términos de contraste para las diferentes fuentes de variación. Para el análisis de covarianza hay que seguir esas mismas generalizaciones, teniendo en cuenta, tal como se ha repetido a lo largo de todo este capítulo, que en el análisis de covarianza se tienen dos variables, en lugar de la única variable que se tiene en el análisis de varianza, por lo que hay que descomponer dos sumas de cuadrados, una para cada variable (*X* e *Y*), y la suma de productos de ambas variables, siguiendo las reglas vistas en el Capítulo 9.

Una vez realizadas la descomposición de las sumas de cuadrados, de la suma de

productos y lo grados de libertad hay que calcular la fracción de la suma de cuadrados total que es debida a la regresión en la fila del *error*, así como las fracciones de la suma de cuadrados total debidas a todas las fuentes de variación que se quieren probar, sumándole previamente el término de *error*. Una vez calculadas las sumas de cuadrados debidas a las regresiones de cada fuente de variación más el *error*, se le resta esta suma de cuadrados a la suma de cuadrados de la variable que bajo estudio (la *Y*) en la línea de la fuente de variación correspondiente. Esto dará la sumas de cuadrados debida a cada fuente de variación más *error* ajustadas para la regresión, por lo que si se le quita la suma de cuadrados ajustadas del *error* queda la sumas de cuadrados ajustadas de cada fuente de variación, las cuales se dividirán por sus grados de libertad para calcular los cuadrados medios y estos se dividirán por los cuadrados medios de los términos de contraste que indique las estimas de cada cuadrado medio.

Partición de la covarianza.-

El término de análisis de la *covarianza* implica un uso al que generalmente no se le da importancia, éste es, la partición efectuada en la columna de las sumas de productos, lo que posibilita el cálculo de los coeficientes de regresión y correlación de cada fuente de variación del modelo. En las tablas siguientes se presentan éstos, realizados con los ejemplos de los modelos estudiados en este capítulo

FV	*gl*	SC_X	SP_{XY}	SC_Y	*b*	*r*
Trata	2	0.867	0.107	0.355	0.1234	0.1929
Error	27	104.34	8.197	1.899	0.0786	0.5823
total	29	103.47	8.304	2.254	0.0802	0.5437

FV	*gl*	SC_X	SP_{XY}	SC_Y	*b*	*r*
Dietas	2	0.8666	0.1073	0.3546	0.1238	0.1936
Jaulas	9	57.8733	3.491	0.4436	0.0603	0.6890
Error	18	46.4667	4.706	1.4554	0.1013	0.5722
total	29	105.207	8.3043	2.2536	0.0789	0.5393

FV	*gl*	SC_X	SP_{XY}	SC_Y	*b*	*r*
Dietas	2	0.8643	0.1072	0.3546	0.1240	0.1936
Raza	1	4.0332	-0.6052	0.0907	-0.1501	-0.9999
DiexRaz	2	5.2695	0.7126	0.1014	0.1352	0.9768
Error	24	95.0400	8.0901	1.7120	0.0851	0.6342
total	29	105.207	8.3047	2.2587	0.0789	0.5387

FV	gl	SC_X	SP_{XY}	SC_Y	b	r
Insecticidas	3	413.41	469.4	530.92	1.1355	0.9998
Animales(insec)	8	498.50	420.5	371.00	0.8435	0.9778
Error	36	1878.91	1147.5	760.00	0.6110	0.9605
total	47	2789.91	2036.4	1661.92	0.7299	0.9457

En el caso del modelo de un solo factor, si la *F* de tratamientos ajustados es no significativa indica que tanto los *tratamientos* como el *error* proporcionan estimaciones independientes de un mismo coeficiente de regresión. La justificación es la siguiente.

La suma de cuadrados de los tratamientos ajustados pueden considerarse como la suma de cuadrados de las desviaciones promedio de los tratamientos respecto de la recta de regresión común. Si se calcula la suma de cuadrados de las desviaciones de los tratamientos con respecto a su recta de regresión, se tendrá *t*-2 grados de libertad. Si se restan ambas sumas de cuadrados, si ambas regresiones son homogéneas, deberán dar un resultado no significativo.

Es útil el procedimiento de comparar la regresión de los *tratamientos* con la regresión del *error*, en el mismo estudio, porque muchas veces, la *Y* y la *X* son mediciones hechas bien en humanos o bien en animales domésticos o animales de experimentación de alto valor, lo que condiciona que el número de individuos disponibles sea limitado, pero sí se puede hacer varias mediciones de *Y* y *X* en cada individuo. En este caso, la regresión entre individuos puede ser de gran interés. El objetivo de este apartado es ver si las regresiones de *individuos* y *error* estiman lo mismo. Si fuese así, pueden combinarse ambas regresiones para dar una mejor estima de la relación *Entre Individuos*.

El modelo más sencillo que se puede usar, tal como se vio al principio del capítulo, es

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij}$$

donde *i* representan los diferentes tratamientos o niveles del factor, es decir, los diferentes individuos.

En este caso, es una recta de regresión que tiene validez para todos los individuos, por lo que la estima de α y de β se obtiene tratando los datos como una sola muestra, es decir, estimándolos a partir de la fila de los *totales* de la descomposición de las sumas de cuadrados y de la suma de productos.

Las consecuencias importantes de este modelo son dos: 1) las líneas de regresión de los tratamientos o *individuos* y la del *error* darán estimas independientes de *b*; a las que se pueden llamar b_T y *b*, respectivamente; 2) el cuadrado medios residuales de la fila de los tratamientos o *individuos* ($S^2_{T:Y.X}$) y el cuadrado medio residual de la fila del *error* ($S^2_{Y.X}$) son, ambos, estimas no sesgadas de σ^2, la varianza del error (ε_{ij})

Para probar si ambas regresiones son la misma, se compara b_T con *b*, y $S^2_{T:Y.X}$ con

$S^2_{Y.X}$. Puede ocurrir que b_T y b concuerden bien, pero que $S^2_{T:Y.X}$ sea mayor que $S^2_{Y.X}$. Una explicación a esto es que todos los valores de la variable Y_{ij} para un individuo estén afectados por un componente adicional de variación d_i, independiente de las ε_{ij}. Este modelo se formaliza así

$$Y_{ij} = \alpha + \beta X_{ij} + d_i + \varepsilon_{ij}$$

Si los individuos constituyen una muestra aleatoria de alguna población, las d_i se consideran generalmente como variables aleatorias de un individuo a otro, con media cero y varianza σ^2_T. De acuerdo con este modelo, b_T y b siguen siendo estimas no sesgadas de β, pero al haber n pares de observaciones por individuo, $S^2_{T:Y.X}$ es una estima no sesgada de $\sigma^2_{T:Y.X} = (\sigma^2 + n\sigma^2_T)$, en tanto que $S^2_{Y.X}$ estima σ^2. Como el método de comparar b y b_T y la mejor manera de combinarlas depende de si el componente d_i está presente, es conveniente comparar primero $S^2_{T:Y.X}$ y $S^2_{Y.X}$ con una prueba F.

Para continuar la explicación es conveniente hacerlo con un ejemplo

Ejemplo.-

Se tiene la abundancia de parásitos en cuatro zonas del cuerpo antes (X) y después (Y) de un tratamiento con un desparasitador externo, aplicado a 10 caballos. Así se tiene n=4 t=10

		Individuo									
		1	2	3	4	5	6	7	8	9	10
Zona 1	*X*	13	6	7	12	17	6	12	10	11	8
	Y	7	5	5	9	12	5	10	5	9	5
Zona 2	*X*	12	9	5	11	12	6	7	8	9	6
	Y	4	4	4	5	9	4	4	4	5	4
Zona 3	*X*	11	7	7	13	18	5	8	6	10	11
	Y	5	4	4	7	11	4	7	4	7	4
Zona 4	*X*	11	9	7	14	12	6	9	11	13	6
	Y	6	5	5	9	8	5	13	5	9	5

Este procedimiento es útil si no se tienen más individuos en los que se puedan medir estas variables, en el caso de que se dispusiera de más individuos se puede estimar con precisión la regresión *entre individuos* de la manera vista en el capítulo 11.

Los resultados de los cálculos son

						ajuste debido a b	
FV	*gl*	SC_X	SP_{XY}	SC_Y	*b*	*gl*	*SC*	*CM*
Individuo	9	292.73	165.85	152.1	0.5666=b_T	8	58.1355	7.2669
Error	30	103.25	49.00	87.0	0.4746=*b*	29	63.7458	2.1981
total	39	395.98	214.85	239.1				

Es decir, se calcula por separado el ajuste de la suma de cuadrados debida a *b* tanto para la fuente de variación *Individuoe* como para la fuente de variación *error.*

La prueba para contrastar si $S^2_{T:Y.X}$ es igual a $S^2_{Y.X}$ es

$$F_o = \frac{S^2_{T:Y.X}}{S^2_{Y.X}} = \frac{7.2669}{2.1981} = 3.31^{**}$$

$$F_{(8,29;\,0.01)} = 3.17$$

se concluye que $\sigma^2_{T:Y.X}$ es mayor que σ^2 por lo que hay que tener en cuenta el modelo $Y_{ij} = \alpha + \beta X_{ij} + d_i + \varepsilon_{ij}$, máxime cuando también es significativo el factor *individuo* para ambas variables, *X* e *Y*.

Para comparar b_T y *b*, bajo este modelo, hay que tener en cuenta que las varianzas estimadas de b_T y *b* son, respectivamente

$$\sigma_{b_T} = \frac{S_{T:Y.X}}{T_{XX}} = \frac{7.2669}{292.73} = 0.0248$$

$$\sigma_b = \frac{S_{Y.X}}{E_{XX}} = \frac{2.1981}{103.25} = 0.0213$$

La prueba *t* sería, por tanto

$$t_o = \frac{b_T - b}{\sqrt{\frac{S^2_{T:Y.X}}{T_{XX}} + \frac{S_{Y.X}}{E_{XX}}}} = \frac{0.4666-0.4746}{\sqrt{\frac{7.2669}{2.9273} + \frac{2.1981}{1.0325}}} = 0.4284 ns$$

Esta *t* no se distribuye exactamente como la *t* de *Student* pero puede aproximarse a la distribución *t* si se realiza la transformación de *Cochran* y *Cox*, tal como se estudió en el capítulo 5. Como $S^2_{T:Y.X}$ tiene *gl*=8 y $S^2_{Y.X}$ tiene *gl*=29, al nivel de significación de α=0.05 la t_T=2.3060 y la t_e=2.0452. Como la t_o es menor que las dos *t* de las tablas, es no significativa, es decir, que los dos coeficientes de regresión son iguales. Si se hubiera tenido que calcular la *t* ponderada de estos dos valores, la ponderación se habría hecho con $S^2_{T:Y.X}/T_{XX}$ y $S^2_{Y.X}/E_{XX}$, es decir

$$t_p = \frac{\left[t_{(0.05;8)}\frac{S^2_{T:Y.X}}{T_{XX}}\right]+\left[t_{(0.05;29)}\frac{S^2_{Y.X}}{E_{XX}}\right]}{\frac{S^2_{T:Y.X}}{T_{XX}}+\frac{S^2_{Y.X}}{E_{XX}}} = 2.185$$

Puesto que no existe diferencias entre ambas *b*, se puede calcular la estima combinada de β a partir de b_T y *b*. Al combinar dos estimas independientes, que son de precisión desigual,la regla general, lo mismo que se ha hecho con la *t* es ponderar cada estima por la inversa de su varianza. En el ejemplo, la estima de la varianza de b_T y de *b* son las dadas más arriba, por lo que los pesos de ponderación son

$$w_T = \frac{1}{\sigma_{b_T}} = \frac{T_{XX}}{S_{T:Y.X}} = \frac{292.73}{7.2669} = 40.2826$$

$$w = \frac{1}{\sigma_b} = \frac{E_{XX}}{S_{Y.X}} = \frac{103.25}{2.1981} = 46.9724$$

Y, por tanto, la estima de β es

$$b_p = \frac{[b_T w_T]+[bw]}{w_T+w} = 0.5171$$

Si $W=w_T+w=87.255$, el error típico de β es

$$S_{Y.X} = \frac{1}{\sqrt{W}}\sqrt{1+\frac{4x w_T xw}{W^2}\times\frac{gl_T+gl_E}{gl_T x gl_E}} = 0.1152$$

Si el primer contraste que se hizo para probar si $\sigma^2_T=\sigma^2$ hubiese salido no significativo, es decir, que ambas varianzas son iguales, se hubiera podido elaborar una estima global de σ^2 a partir de $S^2_{T:Y.X}$ y $S^2_{Y.X}$, esta estima habría sido

$$\hat{\sigma}^2 = \frac{SC_T+SC_{Error}}{gl_T+gl_{Error}} = \frac{121.8813}{37} = 3.2941$$

La varianza de la estima de (b_T-b) hubiera sido

$$S^2_{b_t-b} = \frac{\hat{\sigma}^2}{T_{XX}}+\frac{\hat{\sigma}^2}{E_{XX}} = \hat{\sigma}^2\frac{T_{XX}+E_{XX}}{T_{XX}E_{XX}} = 0.0432$$

Y la prueba *t* para el contraste de diferencia de coeficientes de regresión hubiera sido

$$t_o = \frac{b_T - b}{S_{b_t - b}} = \frac{0.092}{\sqrt{0.0432}} = 0.4428 ns$$

$$t_{(37;\,0.05/2)} = 2.0262$$

La estima total de β hubiera sido simplemente S_{XY}/S_{XX}, es decir, en este ejemplo, tomando los valores de la línea de los totales

$$b = \frac{214.85}{395.98} = 0.5426$$

con un error típico de

$$S_{Y.X} = \sqrt{\frac{\hat{\sigma}^2}{T_{XX} + E_{XX}}} = 0.0912$$

Prueba de homogeneidad de coeficientes de regresión.-

A menudo la relación funcional entre Y y X se estudia en muestras sacadas por investigadores diferentes, o bien, en ambientes diferentes o en épocas diferentes. En casos como estos, en el que el experimentador maneja dos o más líneas de regresión a partir de datos análogos, éste, naturalmente, desea saber si las relaciones funcionales descritas por las ecuaciones de regresión son las mismas o diferentes. Por ejemplo, se puede haber establecido la regresión de la concentración sanguínea de colesterol sobre la edad en una muestra de individuos y se puede desear, ahora, comparar esta ecuación de regresión con la de otra u otras muestras sometidas a una dieta diferente. Por lo que lo que se desea es contrastar la *homogeneidad* de los b, es decir, determinar si pueden considerarse o no estimaciones de un β común. El modelo básico de tal tipo de prueba es el del análisis de covarianza para el diseño completamente aleatorio, es decir

$$Y_{ij} = \mu + T_i + \beta(X_{ij} - X_{..}) + \varepsilon_{ij}$$

Existirán, por tanto, t muestras representando los *tratamientos.*

Dentro de cada una de estas t muestras se calculan las sumas de cuadrados para las dos variables

$$SC_{(X)} \text{ del tratamiento } i = E_{X_i X_i}$$
$$SC_{(Y)} \text{ del tratamiento } i = E_{Y_i Y_i}$$

la suma de productos

$$SP \text{ del tratamiento } i = E_{X_i Y_i}$$

la regresión

$$b_i = \frac{E_{X_i Y_i}}{E_{X_i X_i}}$$

la suma de cuadrados debida a la regresión

$$SC_{(\text{Regresión}_i)} = \frac{E^2_{X_i Y_i}}{E_{X_i X_i}}$$

y la suma de cuadrados residual

$$SC_{(\text{Error}_i)} = E_{Y_i Y_i} - \frac{E^2_{X_i Y_i}}{E_{X_i X_i}}$$

La suma de las t sumas de cuadrados residuales

$$SC_{(\text{Error})} = \sum_i SC_{(\text{Error}_i)} = \sum_i E_{Y_i Y_i} - \frac{E^2_{X_i Y_1}}{E_{X_i X_i}}$$

$$gl = \sum_i n_i - 2t = N - 2t$$

hará de termino de error en la prueba F de contraste de homogeneidad de regresiones.

El numerador de esta prueba es la suma de cuadrados debida a los diferentes coeficientes de regresión, esta es

$$SC_{(\text{Regresión})} = \sum_i SC_{(\text{Regresión}_i)} - \frac{\left[\sum_i E_{X_i Y_i}\right]^2}{\sum_i E_{X_i X_i}} =$$

$$= \sum_i \frac{E^2_{X_i Y_i}}{E_{X_i X_i}} - \frac{\left[\sum_i E_{X_i Y_i}\right]^2}{\sum_i E_{X_i X_i}}$$

$$gl = t - 1$$

Esto se puede resumir en las siguientes tablas

Muestra	*gl*	$SC_{(X)}$	*SP*	$SC_{(Y)}$	*gl*	$SC_{(regresión)}$	*gl*	$SC_{(error)}$
1	n_1-1	$E_{X_1X_1}$	$E_{X_1Y_1}$	$E_{Y_1Y_1}$	1	$\frac{E^2_{X_1Y_1}}{E_{X_1X_1}}$	n_1-2	$E_{Y_1Y_1}-\frac{E^2_{X_1Y_1}}{E_{X_1X_1}}$
2	n_2-1	$E_{X_2X_2}$	$E_{X_2Y_2}$	$E_{Y_2Y_2}$	1	$\frac{E^2_{X_2Y_2}}{E_{X_2X_2}}$	n_2-2	$E_{Y_2Y_2}-\frac{E^2_{X_2Y_2}}{E_{X_2X_2}}$
...	...	...	...	...	1	...	...	...
t	n_t-1	$E_{X_tX_t}$	$E_{X_tY_t}$	$E_{Y_tY_t}$	1	$\frac{E^2_{X_tY_t}}{E_{X_tX_t}}$	n_t-2	$E_{Y_tY_t}-\frac{E^2_{X_tY_t}}{E_{X_tX_t}}$
Σ	*N-t*	$\sum_i E_{X_iX_i}$	$\sum_i E_{X_iY_i}$	$\sum_i E_{Y_iY}$	*t*	$\sum_i \frac{E^2_{X_iY_i}}{E_{X_iX_i}}$	*N-2t*	$\sum_i E_{Y_iY_i}-\frac{E^2_{X_iY_i}}{E_{X_iX_i}}$

FV	*gl*	*SC*	F_o
Regresiones	*t*-1	$\sum_i \frac{E^2_{X_iY_i}}{E_{X_iX_i}}-\frac{\left[\sum_i E_{X_iY_i}\right]^2}{\sum_i E_{X_iX_i}}$	$\frac{\sum_i \frac{E^2_{X_iY_i}}{E_{X_iX_i}}-\frac{\left[\sum_i E_{X_iY_i}\right]^2}{\sum_i E_{X_iX_i}}}{t-1}$ / $\frac{\sum_i E_{Y_iY_i}-\frac{E^2_{X_iY_i}}{E_{X_iX_i}}}{N-2t}$
Residuo	*N-2t*	$\sum_i E_{Y_iY_i}-\frac{E^2_{X_iY_1}}{E_{X_iX_i}}$	
Total	*N-t*-1	$\sum_i E^2_{Y_iY_i}-\frac{\left[\sum_i E_{X_iY_i}\right]^2}{\sum_i E_{X_iX_i}}$	

Ver también el epígrafe *Prueba de homogeneidad de dos o más líneas de regresión* del Capítulo 11 y del Capítulo 13.

Ejemplo.-

Volvemos al mismo ejemplo del principio. Este ejemplo también esta resuelto en el Capítulo 11.

En un estudio de la ganancia de peso (*Y*) de lechones se probaron tres dietas.

Como el peso al inicio de la experiencia influye en la ganancia de peso, se tomó este peso inicial como covariable (X). Se desea saber si la regresión del peso inicial sobre el incremento de peso es la misma para las tres dietas.

	Dietas							
	A		*B*		*C*		*Global*	
	X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
	14.8	4.22	15.8	3.78	19.8	4.07		
	18.9	3.84	19.9	4.65	19.9	4.02		
	19.2	4.43	21.2	4.04	19.2	4.28		
	19.7	4.46	15.7	3.71	16.7	4.02		
	21.4	4.74	19.4	4.20	19.4	4.02		
	19.8	4.41	19.8	4.46	19.8	4.36		
	17.9	4.42	17.9	4.33	15.9	4.09		
	16.2	4.43	19.2	4.46	15.2	3.70		
	18.7	4.61	18.7	4.08	20.7	4.41		
	16.4	4.01	18.4	4.42	15.4	3.94		
Σ	183.0	43.57	186.0	42.13	182.0	40.91	551.00	126.61
Σ^2	3384.48	190.48	3487.28	178.35	3353.48	167.76	10225.24	536.59
ΣXY	799.559		786.705		747.444		2333.708	
$\overline{m}$	18.30	4.357	18.6	4.213	18.2	4.091	18.366	4.22

Dieta	*gl*	$SC_{(X)}$	*SP*	$SC_{(Y)}$	*gl*	$SC_{(regresión)}$	*gl*	$SC_{(error)}$
1	9	35.58	2.228	0.6455	1	0.1395	8	0.5060
2	9	27.28	3.087	0.8563	1	0.3443	8	0.5120
3	9	41.08	2.882	0.3972	1	0.2022	8	0.1950
Σ	27	103.94	8.197	1.899		0.686	24	1.213

$$SC_{(\text{entre b})} = 0.686 - \frac{8.197^2}{103.94} = 0.040$$

FV	*gl*	*SC*	*CM*	F_o
Entre b	2	0.040	0.020	0.40*ns*
Error	24	1.213	0.0505	
Total	26	1.253		

Se concluye que en las tres dietas existe la misma relación funcional entre el peso

inicial y el peso final, por lo que se puede hacer una estima conjunta de la regresión.

Archivo del programa *SAS* (C17-5.SAS).-

```
title 'Homogeneidad de coeficiente de regresión';
options ls=75 ps=60;
data homoregr;
infile 'c17-5.dat';
input dieta $ X Y @@;
title 'Coeficiente de regresión en cada dieta';
proc glm;
  class dieta;
  model Y = dieta X(dieta) / solution;
run;
title 'Homogeneidad de coeficientes de regresión';
proc glm;
  class dieta;
  model Y = dieta X dieta*X;
run;
```

Archivo de datos (C17-5.DAT).-

```
A 14.8  4.22  B 15.8  3.78  C 19.8  4.07
A 18.9  3.84  B 19.9  4.65  C 19.9  4.02
A 19.2  4.43  B 21.2  4.04  C 19.2  4.28
A 19.7  4.46  B 15.7  3.71  C 16.7  4.02
A 21.4  4.74  B 19.4  4.20  C 19.4  4.02
A 19.8  4.41  B 19.8  4.46  C 19.8  4.36
A 17.9  4.42  B 17.9  4.33  C 15.9  4.09
A 16.2  4.43  B 19.2  4.46  C 15.2  3.70
A 18.7  4.61  B 18.7  4.08  C 20.7  4.41
A 16.4  4.01  B 18.4  4.42  C 15.4  3.94
```

Archivo de resultados (C17-5.LST).-

```
                Coeficiente de regresión en cada dieta
                   General Linear Models Procedure
Dependent Variable: Y

                                    Sum of          Mean
Source                      DF      Squares        Square   F Value    Pr > F
Model                        5   1.04056813   0.20811363      4.10    0.0078
Error                       24   1.21812853   0.05075536
Corrected Total             29   2.25869667

                     R-Square         C.V.      Root MSE             Y Mean
                     0.460694     5.338192       0.22529            4.22033

Source                      DF    Type I SS   Mean Square   F Value    Pr > F
DIETA                        2   0.35458667   0.17729333      3.49    0.0466
X(DIETA)                     3   0.68598147   0.22866049      4.51    0.0121

Source                      DF   Type III SS   Mean Square   F Value    Pr > F
DIETA                        2   0.05232051   0.02616026      0.52    0.6037
X(DIETA)                     3   0.68598147   0.22866049      4.51    0.0121

                                                T for H0:     Pr > |T|    Std Error of
Parameter                Estimate               Parameter=0                Estimate
INTERCEPT          2.814164557 B                       4.37       0.0002    0.64368528
DIETA         A    0.396899524 B                       0.42       0.6789    0.94717023
              B   -0.675521493 B                      -0.66       0.5168    1.02653538
              C    0.000000000 B                          .            .             .
X(DIETA)      A    0.062619449                         1.66       0.1103    0.03776922
              B    0.111524566                         2.60       0.0155    0.04282111
              C    0.070155794                         2.00       0.0574    0.03515003
NOTE: The X'X matrix has been found to be singular and a generalized
```

```
      inverse was used to solve the normal equations.   Estimates followed
      by the letter 'B' are biased, and are not unique estimators of the
      parameters.

                  Homogeneidad de coeficientes de regresión
Dependent Variable: Y
                                        Sum of           Mean
Source                        DF       Squares         Square  F Value     Pr > F
Model                          5    1.04056813     0.20811363     4.10     0.0078
Error                         24    1.21812853     0.05075536
Corrected Total               29    2.25869667

                    R-Square             C.V.      Root MSE              Y Mean
                    0.460694         5.338192       0.22529             4.22033

Source                        DF     Type I SS   Mean Square  F Value     Pr > F
DIETA                          2    0.35458667     0.17729333     3.49     0.0466
X                              1    0.64396022     0.64396022    12.69     0.0016
X*DIETA                        2    0.04202125     0.02101062     0.41     0.6657

Source                        DF    Type III SS   Mean Square  F Value     Pr > F
DIETA                          2    0.05232051     0.02616026     0.52     0.6037
X                              1    0.67380185     0.67380185    13.28     0.0013
X*DIETA                        2    0.04202125     0.02101062     0.41     0.6657
```

Análisis de covarianza con varias covariables.-

Una variable dependiente puede verse afectada por más de una covariable, en este caso, conceptualmente, no cambia nada, lo único es que se tiene tantas regresiones como covariables hay, y los datos se ajustan para todas ellas. Como la magnitud de los cálculos si es considerable, es mejor realizarlos directamente con un paquete estadístico.

Ejemplo.-

En un estudio de la producción de leche en tres razas, se midieron en 10 individuos de cada raza la producción total en *Kg* (**prod**), y además, como covariables se tomaron medidas productiva de la fracción de la caseína, estas fueron, porcentaje de caseína α (**alfa**), porcentaje de la caseína β (**beta**) y porcentaje de la caseína κ (**kapa**). Estas covariables se incluyen para controlar el error y ajustar las medias de las razas.

Se quiere saber si la producción es diferente en las tres razas.

Archivo del programa *SAS* (C17-6.SAS).-

```
title 'Análisis de covarianza de una vía con tres covariables';
options ls=75 ps=60;
data ancova;
infile 'c17-6.dat';
input raza $ prod c_alfa c_beta c_kapa;
title 'ANOVA de las cuatro variables';
proc anova;
  class raza;
  model prod c_alfa c_beta c_kapa = raza;
run;
title 'ANCOVA de la producción con tres covariables';
proc glm;
  class raza;
   model prod = raza c_alfa c_beta c_kapa / solution;
   lsmeans raza / stderr tdiff;
run;
```

Archivo de datos (C17-6.DAT).-

```
A 178.60   1.87   2.81   0.61
A 278.18   2.82   4.64   1.45
A 346.25   3.23   4.49   1.39
A 402.53   3.55   5.30   1.03
A 321.38   3.25   4.78   2.02
A 272.25   2.72   4.27   1.16
A 268.90   2.70   4.11   1.23
A 304.75   2.88   4.13   0.87
A 325.95   2.94   5.98   1.09
B 330.23   3.06   5.33   1.55
B 355.45   3.43   4.82   1.88
B 273.40   1.97   4.50   0.94
B 410.98   3.95   5.08   1.94
B 497.65   4.21   6.90   2.05
B 521.58   4.22   7.39   2.46
B 457.10   4.01   7.08   1.82
B 596.23   4.50   8.93   2.54
B 326.20   3.11   4.05   1.33
B 256.55   2.45   3.29   1.02
B 256.95   2.45   3.16   1.20
C 317.85   3.02   3.74   1.71
C 312.90   2.88   3.96   1.55
C 222.25   2.22   3.56   0.57
C 174.08   1.24   4.47   0.17
C 224.85   2.13   2.86   0.81
C 560.10   4.41   8.76   1.70
C 254.20   2.45   2.93   1.17
C 682.75   5.43   8.93   2.10
C 400.98   3.85   4.98   1.84
C 229.25   2.18   2.85   0.97
```

Archivo de resultados (C17-6.LST).-

```
                    ANOVA de las cuatro variables

                    Analysis of Variance Procedure
Dependent Variable: PROD

                                  Sum of          Mean
Source                    DF     Squares        Square   F Value    Pr > F
Model                      2   40420.7970   20210.3985      1.33    0.2812
Error                     27  410172.0925   15191.5590
Corrected Total           29  450592.8895

                  R-Square         C.V.      Root MSE          PROD Mean
                  0.089706     35.69022       123.254            345.344

Source                    DF    Anova SS  Mean Square  F Value    Pr > F
RAZA                       2   40420.7970   20210.3985      1.33    0.2812

Dependent Variable: C_ALFA

                                  Sum of          Mean
Source                    DF     Squares        Square   F Value    Pr > F
Model                      2  1.52536990   0.76268495      0.89    0.4221
Error                     27 23.12136677   0.85634692
Corrected Total           29 24.64673667

                  R-Square         C.V.      Root MSE        C_ALFA Mean
                  0.061889     29.80963        0.92539            3.10433

Source                    DF    Anova SS  Mean Square  F Value    Pr > F
RAZA                       2  1.52536990   0.76268495      0.89    0.4221
```

Dependent Variable: C_BETA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.77337293	2.88668646	0.89	0.4211
Error	27	87.26714707	3.23211656		
Corrected Total	29	93.04052000			

R-Square	C.V.	Root MSE	C_BETA Mean
0.062052	36.42238	1.79781	4.93600

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAZA	2	5.77337293	2.88668646	0.89	0.4211

Dependent Variable: C_KAPA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.54620626	0.77310313	2.71	0.0848
Error	27	7.71173040	0.28561964		
Corrected Total	29	9.25793667			

R-Square	C.V.	Root MSE	C_KAPA Mean
0.167014	38.01996	0.53443	1.40567

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAZA	2	1.54620626	0.77310313	2.71	0.0848

ANCOVA de la producción con tres covariables
General Linear Models Procedure

Dependent Variable: PROD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	443060.873	88612.175	282.35	0.0001
Error	24	7532.016	313.834		
Corrected Total	29	450592.890			

R-Square	C.V.	Root MSE	PROD Mean
0.983284	5.129772	17.7154	345.344

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RAZA	2	40420.797	20210.399	64.40	0.0001
C_ALFA	1	384058.164	384058.164	1223.76	0.0001
C_BETA	1	18355.910	18355.910	58.49	0.0001
C_KAPA	1	226.003	226.003	0.72	0.4045

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	3218.6614	1609.3307	5.13	0.0140
C_ALFA	1	25971.6409	25971.6409	82.76	0.0001
C_BETA	1	16810.4994	16810.4994	53.56	0.0001
C_KAPA	1	226.0028	226.0028	0.72	0.4045

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		-44.89258894 B	-3.54	0.0017	12.69380014
RAZA	A	-24.48287614 B	-3.00	0.0061	8.14960480
	B	-2.75713858 B	-0.32	0.7484	8.49849779
	C	0.00000000 B	.	.	.
C_ALFA		92.07064593	9.10	0.0001	10.12095498
C_BETA		25.94656951	7.32	0.0001	3.54519098
C_KAPA		-10.88297820	-0.85	0.4045	12.82451183

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

```
                        Least Squares Means
RAZA          PROD          Std Err      Pr > |T|    LSMEAN
             LSMEAN           LSMEAN    H0:LSMEAN=0   Number
A        329.216938        6.089603          0.0001      1
B        350.942675        5.816952          0.0001      2
C        353.699814        5.731331          0.0001      3

      T for H0: LSMEAN(i)=LSMEAN(j) / Pr > |T|
            i/j          1           2           3
            1            .     -2.46411    -3.00418
                               0.0213      0.0061
            2      2.464114          .     -0.32443
                     0.0213                0.7484
            3      3.00418    0.324427           .
                     0.0061      0.7484
NOTE: To ensure overall protection level, only probabilities associated
      with pre-planned comparisons should be used.
```

En el archivo de resultados (C17-6.LST) se observa que no existe diferencia entre las razas para la producción, ni para ninguna de las covariables. Sin embargo, si la producción se ajusta para las tres covariables, sí son significativas las razas (F=5.13*), siendo las mas productiva la raza C con 352.70, y la menos productiva la raza A con 329.22; la diferencia entre la producciones de la razas C y B no es significativa (t=-0.324*ns*).

Las regresiones de las covariables con la producción valen: 92.07*** con la alfa caseína; 25.95*** con la beta caseína y –10.88*ns* con la kappa caseína.

Análisis multivariante de la covarianza. Varias variable dependientes con varias covariables.-

Puede ocurrir que varias variable dependiente se vean afectada por una o más de una covariable, en este caso, conceptualmente, no cambia nada, lo único es que se tienen tantas regresiones como pares de variables covariables hay, y los datos se ajustan para todas ellas. Como la magnitud de los cálculos si es considerable, es mejor realizarlos directamente con un paquete estadístico.

Ejemplo.-

En un estudio de las características productivas de la leche en tres razas, se midieron en 10 individuos de cada raza estas características productivas (producción total en *Kg*, **prod**, porcentaje de proteína total, **prot**, porcentaje de caseína total, **cas**, porcentaje de grasa, **gras,** y porcentaje de lactosa, **lac;**, y además, como covariables se tomaron medidas productiva de la fracción de la caseína, estas fueron, porcentaje de caseína α (**alfa**), porcentaje de la caseína β (**beta**) y porcentaje de la caseína κ (**kapa**). Estas covariables se incluyen para controlar el error y ajustar las medias de las razas.

Se quiere saber si la producción es diferente en las tres razas.

Archivo del programa *SAS* (C17-7.SAS).-

```
title 'Análisis de Multivariante de la covarianza de una vía';
options ls=75 ps=60;
data ancova;
infile 'c17-7.dat';
input raza $ prod prot cas gras lac c_alfa c_beta c_kapa;
title 'ANOVA de las ocho variables';
proc anova;
  class raza;
  model prod prot cas gras lac c_alfa c_beta c_kapa = raza;
run;
title 'ANCOVA de la variables productivas ajustando con las covariables';
proc glm;
  class raza;
   model prod prot cas gras lac = raza c_alfa c_beta c_kapa / solution;
manova h=raza;
run;
```

Si se quiere evitar las salidas univariantes se introduce en el modelo la opción **NOUNI** (.model prod prot cas gras lac = raza c_alfa c_beta c_kapa / nouni solution;).

Archivo de datos (C17-7.DAT).-

```
A 178.60   5.64   4.92  12.23   6.92   1.87   2.81   0.61
A 278.18   9.17   8.06  20.16  10.68   2.82   4.64   1.45
A 346.25   9.85   8.74  19.14  13.41   3.23   4.49   1.39
A 402.53  10.69   9.55  22.80  15.51   3.55   5.30   1.03
A 321.38  11.85   9.68  24.44  12.37   3.25   4.78   2.02
A 272.25   9.89   7.90  17.87  10.04   2.72   4.27   1.16
A 268.90   8.22   7.66  16.47  10.39   2.70   4.11   1.23
A 304.75   8.92   7.56  18.32  11.70   2.88   4.13   0.87
A 325.95   9.98   9.49  24.26  12.00   2.94   5.98   1.09
B 330.23  11.69   9.68  23.24  11.96   3.06   5.33   1.55
B 355.45  11.85   9.76  24.79  12.69   3.43   4.82   1.88
B 273.40   8.58   7.31  13.03  12.88   1.97   4.50   0.94
B 410.98  14.27  10.97  24.80  15.43   3.95   5.08   1.94
B 497.65  19.02  13.00  27.98  19.74   4.21   6.90   2.05
B 521.58  19.85  13.71  28.81  20.82   4.22   7.39   2.46
B 457.10  14.94  12.77  30.87  17.10   4.01   7.08   1.82
B 596.23  22.95  15.66  30.86  24.11   4.50   8.93   2.54
B 326.20  10.48   8.46  21.47  12.46   3.11   4.05   1.33
B 256.55   8.59   6.77  18.05   9.84   2.45   3.29   1.02
B 256.95   8.86   6.95  17.36   9.89   2.45   3.16   1.20
C 317.85  11.06   8.58  21.43  12.56   3.02   3.74   1.71
C 312.90  10.65   8.50  22.09  12.24   2.88   3.96   1.55
C 222.25   8.47   6.66  16.77   8.78   2.22   3.56   0.57
C 174.08   5.71   6.08  15.68   5.38   1.24   4.47   0.17
C 224.85   7.07   6.07  14.47   8.97   2.13   2.86   0.81
C 560.10  19.75  14.55  29.69  22.87   4.41   8.76   1.70
C 254.20   9.00   6.74  19.32  10.20   2.45   2.93   1.17
C 682.75  23.97  19.09  40.35  25.89   5.43   8.93   2.10
C 400.98   7.37   5.96  16.36   8.90   3.85   4.98   1.84
C 229.25   8.89   8.57  22.08  10.22   2.18   2.85   0.97
```

Archivo de resultados (C17-7.LST).-

ANCOVA de la variables productivas ajustando con las covariables

Dependent Variable: PROD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	443060.873	88612.175	282.35	0.0001
Error	24	7532.016	313.834		
Corrected Total	29	450592.890			

R-Square	C.V.	Root MSE	PROD Mean
0.983284	5.129772	17.7154	345.344

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	3218.6614	1609.3307	5.13	0.0140
C_ALFA	1	25971.6409	25971.6409	82.76	0.0001
C_BETA	1	16810.4994	16810.4994	53.56	0.0001
C_KAPA	1	226.0028	226.0028	0.72	0.4045

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		-44.89258894 B	-3.54	0.0017	12.69380014
RAZA	A	-24.48287614 B	-3.00	0.0061	8.14960480
	B	-2.75713858 B	-0.32	0.7484	8.49849779
	C	0.00000000 B	.	.	.
C_ALFA		92.07064593	9.10	0.0001	10.12095498
C_BETA		25.94656951	7.32	0.0001	3.54519098
C_KAPA		-10.88297820	-0.85	0.4045	12.82451183

Dependent Variable: PROT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	612.992713	122.598543	41.21	0.0001
Error	24	71.401823	2.975076		
Corrected Total	29	684.394537			

R-Square	C.V.	Root MSE	PROT Mean
0.895672	14.90229	1.72484	11.5743

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	11.3984763	5.6992381	1.92	0.1691
C_ALFA	1	5.1258742	5.1258742	1.72	0.2017
C_BETA	1	54.6832014	54.6832014	18.38	0.0003
C_KAPA	1	4.4861428	4.4861428	1.51	0.2314

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		-1.553448859 B	-1.26	0.2209	1.23592094
RAZA	A	-1.330251094 B	-1.68	0.1066	0.79347926
	B	0.140925146 B	0.17	0.8662	0.82744893
	C	0.000000000 B	.	.	.
C_ALFA		1.293467334	1.31	0.2017	0.98541808
C_BETA		1.479846277	4.29	0.0003	0.34517447
C_KAPA		1.533300911	1.23	0.2314	1.24864757

Dependent Variable: CAS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	270.235739	54.047148	37.43	0.0001
Error	24	34.655327	1.443972		
Corrected Total	29	304.891067			

R-Square	C.V.	Root MSE	CAS Mean
0.886335	12.90251	1.20165	9.31333

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	1.7097717	0.8548858	0.59	0.5611
C_ALFA	1	4.5910107	4.5910107	3.18	0.0872
C_BETA	1	28.3611961	28.3611961	19.64	0.0002
C_KAPA	1	0.0587274	0.0587274	0.04	0.8419

Parameter		Estimate		T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		0.196765902	B	0.23	0.8212	0.86103547
RAZA	A	-0.562867588	B	-1.02	0.3187	0.55279733
	B	-0.059356301	B	-0.10	0.9188	0.57646315
	C	0.000000000	B	.	.	.
C_ALFA		1.224124669		1.78	0.0872	0.68651634
C_BETA		1.065741579		4.43	0.0002	0.24047449
C_KAPA		0.175432940		0.20	0.8419	0.86990180

Dependent Variable: GRAS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	906.928856	181.385771	21.16	0.0001
Error	24	205.734241	8.572260		
Corrected Total	29	1112.663097			

R-Square	C.V.	Root MSE	GRAS Mean
0.815097	13.40608	2.92784	21.8397

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	13.3665566	6.6832783	0.78	0.4698
C_ALFA	1	36.7565529	36.7565529	4.29	0.0493
C_BETA	1	40.8173703	40.8173703	4.76	0.0391
C_KAPA	1	0.4948154	0.4948154	0.06	0.8122

Parameter		Estimate		T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		4.843402854	B	2.31	0.0299	2.09792195
RAZA	A	-1.681834797	B	-1.25	0.2238	1.34689649
	B	-0.758939449	B	-0.54	0.5939	1.40455851
	C	0.000000000	B	.	.	.
C_ALFA		3.463688942		2.07	0.0493	1.67270426
C_BETA		1.278533609		2.18	0.0391	0.58591863
C_KAPA		0.509228207		0.24	0.8122	2.11952485

Dependent Variable: LAC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	641.586664	128.317333	35.23	0.0001
Error	24	87.411553	3.642148		
Corrected Total	29	728.998217			

R-Square	C.V.	Root MSE	LAC Mean
0.880094	14.45971	1.90844	13.1983

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RAZA	2	5.1060905	2.5530453	0.70	0.5060
C_ALFA	1	19.7975930	19.7975930	5.44	0.0285
C_BETA	1	48.4109018	48.4109018	13.29	0.0013
C_KAPA	1	0.0528208	0.0528208	0.01	0.9051

Parameter		Estimate		T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		-1.317078319	B	-0.96	0.3451	1.36747834
RAZA	A	-0.635279151	B	-0.72	0.4763	0.87794104
	B	0.479371174	B	0.52	0.6054	0.91552660
	C	0.000000000	B	.	.	.
C_ALFA		2.542012555		2.33	0.0285	1.09031075
C_BETA		1.392391055		3.65	0.0013	0.38191651
C_KAPA		-0.166376989		-0.12	0.9051	1.38155965

Manova Test Criteria and F Approximations for
the Hypothesis of no Overall RAZA Effect
H = Type III SS&CP Matrix for RAZA E = Error SS&CP Matrix
S=2 M=1 N=9

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.39558430	2.3598	10	40	0.0266
Pillai's Trace	0.63254738	1.9428	10	42	0.0658
Hotelling-Lawley Trace	1.45679190	2.7679	10	38	0.0114
Roy's Greatest Root	1.40622070	5.9061	5	21	0.0015

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Los ANOVA de cada variable nos confirma que las razas no son significativas para ninguna de ellas (se ha quitado de la salida).

La salida univariante de cada variable productiva con las tres covariables, nos muestra el ANCOVA para cada variable dependiente.

El valor de la Lambda de Wilk es 0.3956 y su *F* vale 2.35*, por lo que, efectivamente, las variables productivas, tomadas en conjunto, son diferentes en las diferentes razas cuando se ajustan con la regresión de las variables de las fracción de la caseína.

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

Freund, R.J., and Littel, R.C. 1991. *SAS* SYSTEM FOR REGRESION. SAS Institute Inc., Cary, NC, USA.

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

Lite, TM, y Jackson Hills, F. 1987. METODOS ESTADISTICOS PARA LA INVESTIGACION EN LA AGRICULTURA. Ed TRILLAS. México.

Littel, R.C., Freund, R.J. and Spector, P.C. 1991. *SAS* FOR LINEAR MODELS. SAS Institute Inc., Cary, NC, USA.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. MÉTODOS ESTADÍSTICOS. Ed C.E.C.S.A. México.

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

SAS Institute Inc. 1990. *SAS* PROCEDURE GUIDE. Cary, NC, USA.

CAPÍTULO 18

Análisis de Componentes Principales y Análisis Factorial

CAPÍTULO 18

Análisis de Componentes Principales y Análisis Factorial

El análisis de componentes principales tiene como finalidad el simplificar la descripción que se obtiene de un grupo de variables interelacionadas. En este análisis todas las variables son del mismo tipo, esto es, no hay variables independientes y variables dependientes.

El análisis de componentes principales opera sobre p variables aleatorias medidas en n unidades experimentales o individuos, es decir, a cada uno de los n individuos se le miden p variables. Este análisis trata de encontrar m ($<p$) nuevas variables llamadas componentes principales y determinar su contribución a la explicación de las variables originales.

El método estadístico consiste en transformar las variables originales en las componentes principales por medio de una combinación lineal de las variables originales, midiéndose su contribución a la información total de las variables originales por medio de su varianza. Es por esto por lo que las componentes principales se van a ordenar en orden decreciente de sus varianzas, siendo la componente principal más informativa (más varianza) la primera y la menos informativa (menos varianza) la última.

El análisis de componentes principales es útil para varias finalidades. La primera ya se ha dicho, se puede reducir la dimensión de un problema reduciendo el número de variables si se eliminan las variables que tienen poca información; este objetivo se consigue eligiendo las primeras componentes principales sin que se pierda información pues, al ser las componentes principales variables incorrelacionadas, unas pocas pueden tener prácticamente la misma información que muchas de las variables originales correlacionadas.

También se puede utilizar para probar la multinormalidad de las variables pues si las componentes principales no se distribuyen normalmente es porque las variables originales no se ajustan a una multinormalidad. También se puede utilizar para detectar

datos anómalos o individuos pertenecientes a otra población.

Si se tiene un análisis de regresión múltiple donde se presenta el problema de la multicolinealidad como consecuencia de la correlación entre las variables independientes, se puede evitar este problema utilizando como variables independientes las primeras componentes principales de las variables *X*.

Interpretación geométrica.-

Considerando los valores de las variables como dimensiones o ejes de un espacio *p-dimensional*, desde el punto de vista geométrico, el proceso de elaborar las componentes principales o factores comunes consiste en cambiar los ejes de la nube constituida por los *n* puntos representados *p-dimensionalmente* (una dimensión para cada variable medida). Para que estos nuevos ejes sean diferentes de los originales tiene que haber correlación entre las variables, esto quiere decir que si se representan, por ejemplo, dos variables en un eje de coordenadas bidimensional, la nube de puntos no ocupa todo el espacio sino que tiende a concentrarse en un subespacio en forma de elipse, en ese caso la componente principal o nuevo eje es la recta mínimocuadrada de dicha nube de puntos. Esta recta mínimo cuadrada puede explicar la mayoría de la varianza de las dos variables si la correlación entre las dos variables es alta. Si no existe correlación entre las variables, la representación de estas en un eje de coordenadas daría puntos que tienden a ocupar todo el espacio, por lo que la recta mínimo cuadrada coincidiría con los ejes originales.

En el caso de que se midan 100 variables a cada unidad experimental, se tendría un espacio de 100 dimensiones. Si se pretenden reducir estas 100 dimensiones a un número más pequeño trazando nuevos ejes principales, es poco probable que se logre incluir suficiente información en una sola dimensión o recta minimocuadrada (primer componente principal) que es la que pasará por el eje principal de la nube de puntos y, por lo tanto, será la que englobe una mayor cantidad de información. Se necesitará otro eje. Por conveniencia, se representa la segunda dimensión mediante una recta perpendicular al primer componente principal. Ese segundo eje, o segundo componente principal, se define como la línea que explica más, de la variabilidad restante, que cualquier otra posible recta perpendicular al primer componente principal. Y así sucesivamente.

Por ejemplo, si la nube de puntos *p-dimensional* (tridimensional) tiene forma de lenguado (huso), el primer componente o eje principal pasaría por el centro, en sentido longitudinal (de la boca a la cola), y el segundo componente principal también pasaría por el centro, pero perpendicular a la primera componente, por lo tanto se trazará en sentido transversal (en el plano horizontal y no vertical). Las líneas siguientes tendrían que ser perpendiculares a las dos anteriores y explicarían cantidades cada vez menores de la variabilidad restante. En este caso, los dos primeros ejes principales darían suficiente información como para reconocer un lenguado tridimensional por medio de su representación bidimensional.

Puede suceder que cinco componentes principales expliquen casi la totalidad de la variabilidad de 100 variables, o sea, que la nube de puntos de dimensión 100 se puede reducir a 5 dimensiones de manera que la representación *5-dimensional* se

parezca suficientemente a la representación *100-dimensional* como para que no se pierda información sensible. De la misma manera se puede representar en dos dimensiones la información suficiente de una pizza o un lenguado, que tal como se conocen son tridimensionales, como para que se reconozcan suficientemente. Si se decide representar las 100 dimensiones en 5, se puede lograr una simplificación considerable a cambio de una pérdida pequeña de información. Perdida de información que viene compensada por la posibilidad de entender conceptualmente las cinco dimensiones e incluso poder interpretarlas biológicamente (o socialmente o económicamente, *etc.*).

Ejemplo ilustrativo.-

Repasemos todos estos conceptos más detenidamente con un ejemplo muy simple, supóngase que se toma una muestra aleatoria de tres individuos (n=3) a los que se les mide dos variables (p=2)

Individuo	X_1	X_2
1	5	14
2	4	10
3	7	15
$\overline{X}$	5.333	13.0
s^2	2.333	7

siendo X_1 medidas en *cm* y X_2 pesos en gramos.

Si consideramos estas variables como las coordenadas de los individuos, la distancia euclídea entre los individuos 1 y 2 y entre los individuos 1 y 3 es, respectivamente,

$$d_{(1,2)} = (4-5)^2 + (10-14)^2 = 17$$

$$d_{(1,3)} = (7-5)^2 + (15-14)^2 = 5$$

Esto es, $d_{(1,2)} > d_{(1,3)}$, el individuo uno está más cerca al individuo 3 que al 2.

Si se cambia de escala de la primera variable y las medidas se expresan en *mm*, las distancias anteriores serían

$$d_{(1,2)} = (40-50)^2 + (10-14)^2 = 116$$

$$d_{(1,3)} = (70-50)^2 + (15-14)^2 = 401$$

siendo los valores los mismos, la magnitud relativa de las distancias son las contrarias, esto es, $d_{(1,2)} < d_{(1,3)}$, por lo que en *mm* el individuo mas cercano al 1 es el 2 y no el 3.

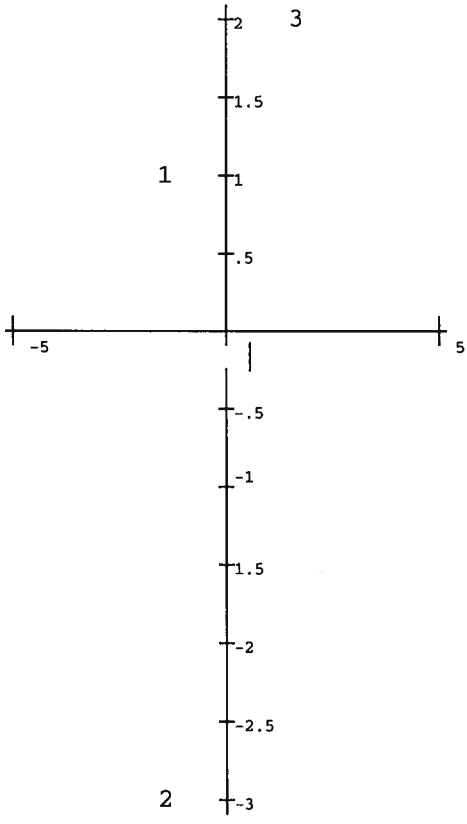
Para evitar estos problemas de escala, dado que las variables pueden ser de

muchos tipos, se le resta a cada observación la media muestral de la variable de manera que los valores del primer individuos son

$$X_1 = 5 - 5.3333 = -0.3333$$
$$X_2 = 14 - 13 = 1$$

Recuérdese que esta operación hace que la media de la nueva variable valga cero (se traslada el *parámetro de localización* al origen de coordenadas) pero no cambian los valores ni de las varianzas muestrales ni de las correlaciones entre las variables.

Si representamos estos valores, corregidos para la media, se tiene la gráfica



La idea básica, tal como se ha dicho anteriormente, es crear nuevas variables, C_1 y C_2, denominadas componentes principales que serán funciones lineales de las variables originales tipificadas, esto es

$$C_1 = a_{11} x_1 + a_{12} x_2$$
$$C_2 = a_{21} x_1 + a_{22} x_2$$

Hay que anotar que cualquier conjunto de valores de los cuatro coeficientes a_{11},

a_{12}, a_{21} y a_{12} se pueden aplicar a las N observaciones de x_1 y x_2 y obtener N valores de C_1 y C_2.

A estos coeficientes (a_{ij}) se les denominan ***vectores propios*** (a_{1j} es el primer vector propio y a_{2j} es el segundo vector propio), también denominados ***vectores característicos***, ***vectores latentes*** o ***autovectores***, aunque también se les denomina en su versión inglesa de ***eigenvectors***.

La media y la varianza de los N valores de las componentes principales son

$$\overline{C_1} = \overline{C_2} = 0$$

$$Var(C_1) = a_{11}^2 s_1^2 + a_{12}^2 s_2^2 + 2\, a_{11} a_{12} r s_1 s_2$$

$$Var(C_2) = a_{21}^2 s_1^2 + a_{22}^2 s_2^2 + 2\, a_{21} a_{22} r s_1 s_2$$

$$Var(C_1) + Var(C_2) = s_1^2 + s_2^2$$

siendo s_i^2 la varianza de la *i-ésima* variable.

A las varianzas de los componentes principales se les conoce comúnmente como ***valores propios*** (primer valor propio y segundo valor propio, respectivamente) también se les conoce como ***raíces características***, ***raíces latentes*** o simplemente ***autovalores***, o por su denominación en ingles, ***eigenvalues***.

Para hallar los valores de los coeficientes o vectores propios se tiene que satisfacer los siguientes requerimientos

1. El valor de la varianza de C_1 tiene que ser el máximo posible.
2. Los N valores de C_1 y C_2 tienen que estar incorrelacionados
3. $a^2_{11}+a^2_{12}=a^2_{21}+a^2_{22}=1$

La manera de hallar estos coeficientes fue deducida matemáticamente por *Hotelling* en 1933 y se explicará más adelante, aunque hoy día hay multitud de paquetes estadísticos que ofrecen estas soluciones. Siguiendo con el ejemplo que se está desarrollando, los dos componentes principales son

$$C_1 = 0.471858\, X_1 + 0.881675\, X_2$$

$$C_2 = 0.881674\, X_1 - 0.471858\, X_2$$

obsérvese que

$$0.471858^2 + 0.881675^2 = 1$$

$$0.881675^2 + (-0.471858)^2 = 1$$

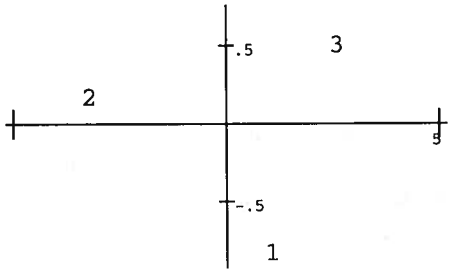
que es el requerimiento número tres. Nótese, así mismo, que para un caso como este, en el que hay solo dos variables, a_{11}=-a_{22} y a_{12}=a_{21}.

Una vez obtenidos los vectores propios se pueden calcular las dos componentes principales de cada individuo

$$C_{11} = 0.471858 \times -0.33333 + 0.881675 \times 1 = 0.72439$$
$$C_{12} = 0.471858 \times -1.33333 + 0.881675 \times -3 = -3.27417$$
$$C_{13} = 0.471858 \times 1.66667 + 0.881675 \times 2 = 2.54978$$

$$C_{21} = 0.881675 \times -0.33333 - 0.471858 \times 1 = -0.76575$$
$$C_{22} = 0.881675 \times -1.33333 - 0.471858 \times -3 = 0.24001$$
$$C_{23} = 0.881675 \times 1.66667 - 0.471858 \times 2 = 0.52574$$

que representadas en una gráfica se obtiene



Compárese esta gráfica con la anterior y se observará, en primer lugar, que lo que ha ocurrido es un giro de manera que los individuos están más próximos a los ejes, esto es, la suma de los cuadrados de las distancias a los ejes es mínima (mínimo cuadrados). En segundo lugar, es posible observar (o al menos atisbar) a simple vista que la varianza total se mantiene igual a la de las variables originales, pero la elipse ha girado de manera que ahora la varianza más grande la tiene el primer eje (en las variables originales estaba en el segundo eje), tal como exigían los requerimientos previos.

La varianza de las componentes principales o valores propios son

$$Var(C_1) = 0.471858^2 \times 2.333 + 0.881675^2 \times 7.0 + 2 \times$$
$$\times 0.471858 \times 0.881675 \times 0.8663 \times 1.5275 \times 2.6457 = 8.87314$$
$$Var(C_2) = 0.881675^2 \times 2.333 + (-0.471858)^2 \times 7.0 + 2 \times$$
$$\times 0.881675 \times -0.471858 \times 0.8663 \times 1.5275 \times 2.6457 = 0.46019$$

Compruébese que la varianza de la primera componente es mucho mayor que la varianza de la segunda componente y que la suma de estas dos varianzas (=9.3333) o valores propios es igual a la suma de las varianzas de las dos variables originales, esto quiere decir que la varianza total se mantiene en el proceso de rotación de los ejes pero que el primer eje absorbe un máximo de la varianza total, de manera que se puede decir que el primer eje o componente explica el 95% de la varianza total

$$\frac{8.87314}{9.333333} = 0.9507$$

Estas ideas básicas son fácilmente extensibles al caso de p variables x_1, x_2,..., x_p. Cada componente principal es una combinación lineal de las variables originales. Los coeficientes de estas funciones lineales se eligen de manera que cumplan

1. $Var\ C_1 \geq Var\ C_2 \geq ... \geq Var\ C_p$
2. Los valores de todas las componentes están incorrelacionados.
3. La suma de cuadrados de los coeficientes (vectores propios) de todas las componentes principales valen uno.

En otras palabras, y resumiendo, C_1 es la combinación lineal que tiene varianza máxima. C_2 es la combinación lineal que, cumpliendo la condición de no estar correlacionada con C_1, tiene varianza máxima. Similarmente, C_3 es la combinación lineal que, cumpliendo la condición de no estar correlacionada ni con C_1 ni con C_2, tiene varianza máxima, *etc.* La varianza de C_j es el *j-ésimo* valor propio. Los p valores propios suman lo mismo que la suma de las varianzas originales. Los coeficientes de las combinaciones lineales son los vectores propios.

Cálculo matemático.-

Como se ha dicho anteriormente, fue *Hotelling* quien dedujo la manera de obtener los valores y los vectores propios. Se va ha exponer a continuación la metodología teórica para obtener estos resultados, con objeto de tener una idea intuitiva de donde proceden los valores, aunque nunca se tendrá que realizar estas operaciones pues, como se ha dicho anteriormente, hoy día hay multitud de paquetes estadísticos que realizan estos cálculos.

Para ello, lo primero es obtener la matriz de varianzas- covarianzas de los datos, esta es

$$\mathbf{V} = \begin{pmatrix} 2.3333 & 3.5000 \\ 3.5000 & 7.0000 \end{pmatrix}$$

Como se ve, en la diagonal van los valores de las varianzas y en los demás miembros los valores de la covarianzas.

Si **V** es no singular, Los coeficientes del *i-ésimo* vector propio, a_i, satisfacen el sistema de ecuaciones

$$(\mathbf{V} - \lambda_i \mathbf{I})\mathbf{a}_i = 0$$

siendo λ_i el *i-ésimo* valor propio o varianza de la *i-ésima* componente, que se pueden obtener resolviendo el determinante

$$|\mathbf{V} - \lambda \mathbf{I}| = 0$$

Para el ejemplo

$$\begin{vmatrix} 2.3333 & 3.5000 \\ 3.5000 & 7.0000 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = \begin{vmatrix} 2.3333 - \lambda & 3.50 \\ 3.50 & 7.0 - \lambda \end{vmatrix} = 0$$

operando se obtiene la ecuación

$$\lambda^2 - 9.333\lambda + 4.08333 = 0$$

y resolviendo esta ecuación de segundo grado se obtienen los valores propios

$$\lambda_1 = 8.87314$$
$$\lambda_2 = 0.46019$$

Ahora se puede hallar el primer vector propio sustituyendo valores en la ecuación matricial anteriormente expuesta

$$\left[\begin{pmatrix} 2.333 & 3.50 \\ 3.50 & 7.00 \end{pmatrix} - 8.87314 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = 0$$

dando el sistema de ecuaciones

$$-6.5398\, a_{11} + 3.50\, a_{12} = 0$$
$$3.500\, a_{11} - 1.87314\, a_{12} = 0$$

cuyas soluciones son

$$a_{11} = 0.471858$$
$$a_{12} = 0.881675$$

El segundo vector propio se hallaría de la misma manera, usando el segundo valor propio.

Número de componentes principales a retener.-

En el análisis de componentes principales se obtienen tantas componentes como variables hay, pero como se dijo al principio, uno de los objetivos de este análisis es reducir el número de variables. Puesto que los componentes principales se ordenan en orden decreciente de sus varianzas, hay que elegir unas pocas como representativa de todas las variables originales.

El primer criterio debe ser el de la proporción de la varianza total explicada por las *m* componentes retenidas. La proporción acumulada de la varianza total indica la cantidad de información retenida con las componentes elegidas.

En el ejemplo, la varianza total es 9.33 y la varianza de la primera componente es 8.87314, lo cual es 8.87314/9.33=0.9507, el 95.1% de la varianza total. Se puede argumentar que esta cantidad es un porcentaje de la varianza total suficiente como

para que la primera componente principal sea razonablemente representativa de las dos variables originales.

Otro criterio puede ser el de coger las componentes cuyos valores propios sean mayores que el mayor elemento de la diagonal de la matriz de varianzas-covarianzas, esto es, elegir las componentes principales cuyas varianzas sean mayores que la mayor de la varianza de las variables originales. Existen otros criterios poco utilizados por la falta de consenso sobre su idoneidad.

Existe un método que puede ser de utilidad por su visualidad, este es un método grafico consistente en representar los valores propios en una gráfica en la que la Y es la magnitud del valor propio y la X es el número del valor propio. Si se unen estos puntos con trazos se observa que hay un punto de inflexión (un cambio de una pendiente a otra), los valores que están antes que esta inflexión son los que hay que conservar.

Interpretación de las componentes.-

Para interpretar el significado de la primera componente principal, recuerdes que esta era

$$C_1 = 0.471858x_1 + 0.881675x_2$$

El coeficiente 0.471858 puede transformarse en la correlación entre x_1 y C_1. En general, la correlación entre la *i-ésima* componente y la *j-ésima* variable es

$$r_{ij} = \frac{a_{ij}\sqrt{\lambda_i}}{\sqrt{s_j^2}}$$

siendo a_{ij} el coeficiente de x_j de la *i-ésima* componente. La transformación de todos los coeficientes (la matriz de vectores propios) en correlaciones entre las variables y las componentes da la denominada ***matriz del modelo factorial***

Siguiendo con el ejemplo, la correlación entre la primera componente y x_1 es

$$r_{ij} = \frac{0.471858\sqrt{8.87314}}{\sqrt{2.33333}} = 0.92016$$

y la correlación entre la primera componente y x_2 es

$$r_{ij} = \frac{0.881675\sqrt{8.87314}}{\sqrt{7}} = 0.99265$$

estos dos son los elementos del primer vector de la ***matriz del modelo factorial***

Se puede calcular de la misma manera el segundo vector de la matriz factorial que dará las correlaciones de las variables con la segunda componente principal, *etc.*, obteniendo al final la matriz del modelo factorial, que la del ejemplo es la siguiente

$$\begin{pmatrix} 0.92016 & 0.39155 \\ 0.99265 & -0.12098 \end{pmatrix}$$

Esta matriz del modelo factorial puede representarse en unos ejes de coordenadas en el que los puntos serán las variables y, por lo tanto, se mostrará la magnitudes y sentidos de las correlaciones entre las variables originales y las componentes. Observando esta gráfica a la vez que se observa la de los componentes principales se puede visualizar la relación existente entre los individuos, las variables y las componentes, si bien ambas gráficas, en este caso, no se pueden superponer como si se podrá hacer si se está trabajando con datos tipificados (ver más adelante).

Obsérvese que los resultados de la matriz del modelo factorial del ejemplo que se está desarrollando, da que las correlaciones con la primera componente son positivas y de valor elevado. Este es un resultado que se presenta con cierta frecuencia cuando se analizan medidas morfométricas y significa que la primera componente está correlacionada positivamente con todas las variables originales y que estas aumentan conforme aumenta la primera componente (ver más adelante el análisis de *tamaño y forma*).

Ejemplo.-

Se tiene el gasto anual medio que realizan 112 familias en siete productos o categorías de productos alimenticios. Las familias están clasificadas según el nivel profesional del padre y según el número de hijos, habiendo doce tipos en total: **T2** trabajador manual con dos hijos, **O2** empleado de oficina con dos hijos, y **D2** directivo con dos hijos; y los mismos niveles profesionales del padre para 3, 4 y 5 hijos. Las categorías de productos alimenticios son: pan, legumbres, fruta, carne, pollo, leche y vino.

Archivo de programa *SAS* (C18-1.SAS).-

```
title 'Análisis de Componente Principales';
option ls=75 ps=60;
data vino;
infile 'c18-1.dat';
input familia $ simfam $ pan legumbre fruta carne pollos leche vino @@;
proc princomp cov out=prefix;
var pan legumbre fruta carne pollos leche vino ;
run;
proc print;
var prin1 prin2;
run;
proc plot;
plot prin2*prin1=simfam / vspace =3 hspace=5;
run;
proc factor cov scree data=vino n=2 preplot;
var pan legumbre fruta carne pollos leche vino;
run;
```

la opción **cov,** en ambos procedimientos (**PRINCOMP** y **FACTOR**), se refiere a que se haga el análisis partiendo de la matriz de varianzas-**cov**arianzas.

Archivo de datos (C18-1.DAT).-

t2	A	248	383	370	1497	518	237	437	t4	I	553	664	401	1634	655	372	375
t2	A	326	390	307	1316	539	131	451	t4	I	530	695	406	1510	639	365	447
t2	A	212	389	317	1505	528	356	423	t4	I	553	661	293	1522	626	393	251
t2	A	266	448	328	1490	533	293	340	t4	I	500	671	318	1631	642	346	394
t2	A	289	459	364	1473	536	336	377	t4	I	623	717	322	1670	626	282	360
t2	A	337	597	346	1393	513	293	363	t4	I	604	668	316	1365	660	399	374
t2	A	338	289	306	1392	517	205	495	t4	I	423	708	398	1742	626	412	450
t2	A	424	359	457	1410	529	208	436	t4	I	506	661	453	1827	644	482	397
t2	A	313	545	311	1559	538	240	360	t4	I	497	661	367	1634	628	362	427
t2	A	347	371	387	1642	513	261	399	t4	I	584	668	294	1820	623	483	406
o2	B	245	503	315	1316	574	211	150	o4	D	525	711	401	1860	758	425	460
o2	B	350	697	421	1644	543	170	318	o4	D	528	783	543	1709	769	335	446
o2	B	233	689	343	1515	563	236	227	o4	D	501	620	590	1884	759	305	388
o2	B	262	590	457	1523	574	212	284	o4	D	538	682	506	2007	771	345	495
o2	B	260	508	349	1545	555	203	284	o4	D	478	833	524	1911	768	365	498
o2	B	266	488	365	1507	573	427	285	o4	D	407	624	501	2170	743	466	439
o2	B	314	474	365	1289	591	121	249	o4	D	434	516	499	1912	762	364	364
o2	B	265	565	397	1687	555	201	215	o4	D	419	684	529	2012	772	453	375
o2	B	291	488	360	1209	548	260	224	o4	D	450	759	454	1715	769	344	416
o2	B	309	658	438	1494	569	211	321	o4	D	437	716	515	2254	750	333	436
d2	V	320	727	591	2189	924	199	450	d4	X	406	612	663	2619	1145	228	301
d2	V	345	728	551	1810	931	111	405	d4	X	328	786	647	2417	1156	253	354
d2	V	323	778	462	2104	922	214	516	d4	X	395	760	670	2558	1139	253	289
d2	V	327	798	569	2263	924	254	424	d4	X	465	728	588	2531	1149	249	207
d2	V	346	701	552	1686	938	283	481	d4	X	374	858	582	2430	1163	264	217
d2	V	318	725	524	1800	919	282	459	d4	X	363	873	649	2323	1139	351	312
d2	V	397	753	620	1809	921	226	482	d4	X	403	725	674	2611	1132	230	209
d2	V	361	934	601	1922	933	303	421	d4	X	300	747	658	2483	1152	284	304
d2	V	404	662	606	1938	924	195	469	d4	X	302	777	615	2213	1142	274	239
d2	V	334	766	546	1983	944	304	538	d4	X	399	762	590	2481	1156	386	347
t3	E	345	685	377	1559	540	384	401	t5	O	649	856	456	1892	755	525	500
t3	E	463	455	372	1721	539	341	389	t5	O	657	833	386	1592	739	536	387
t3	E	466	523	342	1206	546	347	446	t5	O	617	634	446	2229	744	506	408
t3	E	348	617	192	1371	540	271	367	t5	O	711	854	376	1634	753	523	450
t3	E	372	433	385	1460	549	321	458	t5	O	619	726	352	1418	741	416	447
t3	E	453	643	367	1731	539	375	459	t5	O	628	685	483	1698	771	489	537
t3	E	257	532	386	1442	536	313	405	t5	O	611	819	479	2134	769	458	453
t3	E	404	488	343	1399	568	285	454	t5	O	544	809	479	1694	758	451	410
t3	E	379	662	359	1939	559	241	362	t5	O	528	827	392	1781	778	468	447
t3	E	315	512	347	1306	532	399	345	t5	O	748	987	359	1611	737	503	516
o3	C	366	458	368	1567	553	286	294	o5	F	632	840	605	2059	882	465	269
o3	C	411	566	390	1527	573	371	320	o5	F	590	1047	566	1969	890	545	341
o3	C	409	435	329	1796	582	251	459	o5	F	578	1056	542	2016	887	558	209
o3	C	375	667	509	1501	566	378	282	o5	F	641	1104	562	1711	903	390	323
o3	C	347	621	300	1773	541	418	408	o5	F	563	1020	581	2059	900	618	359
o3	C	350	548	287	1432	548	330	334	o5	F	560	1093	514	1805	890	513	321
o3	C	349	655	370	1426	559	264	425	o5	F	574	896	612	2081	905	425	358
o3	C	463	541	346	1726	568	212	309	o5	F	656	855	572	1962	889	537	355
o3	C	385	498	397	1525	570	310	372	o5	F	530	1097	564	2127	901	506	312
o3	C	458	714	404	1211	572	349	393	o5	F	576	991	378	2075	888	557	304
d3	W	511	997	649	2044	1134	169	291	d5	Z	504	1090	834	2704	1163	613	283
d3	W	371	788	699	2428	1149	252	344	d5	Z	497	1037	900	2672	1176	545	328
d3	W	414	665	603	2319	1150	340	314	d5	Z	474	965	906	2685	1156	599	272
d3	W	422	908	717	2556	1149	202	418	d5	Z	522	1017	890	2991	1156	597	286
d3	W	444	830	693	2199	1125	182	328	d5	Z	498	872	903	2661	1178	534	313
d3	W	481	784	719	2462	1150	175	286	d5	Z	496	971	896	2780	1153	497	202
d3	W	400	762	758	2620	1144	397	346	d5	Z	554	1192	835	2753	1165	578	316
d3	W	384	938	645	2264	1167	332	338	d5	Z	437	1057	950	2695	1182	563	338
d3	W	387	804	607	2276	1142	308	407	d5	Z	634	1076	964	2423	1183	449	260
d3	W	386	702	681	2639	1156	246	355	d5	Z	499	847	947	2607	1148	540	330

Los doce tipos de familia se han simbolizado: con las vocales A, E, I y O los trabajadores de 2, 3, 4 y 5 hijos respectivamente; con las primeras consonantes del abecedario, B, C, D y F la de los oficinistas con 2, 3, 4 y 5, respectivamente; y con las últimas consonantes del abecedario, V, W, X y Z la de los directivos con 2, 3, 4 y 5

hijos respectivamente

Las salidas que da este programa (**C18-1.LST**) son:

Proc PRINCOMP

Principal Component Analysis

120 Observations
7 Variables

Simple Statistics

	PAN	LEGUMBRE	FRUTA	CARNE	POLLOS	LECHE	VINO
Mean	436.9750000	719.8000000	504.6750000	1902.566667	802.8166667	351.1583333	366.6916667
StD	117.7969669	190.5390672	171.6043996	434.737837	239.5378949	123.3223552	82.4063304

Covariance Matrix

	PAN	LEGUMBRE	FRUTA	CARNE	POLLOS	LECHE	VINO
PAN	13876.1254	12610.2134	3734.7313	9923.0143	6107.5920	9482.6258	1393.1351
LEGUMBRE	12610.2134	36305.1361	21844.3210	52050.2908	32453.1479	13215.1832	-2920.0706
FRUTA	3734.7313	21844.3210	29448.0700	65158.7655	36484.5786	6196.9511	-4030.6641
CARNE	9923.0143	52050.2908	65158.7655	188996.9871	93779.7686	15513.6658	-9671.4373
POLLOS	6107.5920	32453.1479	36484.5786	93779.7686	57378.4031	6073.3486	-5197.9730
LECHE	9482.6258	13215.1832	6196.9511	15513.6658	6073.3486	15208.4033	26.6459
VINO	1393.1351	-2920.0706	-4030.6641	-9671.4373	-5197.9730	26.6459	6790.8033

Total Variance = 348003.92829

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	282129	249178	0.810705	0.81071
PRIN2	32951	19776	0.094686	0.90539
PRIN3	13175	6077	0.037860	0.94325
PRIN4	7098	1764	0.020396	0.96365
PRIN5	5334	985	0.015327	0.97897
PRIN6	4349	1380	0.012496	0.99147
PRIN7	2968	.	0.008530	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4
PAN	0.058502	0.520585	0.226809	0.425917
LEGUMBRE	0.261165	0.632189	-.403431	-.162596
FRUTA	0.295704	-.007173	-.229478	-.192637
CARNE	0.806593	-.272599	0.446485	0.002278
POLLOS	0.426984	-.023525	-.540535	0.304809
LECHE	0.078465	0.500288	0.469800	-.320916
VINO	-.043188	0.064226	0.144571	0.747761

. Número de observaciones o familias = 120
. Número de variables = 7
. La media (**Mean**) y la desviación típica (**Std**) de cada variable
. La matriz de varianzas-covarianzas (**Covariance Matrix**) . La varianza mayor es la de la carne que vale 188996.9871
. La varianza total (suma de las 7 varianzas)=348003.92829
. Los valores propios (**Eigenvalues**). El primer valor propio vale 282129 y explica una proporción de la varianza total del 81.07%. El segundo valor propio, que ya es mucho más pequeño que la mayor de las varianzas, vale 32951 y explica el 9.47% de la varianza total, Entre los dos primeros explican el 90.54% de la varianza total. Como esa proporción es razonablemente grande, se va a utilizar las dos primeras componentes.
. Los vectores propios (**Eigenvectors**) o coeficientes, a_{ij}, de las funciones lineales

Proc PRINT

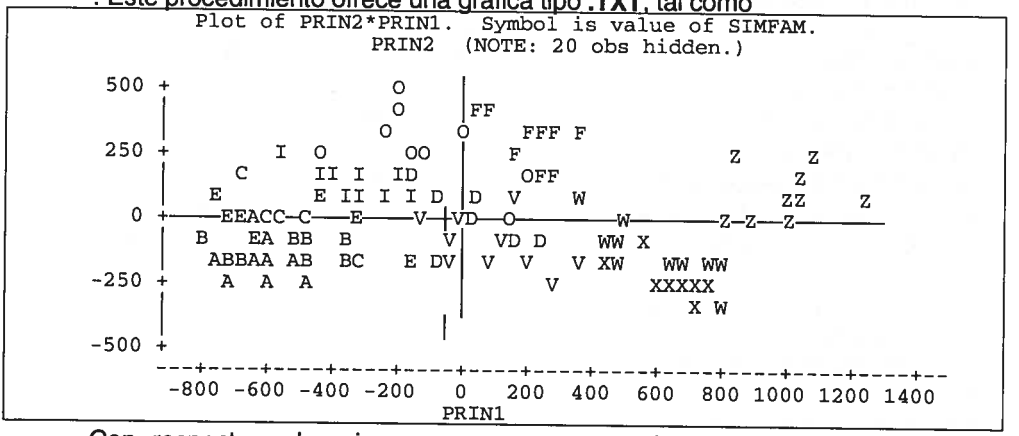
OBS	PRIN1	PRIN2	OBS	PRIN1	PRIN2	OBS	PRIN1	PRIN2
1	-599.573	-245.672	41	287.805	-208.668	81	-492.25	-141.390
2	-316.905	113.517	42	734.002	-354.593	82	245.61	182.970
3	-757.760	-203.474	43	-29.971	-138.499	83	-473.08	4.783
4	-419.184	156.407	44	611.588	-214.366	84	219.67	361.241
5	-595.120	-204.021	45	192.067	-138.978	85	-320.05	-203.285
6	-445.343	146.151	46	724.797	-238.300	86	257.57	346.140
7	-584.622	-171.566	47	365.378	-154.307	87	-435.39	57.353
8	-353.552	80.283	48	682.006	-221.643	88	22.42	415.178

9	-580.413	-124.445	49	-123.483	-30.487	89	-304.43	-30.260
10	-312.088	128.910	50	634.117	-151.439	90	297.29	342.916
11	-624.005	-12.148	51	-42.422	-62.232	91	-602.93	-30.741
12	-550.689	229.861	52	563.373	-68.804	92	80.62	402.490
13	-727.917	-241.424	53	0.626	-33.133	93	-559.75	9.990
14	-239.277	69.749	54	758.718	-287.218	94	279.49	167.294
15	-637.527	-160.961	55	145.117	66.207	95	-343.20	-118.018
16	-146.405	90.872	56	659.14	-259.276	96	167.86	273.005
17	-508.799	-129.812	57	76.59	-138.497	97	-500.22	-78.353
18	-345.578	81.684	58	434.35	-174.302	98	358.96	296.926
19	-473.543	-231.685	59	132.32	-62.534	99	-687.73	202.561
20	-201.954	140.524	60	654.98	-143.526	100	235.82	294.677
21	-696.400	-154.394	61	-440.46	49.507	101	363.87	70.405
22	-79.512	96.598	62	12.52	296.368	102	1022.37	165.244
23	-367.393	-76.253	63	-367.72	-100.852	103	636.24	-195.066
24	-142.098	137.637	64	-256.78	366.897	104	1000.10	104.909
25	-485.794	-79.790	65	-773.06	90.800	105	498.95	-177.963
26	-35.315	-46.006	66	219.32	32.414	106	990.32	67.720
27	-469.438	-138.883	67	-671.79	1.940	107	772.01	-147.937
28	61.055	6.126	68	-214.98	394.120	108	1248.04	42.180
29	-513.980	-201.042	69	-585.75	-96.888	109	449.95	-103.526
30	25.030	106.661	70	-448.50	270.918	110	949.72	-2.419
31	-519.553	-88.613	71	-312.98	31.609	111	671.86	-192.013
32	168.197	-85.539	72	-179.45	214.012	112	1060.58	1.686
33	-711.008	-168.873	73	-584.73	-96.369	113	812.61	-176.380
34	-63.773	-125.708	74	205.39	150.194	114	1088.44	226.955
35	-367.248	-206.890	75	-625.67	-47.243	115	542.15	-9.162
36	79.534	-9.787	76	-159.44	222.932	116	1038.69	89.195
37	-781.078	-81.206	77	-144.72	-125.285	117	490.23	-102.299
38	-172.405	83.438	78	-102.95	213.300	118	834.77	215.742
39	-481.750	-61.392	79	-700.16	-2.200	119	781.57	-301.475
40	258.556	-101.654	80	-212.91	498.464	120	899.63	1.501

. La primera componentes (**PRIN1**) y la segunda componentes (**PRIN2**). Se han elegido dos en función de la varianza absorbida. Esta salida puede servir para ser tomada por cualquier paquete de gráficos y realizar una gráfica de calidad de las componentes principales.

Proc PLOT

. Este procedimiento ofrece una gráfica tipo **.TXT**, tal como



Con respecto a la primera componente, se observa que los trabajadores manuales (los simbolizados con vocales) están a la izquierda (valores negativos), mientras que los directivos (simbolizados con consonantes del final del abecedario) están a la derecha (valores positivos) y los oficinistas (simbolizados con consonantes del principio del abecedario) están hacia el centro izquierda, esto es, se entremezclan más con los trabajadores manuales que con los directivos. Por lo observado, esta

primera componente indica el *nivel de consumo relacionado con la categoría profesional*, los trabajadores manuales están en el extremo inferior, los directivos están en el extremo superior y los oficinistas ocupan una posición intermedia más cercana a los trabajadores manuales.

Con respecto a la segunda componente, se observa que las familias con menos hijos tienen valores más negativos (hacia abajo), mientras que las familias con más hijos tienen valores más positivos. También se observa que la diferencia de los valores de esta segunda componente entre las familias que tiene menos y más hijos, son mayores en trabajadores manuales que en directivos en los cuales las diferencias entre los tamaños de familia son mínimas. Por lo observado, esta segunda componente indica el *nivel de consumo relacionado con el número de hijos*, las familias pequeñas están en el extremo inferior y las familias mayores en el extremo superior. Y también se observa que esta segunda componente afecta más a los trabajadores manuales y oficinistas que a los directivos.

Proc FACTOR: da información redundante con el Proc PRINCOMP. Más adelante se estudiará el Análisis Factorial, lo que interesa de esta salida para el análisis de componentes principales es la matriz del modelo factorial (**Factor Patern**) y su representación gráfica (**Preplot of Pattern for FACTOR1 and FACTOR2**). Se han elegido dos factores (**n**=2) en función del porcentaje de varianza absorbida y en función de la representación de los valores propios (**SCREE**), donde se observa que los dos primeros valores los une una line casi vertical y los cinto restantes los une una linea casi horizontal..

Con respecto a la primera componente, se observa en la matriz del modelo factorial que la correlación mayor es con la variable **CARNE**, que junto con las variables **FRUTA** y **POLLO**, tienen una correlación superior al 0.90. Las correlaciones menores corresponden a las variables **PAN**, **LECHE** y **VINO**, si bien, el pan y la leche están muy correlacionadas con la segunda componente. También se observa todas las correlaciones con la primera componente son positivas menos la del vino que es negativa, esto es, el vino esta correlacionado negativamente con la primera componente y las demás variables, Esto significa que conforme aumenta la primera componente (que se ha identificado con el *nivel de consumo relacionado con la categoría profesional*), aumenta el consumo de legumbre, fruta, carne y pollo, y con menos intensidad aumenta el consumo de pan y la leche, mientras que disminuye el consumo de vino. Por lo que esta primera componente separa los grandes consumidores de los pequeños consumidores en función de su nivel adquisitivo.

Con respecto a la segunda componente se observa en al matriz del modelo factorial que las correlaciones mayores son con el pan y la leche, aunque también son bastantes significativas las correlaciones con la legumbre y no tanto con el vino, mientras que las correlaciones con la fruta, carnes y pollo son negativas, cercanas a cero. Esto significa que conforme aumenta la segunda componente (que se ha identificado con el *nivel de consumo relacionado con el número de hijos*), aumenta sobre todo el consumo de pan y leche y en menor medida aumenta el consumo de legumbres y vino, mientras que el consumo de fruta, carne y pollo disminuye levemente.

. La salida de la matriz factorial puede ser tomada por un paquete gráfico para su representación. Con la opción **preplot** el *SAS* ofrece una gráfica tipo TXT, tal como

```
                Initial Factor Method: Principal Components

              Plot of Factor Pattern for FACTOR1 and FACTOR2

                              FACTOR2
                                 1
                                .9
                                .8       Pan
                                .7         Lec
                                .6                    Leg
                                .5
                                .4
                                .3                                  F
                                .2                                  A
                          Vin   .1                                  C
-1-.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 Fru1.0T
                               -.1                        PolCarO
                               -.2                               R
                               -.3                               1
                               -.4
                               -.5
                               -.6
                               -.7
                               -.8
                               -.9
                                -1
```

Si observamos en paralelo estas dos gráficas, se observa que el consumo de legumbres, frutas, carne y pollos se asocia con los directivos mientras que el consumo de leche y pan se asocia con los trabajadores manuales y oficinistas con muchos hijos y el consumo de vino se asocia con los trabajadores manuales y oficinistas con pocos hijos.

Usando variables tipificadas.-

Hay veces que se prefiere estandarizar las variables originales antes de realizar las componentes principales. Estandarizar (o tipificar) se sabe que es restarle a cada valor la media y dividirlo por la desviación típica

$$z_{ij} = \frac{X_{ij} - \overline{X}_i}{s_i^2}$$

de manera que los valores del primer individuos son

$$z_1 = \frac{5 - 5.3333}{1.5275} = -2.2182$$

$$z_2 = \frac{14 - 13}{2.6457} = 0.3780$$

Esta operación tiene el efecto de que las medias de las variables tipificadas valen cero y las varianzas valen uno, sean cuales sean las escalas en las que están tomadas las variables, pero no cambia los valores de las correlaciones entre las

variables.

Por lo demás, se cumple todo lo dicho anteriormente para las variables no tipificadas, teniendo en cuenta que las varianzas valen uno, así la media y la varianza de las dos componentes principales del ejemplo son

$$\overline{C_1} = \overline{C_2} = 0$$

$$Var(C_1) = a_{11}^2 + a_{12}^2 + 2\,a_{11}a_{12}\,r$$

$$Var(C_2) = a_{21}^2 + a_{22}^2 + 2\,a_{21}a_{22}\,r$$

$$Var(C_1) + Var(C_2) = s_1^2 + s_2^2 = 1+1 = 2$$

puesto que $s_i^2=1$.

Para el cálculo de los vectores propios se tiene que satisfacer los mismos requerimientos expuestos anteriormente, estos son

1. El valor de la varianza de C_1 tiene que ser el máximo posible.
2. Los N valores de C_1 y C_2 tienen que estar incorrelacionados
3. $a^2_{11}+a^2_{12}=a^2_{21}+a^2_{22}=1$

La manera de hallar estos coeficientes es la misma de la expuesta anteriormente en el epígrafe *Cálculo matemático*, solo que hay que partir de la matriz de correlación en vez de la matriz de varianzas-covarianzas, esto es, partiendo de

$$R = \begin{pmatrix} 1.0000 & 0.8660 \\ 0.8660 & 1.0000 \end{pmatrix}$$

no hay más que cambiar **R** por **V** y realizar todas las operaciones tal como se ha expresado en el epígrafe *Cálculo matemático* y se obtiene los valores y los vectores propios para los datos tipificados.

Siguiendo con el ejemplo que se está desarrollando, los dos componentes principales de los datos tipificados son

$$C_1 = 0.707107\,x_1 - 0.707107\,x_2$$

$$C_2 = 0.707107\,x_1 + 0.707107\,x_2$$

obsérvese que

$$0.707107^2 + (-0.707107)^2 = 1$$

$$0.707107^2 + 0.707107^2 = 1$$

que es el requerimiento número tres. Nótese así mismo, que para un caso como este, en el que hay solo dos variables, $a_{11}=a_{22}$ y $a_{12}=-a_{21}$.

Una vez obtenidos los vectores propios se pueden calcular las dos

componentes principales de cada individuo

$$C_{11} = 0.707107 \times -0.21822 + 0.707107 \times 0.3779 = 0.11296$$
$$C_{12} = 0.707107 \times -1.87287 + 0.707107 \times -1.13389 = -1.41900$$
$$C_{13} = 0.707107 \times 1.09108 + 0.707107 \times 0.75593 = 1.30604$$

$$C_{21} = -0.707107 \times -0.21822 + 0.707107 \times 0.3779 = 0.42156$$
$$C_{22} = -0.707107 \times -1.87287 + 0.707107 \times -1.13389 = -0.18457$$
$$C_{23} = -0.707107 \times 1.09108 + 0.707107 \times 0.75593 = -0.23699$$

En el caso de variables tipificadas, y si se cumple el supuesto de multinormalidad, la representación de las componentes principales dará una elipse que necesariamente estará dentro de los valores –3,+3, esto es, ± tres veces la desviación típica, con un 99.9% de confianza.

La varianza de las componentes principales o valores propios son

$$Var(C_1) = 0.707107^2 + 0.707107^2 + 2 \times 0.707107 \times \times 0.707107 \times 0.8663 = 1.86603$$
$$Var(C_2) = -0.707107^2 + 0.707107^2 + 2 \times -0.707107 \times \times 0.707107 \times 0.8663 = 0.13397$$

Compruébese que la varianza de la primera componente es mucho mayor que la varianza de la segunda componente y que la suma de estas dos varianzas (=2) o valores propios es igual a la suma de las varianzas de las dos variables originales tipificadas, esto quiere decir que la varianza total se mantiene en el proceso de rotación de los ejes pero que el primer eje absorbe un máximo de la varianza total, de manera que se puede decir que el primer eje o componente explica el 93% de la varianza total

$$\frac{1.86603}{2} = 0.93301$$

Estas ideas básicas son fácilmente extensibles al caso de p variables x_1, x_2,..., x_p. Cada componente principal es una combinación lineal de las variables originales. Los coeficientes de estas funciones lineales se eligen de manera que cumplan

1. $Var\ C_1 \geq Var\ C_2 \geq ... \geq Var\ C_p$

2. La varianza total es simplemente el número de variables, P, y la proporción de la varianza explicada por cada componente es su valor propio dividido por el número de variables.

3. Los valores de todas las componentes están incorrelacionados.

4. La suma de cuadrados de los coeficientes (vectores propios) de todas las componentes principales valen uno.

En otras palabras, y resumiendo, C_1 es la combinación lineal que tiene varianza máxima. C_2 es la combinación lineal que, cumpliendo la condición de no estar

correlacionada con C_1, tiene varianza máxima. Similarmente, C_3 es la combinación lineal que, cumpliendo la condición de no estar correlacionada ni con C_1 ni con C_2, tiene varianza máxima, *etc.* La varianza de C_j es el *j-ésimo* valor propio. Los *p* vales propios suman lo mismo que la suma de las varianzas originales. Los coeficientes de las combinaciones lineales son los vectores propios.

Cálculo matemático para estimar los valores y vectores propios de los datos tipificados.-

Ya se ha dicho anteriormente que el cálculo matemático para hallar los valores y los vectores propios de los datos tipificados es exactamente el mismo que el explicado para los datos sin tipificar, cambiando únicamente que la matriz de partida no es la matriz de varianzas-covarianzas (**V**) sino la matriz de correlaciones (**R**)

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8660 \\ 0.8660 & 1.0000 \end{pmatrix}$$

como se ve, en la diagonal van los valores de la correlación de cada variable consigo misma, que lógicamente es 1, y en los demás miembros van el resto de las correlaciones

Los valores propios se obtienen resolviendo el determinante

$$|\mathbf{R} - \lambda \mathbf{I}| = 0$$

Para el ejemplo

$$\begin{vmatrix} 1.0000 & 0.8660 \\ 0.8660 & 1.0000 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = \begin{vmatrix} 1-\lambda & 0.8660 \\ 0.8660 & 1-\lambda \end{vmatrix} = 0$$

operando se obtiene la ecuación

$$\lambda^2 - 2\lambda + 0.24999 = 0$$

y resolviendo esta ecuación de segundo grado se obtienen los valores propios

$$\lambda_1 = 1.86603$$
$$\lambda_2 = 0.13397$$

como se observa su suma vale dos, que es el valor de la traza de la matriz de correlaciones o el número de variables analizadas.

Ahora se puede hallar el primer vector propio sustituyendo valores en la ecuación matricial

$$(\mathbf{R} - \lambda_i \mathbf{I})\, \mathbf{a_i} = 0$$

dando

$$\left[\begin{pmatrix} 1.0000 & 0.8660 \\ 0.8660 & 1.0000 \end{pmatrix} - 1.86603 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = 0$$

de donde se extrae el sistema de ecuaciones

$$-0.86603\,a_{11}+0.8660\,a_{12}=0$$
$$0.8660\,a_{11}-0.86603\,a_{12}=0$$

cuyas soluciones son

$$a_{11}=0.707107$$
$$a_{12}=0.707107$$

El segundo vector propio se hallaría de la misma manera, usando el segundo valor propio.

Número de componentes principales a retener.-

Lógicamente, los criterios son los mismos que los vistos para las componentes principales de los datos sin tipificar. El primer criterio es el de la proporción de la varianza total explicada por las *m* componentes retenidas. La proporción acumulada de la varianza total indica la cantidad de información retenida con las componentes elegidas.

En el ejemplo, la varianza total es 2 y la varianza de la primera componente es 1.86603, lo cual es 1.86603/2=0.933, el 93.3% de la varianza total. Se puede argumentar que esta cantidad es un porcentaje de la varianza total suficiente como para que la primera componente principal sea razonablemente representativa de las dos variables originales.

Otro criterio puede ser el de coger las componentes cuyos valores propios sean mayores que la mayor de la varianza de las variables tipificadas, como estas valen todas 1, entonces el criterio es elegir las componentes cuyos valores propios sea superior a 1. Existen otros criterios poco utilizados por la falta de consenso sobre su idoneidad.

Existe un método que puede ser de utilidad por su visualidad, este es un método gráfico consistente en representar los valores propios en una gráfica en la que la *Y* es la magnitud del valor propio y la *X* es el número del valor propio. Si se unen estos puntos con trazos se observa que hay un punto de inflexión (un cambio de una pendiente a otra), los valores que están antes que esta inflexión son los que hay que conservar.

Interpretación de las componentes.-

Para interpretar el significado de la primera componente principal, recuerdes que esta era

$$C_1 = 0.707107\,X_1 + 0.707107\,X_2$$

El coeficiente 0.707107 puede transformarse en la correlación entre x_1 y C_1. En general, la correlación entre la *i-ésima* componente y la *j-ésima* variable tipificada es

$$r_{ij} = \frac{a_{ij}\sqrt{\lambda_i}}{\sqrt{s_j^2}} = a_{ij}\sqrt{\lambda_i}$$

siendo a_{ij} el coeficiente de x_j de la *i-ésima* componente. La transformación de todos los coeficientes (la matriz de vectores propios) en correlaciones entre las variables y las componentes da la denominada ***matriz del modelo factorial***

Siguiendo con el ejemplo, la correlación entre la primera componente y x_1 es

$$r_{ij} = 0.707107\sqrt{1.86603} = 0.96593$$

que es la misma que la correlación entre la primera componente y x_2 pues los coeficientes del vector propio son iguales. Estos dos valores son los elementos del primer vector de la ***matriz del modelo factorial***

Se puede calcular de la misma manera el segundo vector de la matriz factorial que dará las correlaciones de las variables con la segunda componente principal, *etc.*, siendo la matriz del modelo factorial del ejemplo la siguiente

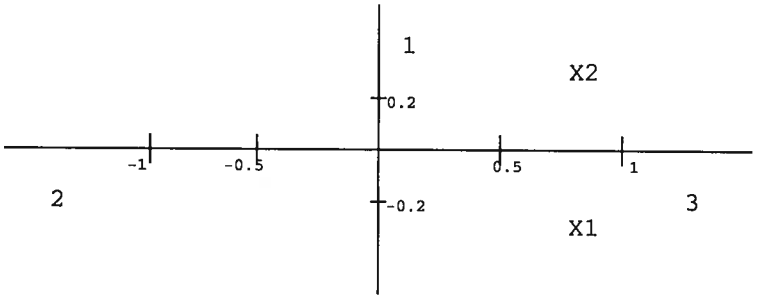
$$\begin{pmatrix} 0.96593 & -0.25882 \\ 0.96593 & 0.25882 \end{pmatrix}$$

Obsérvese que las correlaciones con la primera componente son positivas y de valor elevado. Este es un resultado que se presenta con cierta frecuencia y significa que la primera componente esta correlacionada positivamente con las variables originales y que estas aumentan conforme aumenta la primera componente (ver más adelante el análisis de *tamaño y forma*).

Representación conjunta de individuos y variables.-

La matriz del modelo factorial puede representarse en una gráfica en la que los puntos serán las variable y, por lo tanto, se mostrará la magnitud de la correlación de cada variable con los diferentes ejes factoriales o componentes principales. Pero, además, en el caso en que se estén usando las variables tipificadas se puede representar en la misma gráfica las variable y los individuos, esto es, se puede representar en la misma gráfica las componentes principales y la matriz del modelo factorial, lo que permitirá visualizar la relación entre los individuos, las variables y las componentes principales.

Siguiendo en el ejemplo que se esta desarrollando, la gráfica conjunta de variables e individuos sobre los ejes factoriales es



Como se ve, ambas variables tienen una alta y positiva correlación con la primera componente y una baja correlación con la segunda componente. Con respecto a los individuos y variables se observa que el individuo 3 está fuertemente influido o asociado a la variable X_1, mientras que el individuo 1 está influido por la variable X_2, mientras que el individuo 2 parece estar poco influido o asociado por las dos variables estudiadas.

Esta representación conjunta también puede realizarse con el Análisis de Correpondencias, tal como se verá a partir de la página de 961 en el capítulo ***DATOS CATEGÓRICOS.***

Ejemplo.-

Se tiene el gasto anual medio que realizan 112 familias en siete productos o categorías de productos alimenticios. Las familias están clasificadas según el nivel profesional del padre y según el número de hijos, habiendo doce tipos en total: **T2** trabajador manual con dos hijos, **O2** empleado de oficina con dos hijos, y **D2** directivo con dos hijos; y los mismos niveles profesionales del padre para 3, 4 y 5 hijos. Las categorías de productos alimenticios son: pan, legumbres, fruta, carne, pollo, leche y vino.

Archivo de programa *SAS* (C18-2.SAS).-

```
title 'Análisis de Componente Principales';
option ls=75 ps=60;
data vino;
infile 'c18-1.dat';
input familia $ simfam $ pan legumbre fruta carne pollos leche vino @@;
proc princomp out=prefix;
var pan legumbre fruta carne pollos leche vino ;
run;
proc print;
var prin1 prin2;
run;
proc plot;
plot prin2*prin1=simfam / vspace =3 hspace=5;
run;
proc factor data=vino n=2 preplot;
var pan legumbre fruta carne pollos leche vino;
run;
```

Como se observa, el único cambio con respecto al análisis de los datos sin tipificar es

que no se ha especificado cual es la matriz de partida, esto es, el *SAS* toma por defecto la matriz de correlación.

Recuérdese que los doce tipos de familia se han simbolizado: con las vocales A, E, I y O los trabajadores de 2, 3, 4 y 5 hijos respectivamente; con las primeras consonantes del abecedario, B, C, D y F la de los oficinistas con 2, 3, 4 y 5, respectivamente; y con las últimas consonantes del abecedario, V, W, X y Z la de los directivos con 2, 3, 4 y 5 hijos respectivamente

Las salidas que da este programa (**C18-2.LST**) son:

Proc PRINCOMP

Correlation Matrix

	PAN	LEGUMBRE	FRUTA	CARNE	POLLOS	LECHE	LECHE
PAN	1.0000	0.5618	0.1848	0.1938	0.2165	0.6528	0.6528
LEGUMBRE	0.5618	1.0000	0.6681	0.6284	0.7110	0.5624	0.5624
FRUTA	0.1848	0.6681	1.0000	0.8734	0.8876	0.2928	0.2928
CARNE	0.1938	0.6284	0.8734	1.0000	0.9005	0.2894	0.2894
POLLOS	0.2165	0.7110	0.8876	0.9005	1.0000	0.2056	0.2056
LECHE	0.6528	0.5624	0.2928	0.2894	0.2056	1.0000	1.0000
VINO	0.1435	-.1860	-.2850	-.2700	-.2633	0.0026	0.0026

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	3.78200	2.18026	0.540285	0.54029
PRIN2	1.60174	0.78337	0.228820	0.76911
PRIN3	0.81837	0.45330	0.116910	0.88602
PRIN4	0.36506	0.12244	0.052152	0.93817
PRIN5	0.24263	0.11855	0.034661	0.97283
PRIN6	0.12408	0.05795	0.017725	0.99055
PRIN7	0.06613	.	0.009447	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4
PAN	0.246757	0.608601	-.095809	-.533683
LEGUMBRE	0.448896	0.176205	-.068301	-.315736
FRUTA	0.460732	-.224348	0.165516	0.172966
CARNE	0.457218	-.224568	0.193210	0.193612
POLLOS	0.463972	-.235516	0.227462	-.162813
LECHE	0.278843	0.529010	-.289114	0.712559
VINO	-.152505	0.403080	0.886625	0.117650

. Número de observaciones o familias = 120 (como en C18-1.lst)
. Número de variables = 7 (como en C18-1.lst)
. La media (**Mean**) y la desviación típica (**Std**) de cada variable (como en C18-1.lst).
. La matriz de correlaciones (**Correlation Matrix**) cuya traza es igual al número de variables, es decir, siete, que será la suma de los valores propios.
. Los valores propios (**Eigenvalues**). El primer valor propio vale 3.7820 y explica una proporción de la varianza total del 54.03%. El segundo valor propio, que aún es mayor a uno, vale 1.6017 y explica el 22.88% de la varianza total, Entre los dos primeros explican el 76.91% de la varianza total. Como esa proporción es razonablemente grande, se va a utilizar las dos primeras componentes.
. Los vectores propios (**Eigenvectors**) o coeficientes, a_{ij}, de las funciones lineales

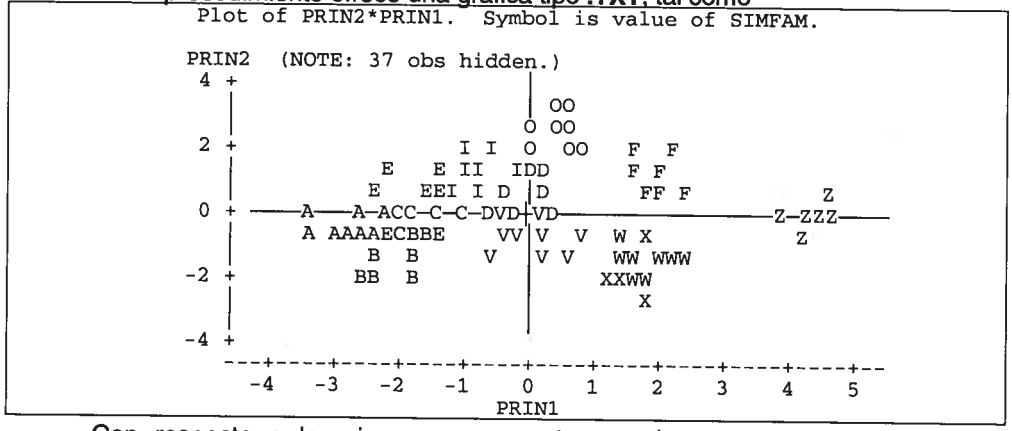
Proc PRINT

OBS	PRIN1	PRIN2	OBS	PRIN1	PRIN2	OBS	PRIN1	PRIN2
1	-2.91737	-0.76800	41	0.04145	-1.22288	81	-1.98201	-0.64622
2	-0.70378	1.09749	42	1.36559	-2.02286	82	1.71717	0.83943
3	-3.32190	-0.58955	43	-0.51196	-1.44920	83	-1.43349	0.15018
4	-0.97598	1.40273	44	1.33648	-1.78401	84	1.98064	1.59883
5	-2.79818	-0.45113	45	-0.35996	-0.55851	85	-2.13831	0.11579
6	-0.89782	0.80583	46	1.71298	-1.86602	86	2.22957	0.96531
7	-2.51146	-0.80536	47	0.41462	-1.02176	87	-0.90621	-0.23352
8	-1.14344	0.93439	48	1.69775	-1.84089	88	1.64770	1.28781
9	-2.32403	-0.35215	49	-0.43938	-0.30358	89	-1.53947	0.52449
10	-0.83844	1.16192	50	1.73359	-2.03131	90	2.14651	1.75962
11	-2.14669	-0.14201	51	-0.39512	-0.54154	91	-2.14717	-0.08071
12	-1.02607	1.72078	52	1.78766	-1.24512	92	1.67873	1.40397
13	-3.41419	-0.10462	53	-0.06177	-0.36734	93	-1.97698	0.05884
14	-0.87144	0.98162	54	1.79625	-2.37275	94	1.55893	0.81218
15	-2.50556	-0.08979	55	0.66731	-0.39935	95	-1.64115	-0.38050
16	-0.28002	1.27450	56	1.43977	-2.12086	96	1.62937	1.79318
17	-2.30473	-0.62069	57	-0.20364	-0.66317	97	-1.87144	-0.05952
18	-1.09059	1.08784	58	1.19354	-2.23811	98	2.12034	0.93611
19	-2.42521	-0.44274	59	-0.06152	-0.08805	99	-1.46789	0.93848
20	-0.58946	1.96441	60	1.65662	-0.86163	100	1.51784	1.53820
21	-2.39815	-2.07754	61	-1.47707	0.40411	101	1.71405	-1.10020
22	-0.25199	1.42140	62	0.67697	2.73528	102	4.18422	0.20381
23	-1.55525	-0.98731	63	-1.69175	0.48175	103	1.58514	-1.67998
24	-0.00992	0.93048	64	0.33909	2.51201	104	3.97638	-0.03549
25	-1.80790	-1.61224	65	-2.22580	1.16305	105	1.26948	-1.15999
26	-0.12012	0.08584	66	0.52049	1.68295	106	3.97535	-0.25807
27	-1.80435	-1.54209	67	-2.51799	0.04467	107	1.90762	-1.24769
28	-0.07704	1.06378	68	0.40014	3.04043	108	4.44684	-0.03924
29	-2.32570	-1.51835	69	-2.32735	0.35096	109	1.40490	-1.49279
30	0.13424	1.02291	70	-0.64537	2.12782	110	3.59299	-0.30367

. La primera componente (**PRIN1**) y la segunda componente (**PRIN2**). Se han elegido dos por la cantidad de varianza absorbida. Esta salida puede servir para ser tomada por cualquier paquete de gráficos y realizar una gráfica de calidad de las componentes principales.

Proc PLOT

. este procedimiento ofrece una gráfica tipo **.TXT**, tal como



Con respecto a la primera componente, se observa que los trabajadores manuales (los simbolizados con vocales) están a la izquierda (valores negativos), mientras que los directivos (simbolizados con consonantes del final del abecedario) están a la derecha (valores positivos) y los oficinistas (simbolizados con consonantes del principio del abecedario) están hacia el centro izquierda, esto es, se entremezclan más con los trabajadores manuales que con los directivos. Por lo observado, esta

primera componente indica el *nivel de consumo relacionado con la categoría profesional*, los trabajadores manuales están en el extremo inferior, los directivos están en el extremo superior y los oficinistas ocupan una posición intermedia más cercana a los trabajadores manuales.

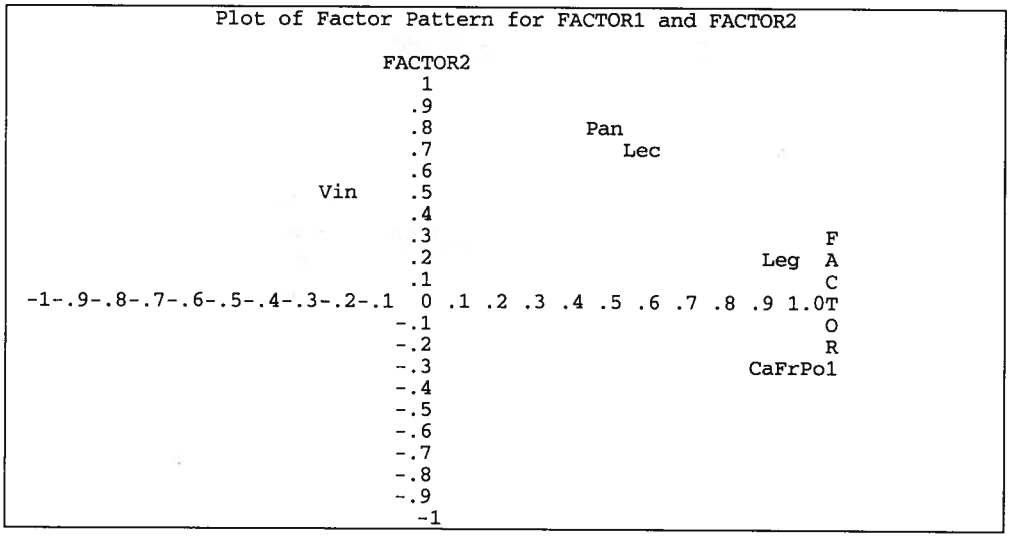
Con respecto a la segunda componente, se observa que las familias con menos hijos tienen valores más negativos (hacia abajo), mientras que las familias con más hijos tienen valores más positivos. También se observa que la diferencia de los valores de esta segunda componente entre las familias que tiene menos y más hijos, son mayores en trabajadores manuales que en directivos en los cuales las diferencias entre los tamaños de familia son mínimas. Por lo observado, esta segunda componente indica el *nivel de consumo relacionado con el número de hijos*, las familias pequeñas están en el extremo inferior y las familias mayores en el extremo superior. Y también se observa que esta segunda componente afecta más a los trabajadores manuales y oficinistas que a los directivos.

Proc FACTOR: da información redundante con el Proc PRINCOMP. Más adelante se estudiará el Análisis Factorial, lo que interesa de esta salida para el análisis de componentes principales es la matriz del modelo factorial (**Factor Patern**) y su representación gráfica (**preplot**), ademas de la representación gráfica de los valores propios (**scree**) para constatar que la elección de dos factores (**n**=2) es correcta pues los dos primeros valores propios tranzan una linea casi vertical y los cinco restantes una linea casi horizontal.

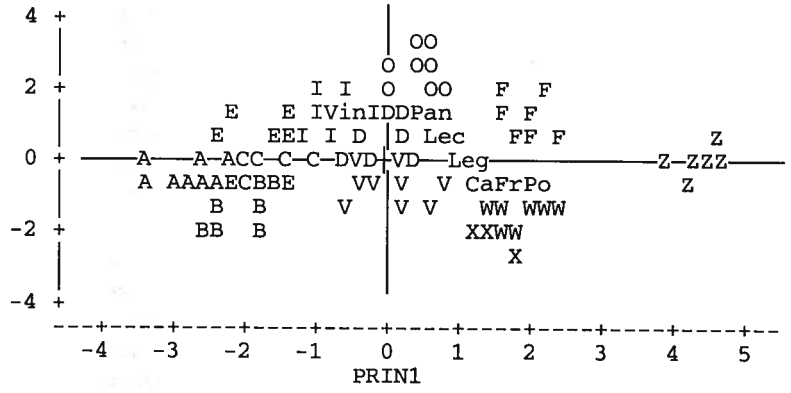
Con respecto a la primera componente, se observa en la matriz del modelo factorial que la correlación mayor es con la variable **POLLO**, que junto con las variables **LEGUMBRE**, **FRUTA** y **CARNE**, tienen una correlación próxima al 0.90. Las correlaciones menores corresponden a las variables **PAN**, **LECHE** y **VINO**, si bien, el pan y la leche están muy correlacionadas con la segunda componente. También se observa todas las correlaciones con la primera componente son positivas menos la del vino que es negativa, esto es, el vino esta correlacionado negativamente con la primera componente y las demás variables, Esto significa que conforme aumenta la primera componente (que se ha identificado con el *nivel de consumo relacionado con la categoría profesional*), aumenta el consumo de legumbre, fruta, carne y pollo, y con menos intensidad aumenta el consumo de pan y la leche, mientras que disminuye el consumo de vino. Por lo que esta primera componente separa los grandes consumidores de los pequeños consumidores en función de su nivel adquisitivo.

Con respecto a la segunda componente se observa en al matriz del modelo factorial que las correlaciones mayores son con el pan y la leche, aunque también es bastante significativa la correlación del vino, mientras que las correlaciones con la fruta, carnes y pollo son negativas. Esto significa que conforme aumenta la segunda componente (que se ha identificado con el *nivel de consumo relacionado con el número de hijos*), aumenta sobre todo el consumo de pan y leche y en menor medida aumenta el consumo de legumbres y vino, mientras que el consumo de fruta, carne y pollo disminuye.

. La salida de la matriz factorial puede ser tomada por un paquete gráfico para su representación. Con la opción **preplot** el *SAS* ofrece una gráfica tipo TXT, tal como



Esta gráfica se puede superponer a la gráfica de las componentes, con lo que quedaría una gráfica semejante a esta



Se observa que el consumo de legumbres, frutas, carne y pollos se asocia con los directivos mientras que el consumo de leche y pan se asocia con los trabajadores manuales y oficinistas con más hijos y el consumo de vino se asocia con los trabajadores manuales y oficinistas con menos hijos.

Ver la solución para este mismo ejemplo que se obtiene con el Análisis de Correspondencias en el capítulo DATOS CATEGÓRICOS.

Datos sin tipificar o datos tipificados.-

Como se ha observado con el ejemplo, los resultados cambian según se use la matriz de varianzas-covarianzas o la matriz de correlaciones. Estas diferencias se generalizan a cualquier ejemplo por lo que hay que plantearse cual método utilizar. Este es un tema controvertido aunque la mayoría de los investigadores prefieren usar

los datos tipificados porque de esta manera se unifica la unidad de medida de las diferentes variables. Si se hace así hay que tener en cuenta que la interpretación debe hacerse en términos de variables tipificadas.

Como regla se puede decir que si las unidades de medida son muy dispares (*Kg*, metros, meses, *etc.*) es preferible el uso de los datos tipificados (matriz de correlación) pues equivale a utilizar variables sin dimensión física. Pero si la unidades de medida son las mismas o razonablemente conmensurables, es preferible realizar el análisis sin tipificar los datos (matriz de varianzas-covarianzas) que es menos artificial y, por lo tanto, dará una mayor absorción de la varianza en menos componentes principales.

También, y dada la facilidad que hoy día ofrecen los paquetes estadísticos, puede ser recomendable utilizar ambos métodos y comparar las interpretaciones que se desprenden de ambos grupos de componentes principales.

ANÁLISIS FACTORIAL

El análisis factorial es un método de análisis multivariante que intenta detectar o explicar una (o pocas) variable hipotética, denominada factor, por medio de un modelo lineal en el que el factor (o pocos factores) es función de un conjunto extenso de variables observables. Este factor es, como en el Análisis de Varianza, un carácter de los individuos que se están estudiando, pero a diferencia del ANOVA, no existe para este factor una variable clara que pueda ser considerada como medida de dicho carácter o factor, por lo que hay que inferir sobre el a través de la combinación lineal de un conjunto de variables medidas.

El ejemplo clásico de utilización del análisis factorial es el de su primer usuario, *Spearman*, que lo utilizó para detectar o medir el llamado *factor de inteligencia* (*g*) a través de los resultados de una batería de tests psicotécnicos.

El análisis factorial es similar al de componentes principales en el sentido que ambos operan sobre *n* individuos a los que se les ha medido *p* variables con objeto utilizar funciones lineales de estas variables. Y ambos difieren del análisis de regresión en que no hay variables *dependiente* y variables *independientes*, sino que todas las variables medidas son variables *independientes* interelacionadas.

Pero ambos análisis son diferentes. En el análisis de componentes principales el objetivo es seleccionar un conjunto pequeño de componentes que expliquen un máximo de la varianza total de las variables. Los valores de las componentes principales de un individuo son relativamente fáciles de computar e interpretar. Por otro lado, el análisis factorial tiene como objetivo detectar unos *factores*, conocidos de antemano, a través de funciones lineales de las variables medidas. En el análisis de componentes principales se conserva toda la información de las medidas originales y se les asigna a los nuevos ejes utilizando el mismo criterio que se usa en el análisis factorial realizado por el método de los componentes principales: el primer eje explica más variabilidad que cualquier otro, y los ejes subsiguientes están situados en ángulo recto con respecto al resto de los ejes y contienen cantidades gradualmente decrecientes de información. En el análisis factorial propiamente dicho se decide de

antemano no incluir toda la información en los ejes factoriales. Sin embargo ambas técnicas desempeñan la misma función conceptual y sólo difieren en la forma en que se realizan los cálculos. En ambos análisis, el primer eje es la dimensión más adecuada para contener un máximo de información.

La idea que subyace en el análisis factorial es que si se tiene un conjunto grande de variables correlacionadas entre sí, estas relaciones reciprocas pueden deberse a la presencia de una o más variables subyacentes o factores correlacionados en grado diverso con las variables originales. Se supone que las correlaciones altas dentro de un grupo de variables se deben a que estas variables están correlacionadas con uno o pocos factores comunes. Uno de los objetivos del análisis factorial es identificar estos factores comunes, que expliquen unas correlaciones en principio inexplicables.

El análisis factorial, como el de componentes principales, esta justificado si existen correlaciones entre las variables medidas. Estas variables pueden ser, o no, de orígenes muy diferentes y no existe una explicación clara a estas correlaciones si no es la del factor hipotético que se trata de detectar con el análisis factorial. Si se encuentra un factor que es común a un grupo de variables, que en otro caso no tendrían relación aparente entre sí, tal factor puede considerarse como la causa común de las correlaciones observadas entre las variables originales.

De lo dicho hasta ahora queda claro que el análisis factorial pone en evidencia los factores comunes, pero no los identifica. Para identificar o dar nombres a estos factores es necesario hipotizar sobre ellos antes de realizar el análisis, sirviendo este para confirmar la existencia del factor hipotizado.

Ejemplo ilustrativo.-

Como se hizo con el análisis de componentes principales, se repasarán detenidamente los conceptos básicos con un ejemplo muy simple, supóngase que se toma una muestra aleatoria de diez individuos (n=10) a los que se les mide cinco variables (p=5)

Indi	X_1	X_2	X_3	X_4	X_5
1	3	22	40	16	32
2	4	20	42	17	30
3	8	23	44	20	45
4	6	25	39	14	36
5	2	20	38	16	31
6	6	23	45	19	40
7	7	29	47	18	45
8	9	32	39	19	39
9	4	23	41	14	38
10	5	24	42	11	33
$\overline{X}$	5.400	24.10	41.70	16.40	36.90
s^2	2.221	3.784	2.908	2.797	5.446

La matriz de correlación es

	X_1	X_2	X_3	X_4	X_5
X_1	1.00000	0.78782	0.38192	0.50797	0.77242
X_2	0.78782	1.00000	0.14438	0.23724	0.56452
X_3	0.38192	0.14438	1.00000	0.37161	0.66201
X_4	0.50797	0.23724	0.37161	1.00000	0.55528
X_5	0.77242	0.56452	0.66201	0.55528	1.00000

Como se observa existen correlaciones altas entre X_1 y X_2, entre X_1 y X_5, entre X_2 y X_5, y entre X_3 y X_5; una correlación media entre X_1 y X_4, entre X_2 y X_4, y entre X_4 y X_5, mientras que el resto de las correlaciones son relativamente bajas.

Conceptos básicos del Análisis Factorial.-

El análisis factorial parte siempre de las variables tipificadas

$$x_i = \frac{X_i - \overline{X}_i}{S_i}$$

por lo que sus medias valdrán cero, sus varianzas valdrán uno y sus correlaciones serán igual a sus covarianzas. En la terminología propia del análisis factorial a estas variables tipificadas se les denomina **variables respuesta.**

El objetivo del análisis factorial es representar cada una de esta variables respuesta como una combinación lineal de un pequeño grupo de *factores comunes* más un factor único. Este modelo lineal es

$$x_1 = l_{11} F_1 + \ldots + l_{1m} F_m + e_1$$
$$x_2 = l_{21} F_1 + \ldots + l_{2m} F_m + e_2$$
$$\ldots\ldots\ldots\ldots$$
$$x_p = l_{p1} F_1 + \ldots + l_{pm} F_m + e_p$$

donde se asume que

1. m es el número de factores comunes (típicamente este número es mucho menor que p).

2. F_1, F_2,..., F_m son los ***factores comunes***, que influyen en la p variables. Se asume que estos factores son variables tipificadas, esto es, de media cero y varianza uno.

3. l_{ij} es el coeficiente de F_j en la combinación lineal que describe la variable x_i. A este coeficiente se le denomina *factor de carga* de la *i-ésima* variable en el *j-ésimo* factor.

4. e_1, e_2,..., e_p son los factores únicos, porque cada uno esta relacionado exclusivamente con una variable original.

Estas ecuaciones y sus supuestos constituyen el llamado *modelo factorial.* Por tanto, cada una de las variables respuesta se componen de una parte debida a los factores comunes y otra parte debida a sus propios factores únicos. Se asume que la parte debida a los factores comunes es una combinación lineal de estos factores.

Como ejemplo, supóngase que x_1, x_2, x_3, x_4 y x_5 son las puntuaciones estandarizadas obtenidas por un individuo en cinco tests. Si m=2, se asume el siguiente modelo

$$x_1 = l_{11} F_1 + l_{12} F_2 + e_1$$
$$x_2 = l_{21} F_1 + l_{22} F_2 + e_2$$
$$\ldots\ldots\ldots\ldots$$
$$x_5 = l_{51} F_1 + l_{52} F_2 + e_5$$

Cada una de las cinco puntuaciones consisten en dos partes: una parte debida a los factores comunes F_1 y F_2 y una parte debida al factor único de éste test. Los factores únicos F_1 y F_2 deben ser considerados la habilidad verbal y la habilidad cuantitativa del individuo, respectivamente, Los factores únicos expresan la variación individual de la puntuación de cada test. El factor único incluye todos los demás efectos no incluidos en los factores comunes.

En una muestra de N individuos se pueden expresar las ecuaciones añádiendo un subíndice a cada x_i, F_j y e_i, para representar al individuo.

El modelo factorial es, en cierto sentido, la imagen especular del modelo de componentes principales, donde cada componente principal se expresa como una combinación lineal de las variables. También, el número de componentes principales es igual al número de variables originales (aunque no se usen todos). Por otro lado, en el análisis factorial se elige un número de factores menor que el número de variables respuesta. Como se ha dicho anteriormente, el número de factores debe ser conocido con anterioridad, aunque esto no ocurre siempre. Sin embargo, como se discutirá más adelante, es posible aceptar que sean los datos mismos los que determinen el número de factores.

El modelo factorial, al descomponiendo cada variable respuesta x_i en dos partes, también descompone la varianza de x_i en dos partes. Puesto que x_i está tipificada, su varianza vale 1. Esta varianza se compone de las siguientes dos partes

1. La *comunalidad.* esto es, la parte de la varianza debida a los factores comunes.

2. La *unicidad*, esto es, la parte de la varianza que es debida al factor único, e_i.

Si se simboliza la *comunalidad* de x_i por h_i^2 y la unicidad por u_i^2, se puede escribir la varianza de x_i de la siguiente manera

$$Var\ x_i = 1 = h_i^2 + u_i^2$$

Esto es, la varianza de x_i es igual a la comunalidad más la unicidad

Como en el análisis factorial lo que interesa es obtener factores comunes de modo que expliquen una buena parte de la variabilidad de las variables, interesa que la comunalidad sea lo mayor posible, por lo que se puede interpretar la comunalidad como el *coeficiente de determinación* del análisis factorial.

Los aspectos numéricos del análisis factorial se refieren a encontrar estimas de los *factores de carga* (a_{ij}) y la *comunalidad* (h_i^2). Existen varias maneras de calcular numéricamente estas cantidades. El proceso de solución se denomina *Extracción del Factor Inicial*, o *Método del Factor Inicial*. Se va ha ver dos de estos métodos: el método de *Componentes Principales* y el método del *Análisis Factorial Principal Iterado*. Una vez obtenidos los factores, el siguiente paso es obtener nuevos factores denominados *factores rotados*, con objeto de mejorar la interpretación de los factores.

En cualquier análisis factorial se requiere conocer el número *m* de factores comunes. Como se ha dicho anteriormente, este número debería conocerse antes de realizar el análisis. Si no se conoce el número de factores, la mayoría de los investigadores (y paquetes estadísticos) usan el criterio que se vio al estudiar el análisis de componentes principales, este es, el de utilizar los factores cuyos valores propios tengan un valor superior a la unidad. También, dado que los resultados numéricos dependen grandemente del número *m* elegido, muchos investigadores realizan varios análisis cada uno con un valor diferente de *m* con objeto de conseguir nuevas perspectivas de sus datos.

Método del Factor Inicial: Análisis de Componentes Principales.-

La idea básica es elegir los primeros *m* componentes principales y modificarlos para ajustarlo al modelo factorial definido anteriormente. La razón de elegir las primeras *m* componentes principales, en vez de otras, es que estas son las que explican una mayor proporción de la varianza y por ello son las más importantes. Nótese que las componentes principales son también incorrelacionadas lo que es muy importante para su elección como factores.

Para satisfacer el supuesto de que las varianzas de los factores valgan la unidad, se divide cada componente principal por su desviación típica. Esto es, se define el *j-ésimo* factor F_j como

$$F_j = \frac{C_J}{\sqrt{var\ C_j}}$$

siendo C_j la *j-ésima* componente principal.

Para expresar cada variable x_i en términos de sus F_j, se debe recordar la

relación entre las variables x_i y sus componentes principales C_j. Específicamente

$$C_1 = a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1p}\,x_p$$

$$C_2 = a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2p}\,x_p$$

$$\ldots\ldots\ldots\ldots$$

$$C_p = a_{p1}\,x_1 + a_{p2}\,x_2 + \ldots + a_{pp}\,x_p$$

Se puede demostrar que este sistema de ecuaciones puede invertirse para expresar las x_i en función de las C_j,

$$x_1 = a_{11}\,C_1 + a_{12}\,C_2 + \ldots + a_{1p}\,C_p$$

$$x_2 = a_{21}\,C_1 + a_{22}\,C_2 + \ldots + a_{2p}\,C_p$$

$$\ldots\ldots\ldots\ldots$$

$$x_p = a_{p1}\,C_1 + a_{p2}\,C_2 + \ldots + a_{pp}\,C_p$$

Nótese que las filas del primer sistema de ecuaciones son las columnas del segundo sistema.

Puesto que $F_j = C_j/(\text{Var } C_j)^{1/2}$, se colige que

$$C_j = \frac{F_j}{\sqrt{\text{Var } C_j}}$$

y se puede expresar la *j-ésima* ecuación de la siguiente manera

$$x_1 = a_{1i}\,F_1(\text{Var } C_1)^{1/2} + a_{2i}\,F_2(\text{Var } C_2)^{1/2} + \ldots + a_{\pi}\,F_p(\text{Var } C_p)^{1/2}$$

Esta última ecuación se puede modificar teniendo en cuenta que:

1. Se ha usado la notación $l_{ij}=a_{ij}(\text{Var } C_j)^{1/2}$ para las m primeras componentes.

2. Se combinan los últimos P-m términos para expresar los resultados de e_i. Esto es,

$$e_j = a_{m+1,i}\,F_{m+1}(\text{Var } C_{m+1})^{1/2} + \ldots + a_{\pi}\,F_p(\text{Var } C_p)^{1/2}$$

Con esta manipulación se puede expresar cada variable x_i como

$$x_i = l_{i1}\,F_1 + l_{i2}\,F_2 + \ldots + l_{im}\,F_m + e_i$$

para i=1, 2, ..., P. En otras palabras, se ha transformado el modelo de componentes principales para producir el modelo factorial. Como ya se vio en el análisis de componentes principales, nótese que si las variables originales están tipificadas, el factor de carga l_{ij} es la correlación entre las x_i y F_j. Además, se puede demostrar que la comunalidad de x_i es $h_i^2=l_{i1}^2+l_{i2}^2+\ldots+l_{im}^2$.

Para el ejemplo simple que se está desarrollando, los valores propios (o Var C_i) de la matriz de correlación son

Var C_1 = 3.05898
Var C_2 = 0.99737
Var C_3 = 0.63441
Var C_4 = 0.17718
Var C_5 = 0.13206
total = 5.0000

Nótese que la suma vale 5 que es igual a P, el número de variables. Basándose en la consabida regla de seleccionar solo las componentes correspondientes a las de valores propios superior a 1, se tendría que seleccionar solo el primer factor, pero como el segundo esta cercano al uno, se va a seleccionar m=2. En este caso los factores de carga (o matriz factorial) así como la comunalidad y unicidad son

	Factores de Carga		***Comunalidad***	***Unicidad***
Variables	F_1	F_2	H_i^2	u_i^2
x_1	0.90750	-0.30075	0.9140	0.0860
x_2	0.72289	-0.63355	0.9239	0.0761
x_3	0.63758	0.62096	0.7921	0.2079
x_4	0.67166	0.32153	0.5545	0.4455
x_5	0.92478	0.12872	0.8718	0.1282
Varianza	3.05898	0.99737	$\sum h_i^2$=4.0563	$\sum h_i^2$=0.9437
explicada	61.18%	19.95%	81.13%	18.87%
Porcentaje				

Por ejemplo, la carga de x_1 sobre F_1 es l_{11}=0.90750 y sobre F_2 es l_{12}=-0.30075. Por lo que la primera ecuación del modelo factorial es

$$x_1 = 0.9075 \times F_1 - 0.30075 \times F_2 + e_1$$

La tabla anterior muestra que la varianza explicada por cada factor. Por ejemplo, la varianza explicada por F_1 es 3.05898 o el 61.18% del total de la varianza que vale 5.

La columna de la comunalidad, h_i^2, muestra la parte de la varianza de cada variable explicada por los factores comunes. Por ejemplo,

$$h_1^2 = 0.9075^2 + (-0.30075)^2 = 0.9140$$

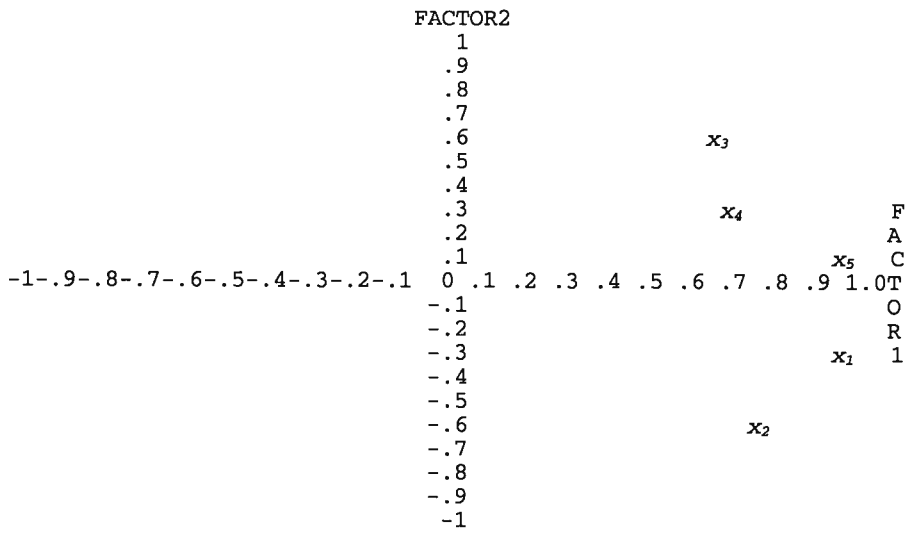
Finalmente, la unicidad u_i^2 es la parte de la varianza no explicada por los factores comunes. En este ejemplo se ha usado variables tipificadas cuyas varianzas valen uno, por lo que para cada x_i, u_i^2=1-h_i^2.

Como se ve en la tabla anterior, la suma de la comunalidad es igual a la suma de la varianza total explicada por los factores comunes. En el ejemplo

$$0.9140 + 0.9239 + 0.7921 + 0.5545 + 0.8718 = 4.0563$$
$$3.05898 + 0.99737 = 4.05635$$

Similarmente, la suma de las unicidades es igual a la varianza total menos la suma de las comunalidades.

Como se hizo con los componentes principales, representar gráficamente la matriz factorial ayuda a la interpretación, esta gráfica es



Puesto que los factores de carga son las correlaciones entre las variables tipificadas y los factores, el rango de valores en los ejes va de -1 a +1. Es posible ver en esta gráfica que las variables x_2 y x_3 cargan por igual en F_1 que en F_2, mientras que las demás variables cargan más sobre F_1 que sobre F_2. Pero puede ocurrir que estas cargas no estén claras por lo que más adelante se hará rotaciones con objeto de clarificar los resultados.

Método del Factor Inicial: Componentes Principales Iterados.-

El segundo método de extracción de los factores iniciales es una modificación del método de componentes principales. Para entender este método, se deber recordar que la comunalidad es la parte de la varianza de cada variable asociada con los factores comunes. El principio que subyace en la *solución iterada* es que se debe realizar el análisis factorial usando la comunalidad en vez de la varianza original. Este principio se consigue substituyendo la comunalidad estimada por los unos, que representan las varianzas de las variables tipificadas, de la diagonal de la matriz de correlaciones. Con unos en la diagonal del la matriz de correlaciones se esta factorizando el total de la varianza de las variables; con las comunalidades en la diagonal, se está factorizando la varianza asociada con los factores comunes. Por tanto, con las comunalidades en la diagonal se seleccionan los factores comunes que maximizan la comunalidad total.

Muchos autores consideran que maximizar la comunalidad total es un objetivo más atractivo que maximizar la proporción total de la varianza explicada, como se realiza en el método de componentes principales. El problema es que la comunalidad no es conocida antes de realizar el análisis. Por lo que deben obtenerse algunas estimas iniciales de la comunalidad. Para ello existen varios métodos y se recomienda,

en ausencia de una estima a priori, que se utilice la opción por defecto de los paquetes estadísticos.

Los pasos realizados por un paquete estadístico para proveer la extracción de los factores iterados se resumen en los siguientes:

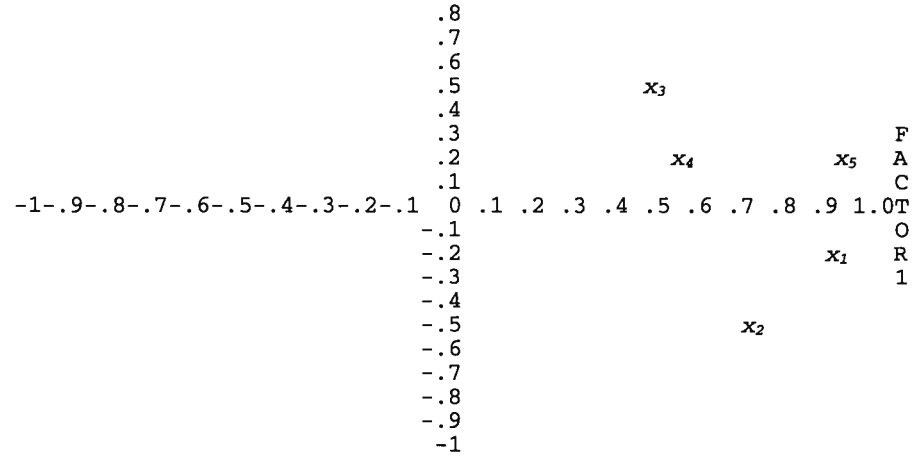
1. Encontrar las estimas de la comunalidades iniciales.
2. Substituir las comunalidades por los unos de la diagonal de la matriz de correlación.
3. Extraer *m* componentes principales de la matriz modificada.
4. Multiplicar los coeficientes de los componentes principales por la desviación típica de los respectivos componentes principales para obtener los factores de carga.
5. Computar las nuevas comunalidades de estos factores de carga.
6. Reemplazar las comunalidades del paso **2** con estas nuevas comunalidades y repetir los pasos **3**, **4** y **5**. Esto constituye una *iteración.*
7. Continuar las iteraciones hasta que se repitan las comunalidades en dos iteraciones seguidas.

Para el ejemplo que se está desarrollando, al usar este método da una varianza total de 3.527798, y los demás resultados son

	Factores de Carga		***Comunalidad***	***Unicidad***
Variables	F_1	F_2	H_i^2	u_i^2
x_1	0.91436	-0.24049	0.8939	0.1061
x_2	0.71962	-0.54321	0.8129	0.1871
x_3	0.55465	0.47924	0.5373	0.4627
x_4	0.54347	0.19416	0.3331	0.6669
x_5	0.94111	0.25445	0.9504	0.0496
Varianza explicada	2.84261	0.68502	$\sum h_i^2$=3.5276	$\sum h_i^2$=1.4724
Porcentaje	80.58%	19.42%	70.55%	29.45%

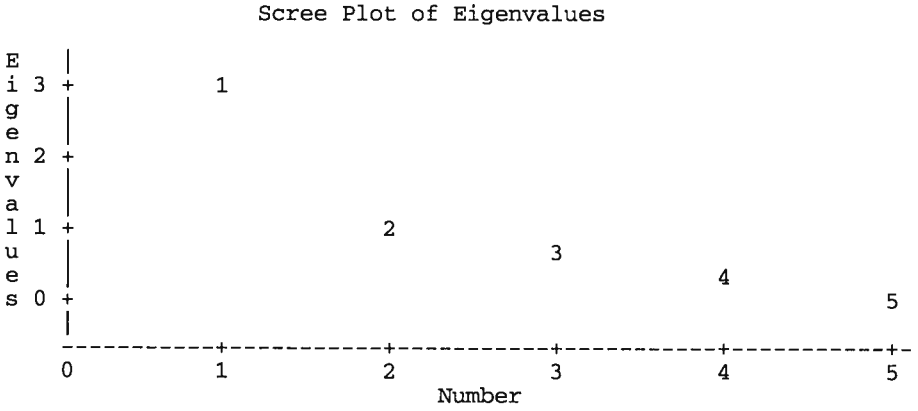
Comparando estos resultados con los de la tabla anterior se observa que la comunalidad total es mayor en el método de los componentes principales (el 81.13% contra el 70.55%). Este resultado es general puesto que en el método iterativo se ha factorizado la comunalidad total, lo cual es necesariamente menor que la varianza total. Los factores de carga no parecen muy diferentes, como se observa comparando la gráfica anterior con esta

FACTOR2
1
.9



Hay que hacer notar que los factores de carga extraídos por el método iterativo dependen del número de factores extraídos. Por ejemplo, las cargas del primer factor pueden depender de que se extraigan tres o dos factores comunes. Esta dependencia no existe en el método no iterativo.

Prescindiendo del método de extracción escogido, si el investigador no tiene una idea preconcebida del número de factores (*m*) procedente del conocimiento de la materia estudiada, entonces hay varios métodos para la elección. Ya se ha dicho varias veces que un criterio es el de elegir los factores que tienen un valor propio igual o superior a la unidad. Este criterio se basa en el teórico desarrollo lógico cuando se usa los verdaderos coeficientes de correlación de la población. Comúnmente se considera para producir un factor a partir de tres o cinco variables y se usa para estimar correctamente el número de factores cuando las comunalidades son altas y el número de variables no es muy grande. Como una alternativa a este método algunos autores utilizan el denominado *método scree*, consistente en representar los valores propios en el eje vertical y el número del valor en el horizontal y comprobar donde cambia la pendiente de la curva cuando se conectan los puntos entre ellos. Para el ejemplo, esta representación utilizando el método iterativo es

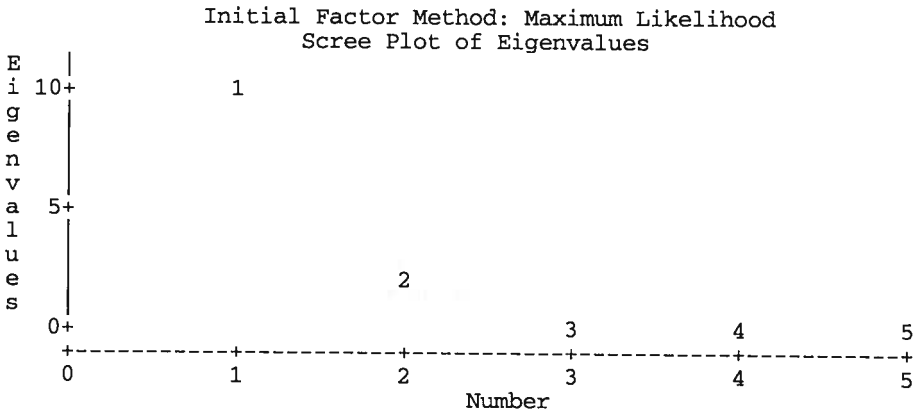


Como se observa a simple vista hay dos pendientes, una la de la recta que une los valores propios 1 y 2, y otra pendiente más suave para la recta que une los valores propios 2, 3, 4 y 5, por lo que parece razonable elegir m=2 factores.

En términos del método de extracción del factor inicial las soluciones dadas por el método iterativo son las más utilizadas por las ciencias sociales. Por razones teóricas, los estadísticos matemáticos están más de acuerdo con el método de componentes principales. Además de los dos estudiados aquí, existen otros métodos para resolver la extracción de factores que están en uso en la mayoría de los paquetes estadísticos. El *SAS*, por ejemplo además del método del componente principal. **PRINCIPAL**, que es el que hace por defecto, y el método del factor principal iterado, **PRINIT**, puede realizar el método alfa, **ALPHA**, el método de máxima verosimilitud, **ML**, el método mínimo cuadrado no ponderado, **ULS**, el método image, **IMAGE**, y los métodos **HARRIS** y **PATTERN**.

Método de máxima verosimilitud.-

Si el investigador está completamente seguro de que es válido el modelo factorial para explicar sus variables, y está completamente seguro de que las variables se ajustan escrupulosamente a una distribución normal multivariante, se debe usar el método de máxima verosimilitud (***ML***). Este método permite hacer pruebas de hipótesis tal como si el modelo factorial es válido y sobre cuantos factores son significativos. Si se utiliza este método para nuestro ejemplo, la opción **SCREE** dá como resultado la siguiente gráfica de los valores propios



Como se observa hay dos pendientes que se cruzan entre los valores 2 y 3, por lo que se tiene por un lado el primero y segundo factor y por otro lado el 3,4 y 5 factor,

Y el método **ML** realiza las siguientes pruebas estadísticas

```
 Initial Factor Method: Maximum Likelihood
Significance tests based on 10 observations:

   Test of H0: No common factors.
        vs HA: At least one common factor.

   Chi-square = 20.115    df = 10    Prob>chi**2 = 0.0282

   Test of H0: 2 Factors are sufficient.
        vs HA: More factors are needed.

   Chi-square = 0.472    df = 1    Prob>chi**2 = 0.4919
```

Es decir, la primera prueba es significativa por lo que se rechaza la hipótesis nula y se confirma que al menos hay un factor común. Y la segunda prueba es no significativa por lo que se acepta la hipótesis nula de que con dos factores es suficiente para explicar los datos.

Rotación de los factores.-

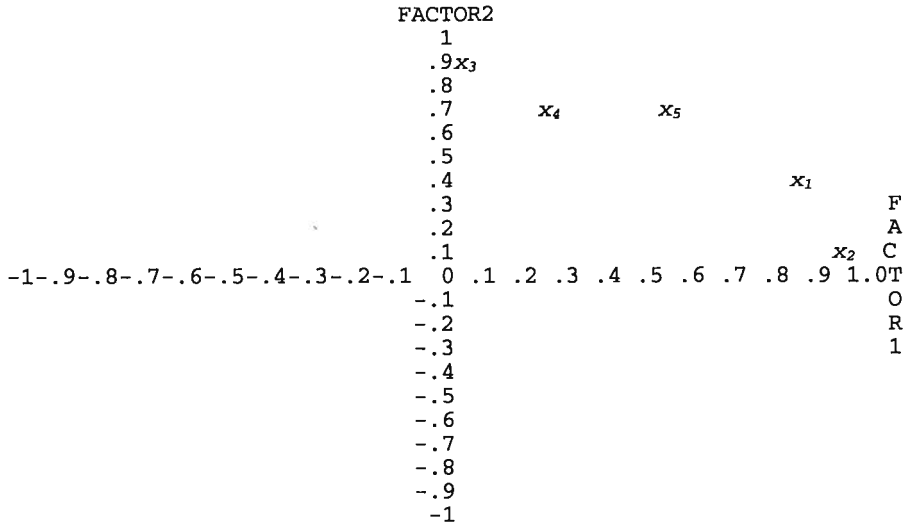
Recuérdese que el objetivo principal del análisis factorial es extraer de los datos factores comunes fácilmente interpretables. Pero ocurre con cierta frecuencia que los factores iniciales son difíciles de interpretar.

Afortunadamente, es posible encontrar nuevos factores cuyas cargas sean fáciles de interpretar. Estos nuevos factores, denominados *factores rotados*, son seleccionados de manera que algunos tienen cargas grandes (cerca de ± 1) y el resto tienen cargas pequeñas (cerca del cero). Recíprocamente, idealmente se podría hacer que, una variable dada, tenga una carga elevada en solo un factor. Si fuera este el caso, es fácil dar a cada factor una interpretación proveniente de las variables con las que esta fuertemente correlacionado (altas cargas).

Teóricamente, las rotaciones de los factores pueden realizarse de multitud de maneras. Todos los paquetes estadísticos traen varias, la más común es la rotación ortogonal ***VARIMAX***.

Rotación varimax.-

Esta consiste en encontrar nuevos ejes o factores. Estos nuevos ejes son seleccionados de manera que pasen a través de subgrupos de puntos que representan a las variables respuestas. La rotación varimax tiene la restricción de que los nuevos ejes sean también ortogonales (perpendiculares). La siguiente figura muestra los nuevos ejes rotados del método de los componentes principales



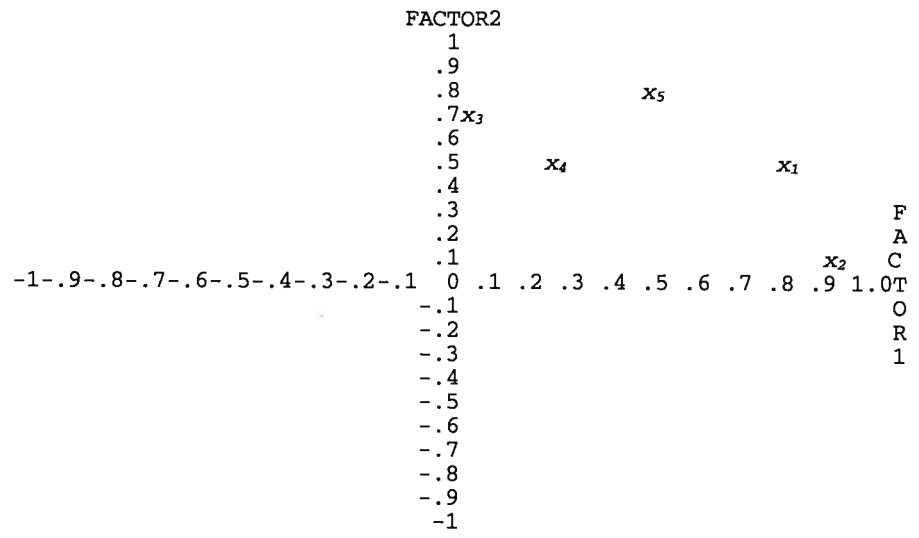
Compárese con la figura de los factores sin rotar y se ve que lo que ha ocurrido es que los ejes han rotado, manteniendo la perpendicularidad, en el sentido de las agujas del reloj hasta quedar el primer factor junto a la variable x_2 y el segundo factor junto a la variable x_3, de esta manera han quedado todos los puntos en el primer cuadrante, Esta gráfica muestra claramente que los factores rotados son ortogonales puesto que el ángulo entre los ejes de los factores rotados es de 90°. Estadísticamente el término de ortogonalidad de los factores rotados equivale a que los factores están incorrelacionados entre ellos.

El resultado de esta rotación varimax es que la variable x_2 tiene una elevada carga del primer factor y una carga cercana al cero para el segundo factor, y x_3 tienen una elevada carga del segundo factor y una carga cercana a cero para el primer factor. x_1 tiene carga alta para el primer factor e intermedias para el segundo mientras que x_5 es al contrario aunque menos acentuado.

Computacionalmente, la rotación varimax consiste en maximizar la suma de las varianzas de los cuadrados de los factores de carga dentro de cada factor. Además, estos factores de carga se ajustan dividiéndolo cada uno por la comunalidad de la

variable correspondiente. Este ajuste es conocido como *normalización de Kaiser*. Este ajuste tiende a igualar el impacto de las variables con comunalidades variadas. Si no se realiza, las variables con altas comunalidades podrían tener una alta influencia en la solución final.

Cualquier método de rotación puede aplicarse a cualquier método de extracción de factores. Por ejemplo la rotación varimax en el método iterativo da la siguiente gráfica



Obsérvese que ocurre lo mismo que con los factores rotados y no rotados del método de componentes principales, esto es, se ve que lo que ha ocurrido es que los ejes han rotado, manteniendo la perpendicularidad, en el sentido de las agujas del reloj hasta quedar el primer factor junto a la variable x_2 y el segundo factor junto a la variable x_3, de esta manera han quedado todos los puntos en el primer cuadrante, Esta gráfica muestra claramente que los factores rotados son ortogonales puesto que el ángulo entre los ejes de los factores rotados es de 90°.

El resultado de esta rotación varimax es que la variable x_2 tiene una elevada carga del primer factor y una carga cercana al cero para el segundo factor, y x_3 tienen una elevada carga del segundo factor y una carga cercana a cero para el primer factor. x_1 tiene carga alta para el primer factor e intermedias para el segundo mientras que x_5 es al contrario.

Los resultados de la rotación varimax del método de componentes principales son

	Factores de Carga		*Comunalidad*	*Unicidad*
Variables	F_1	F_2	h_i^2	U_i^2
x_1	0.86057	0.41645	0.9140	0.0860
x_2	0.95997	0.04909	0.9239	0.0761
x_3	0.02482	0.88965	0.7921	0.2079
x_4	0.25786	0.69858	0.5545	0.4455
x_5	0.57377	0.73660	0.8718	0.1282
Varianza explicada	2.05843	1.99791	$\sum h_i^2$=4.0563	$\sum h_i^2$=0.9437
Porcentaje	41.17%	39.96%	81.13%	18.87%

Y los resultados de la rotación varimax del método de componentes principales iterados son

	Factores de Carga		*Comunalidad*	*Unicidad*
Variables	F_1	F_2	h_i^2	U_i^2
x_1	0.81842	0.47338	0.8939	0.1061
x_2	0.89342	0.12133	0.8129	0.1871
x_3	0.05611	0.73086	0.5373	0.4627
x_4	0.24899	0.52064	0.3331	0.6669
x_5	0.48877	0.84353	0.9504	0.0496
Varianza explicada	1.77206	1.75558	$\sum h_i^2$=3.5276	$\sum h_i^2$=1.4724
Porcentaje	50.23%	49.76%	70.55%	29.45%

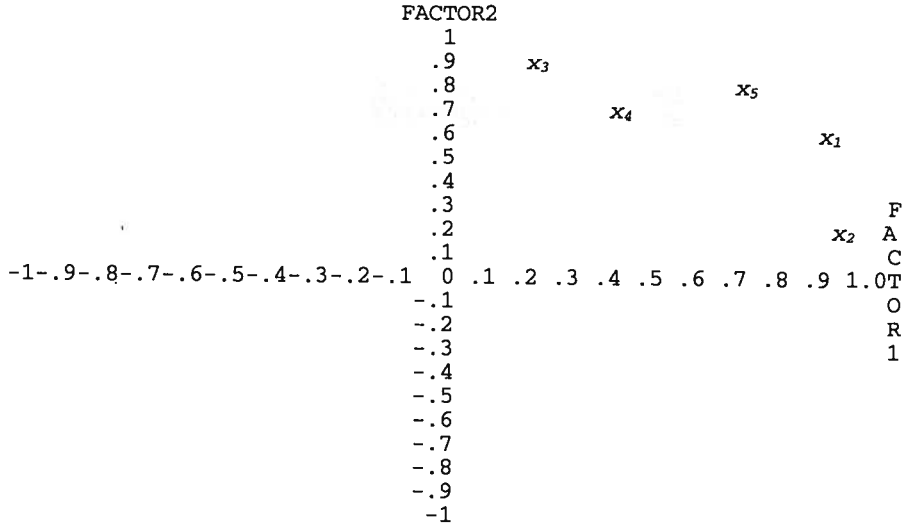
Si se comparan estas tablas con sus respectivas sin rotar, se observa que las comunalidades no han cambiado con la rotación. También se observa que el porcentaje de varianza explicada por el factor uno rotado es mucho menor que la explicada por el factor uno sin rotar, mientras que con el segundo factor ocurre lo contrario, de manera que el porcentaje acumulado de la varianza explicada por los dos primeros factores es el mismo después de la rotación. Como ocurría con los factores sin rotar, las cargas de los factores rotados dependen de cuantos factores se hayan seleccionado, además del método elegido para la extracción del factor inicial.

Rotación oblicua.-

Algunos autores no tienen en cuenta la restricción de la ortogonalidad de los factores rotados, con objeto de permitir un mayor grado de flexibilidad. Las rotaciones no ortogonales se denominan *rotaciones oblicuas*. El origen de este término es geométrico, pues dos líneas se denominan oblicuas si no son perpendiculares entre ellas. Los factores rotados oblicuamente están correlacionados entre ellos, lo que es

conveniente y deseable en algunas aplicaciones.

El resultado de aplicar una rotación oblicua (**PROMAX**) a los datos del ejemplo que se está desarrollando es, para el caso del método de los componentes principales



Aunque se estén representando perpendicularmente, la correlación entre estos ejes es de -0.4300 que se corresponden con un ángulo de 115.4646° (el coseno del ángulo da la correlación entre los ejes), Si se compara esta gráfica con la de la rotación ortogonal, efectivamente se observa que el eje horizontal se ha alejado de las variables hacia abajo y el eje vertical se ha alejado de las variables hacia la izquierda.

La tabla con los resultados de esta rotación en el método de componentes principales es

	Factores de Carga		***Comunalidad***	***Unicidad***
Variables	F_1	F_2	H_i^2	u_i^2
x_1	0.93180	0.59375	0.9140	0.0860
x_2	0.94671	0.25682	0.9239	0.0761
x_3	0.22273	0.87373	0.7921	0.2079
x_4	0.40726	0.73796	0.5545	0.4455
x_5	0.72369	0.84381	0.8718	0.1282
Varianza explicada	2.50370	2.43849	$\sum h_i^2$=4.0563	$\sum h_i^2$=0.9437
Porcentaje	50.07%	48.77%	81.13%	18.87%

Si se comparan esta tabla con la de los ejes sin rotar y con la de los ejes rotados ortogonalmente, se observa que las comunalidades no han cambiado con los dos tipos de rotaciones. También se observa que el porcentaje de varianza explicada por el factor uno rotado oblicuamente es mucho que la explicada por el factor uno sin rotar, mientras que la varianza explicada por el segundo factor rotado es mucho mayor

que la del segundo factor sin rotar, y además ha aumentado el porcentaje acumulado de la varianza explicada por los dos primeros factores.

Asignando puntuaciones de los factores a los individuos.-

Una vez realizada la extracción del los factores y las rotaciones, puede ser interesante obtener las puntuaciones de los individuos para cada factor. Para el ejemplo, si los dos factores, que son habilidad verbal y cuantitativa, proceden de un conjunto de puntuaciones de tests psicológicos, sería deseable determinar las ecuaciones para computar las puntuaciones de los individuos para estos dos factores. Dichas ecuaciones son funciones lineales de las variables originales.

Teóricamente, es posible construir las ecuaciones de las puntuaciones factoriales de muchas maneras. Pero quizás la manera más simple es añadir los valores de las variables que tienen una gran carga en un factor dado. Por ejemplo, para los datos hipotéticos que se están desarrollando, se puede decir que para obtener las puntuaciones del factor primero rotado para un individuo dado sería suficiente con sumar $x_1+x_2+x_5$. Similarmente, la puntuación del factor segundo rotado sería $x_3+x_4+x_5$. En efecto, el análisis factorial identifica a x_1, x_2 y x_5, como un subgrupo de variables intercorrelacionadas. Una manera de combinar la información aportada por las tres variables es simplemente sumarlas. En algunas aplicaciones y como una simple aproximación, esto puede ser suficiente.

Una aproximación más correcta es la que provee los paquetes estadísticos, como es el caso del *SAS*, y que denominan ***score***. Por ejemplo, la opción **SCORE** del *SAS* provee los *coeficientes de las puntuaciones de los factores* en base al cuadrado de la correlación múltiple de cada factor con la variable. Usando estos coeficientes en una combinación lineal de los valores de las variables tipificadas se obtienen las puntuaciones de cada individuo para cada factor. Para el ejemplo que se viene desarrollando, los coeficientes para las puntuaciones de los tres métodos estudiados son

Método de Componentes Principales

$$\textit{puntuación factor}\ 1 = 0.2967\,x_1 + 0.2363\,x_2 + 0.2084\,x_3 + 0.2196\,x_4 + 0.3023\,x_5$$
$$\textit{puntuación factor}\ 2 = -0.3015\,x_1 - 0.6352\,x_2 + 0.6226\,x_3 + 0.3224\,x_4 + 0.1291\,x_5$$

Método de Componentes Principales Iterados

$$\textit{puntuación factor}\ 1 = 0.4120\,x_1 + 0.0621\,x_2 - 0.0013\,x_3 - 0.0100\,x_4 + 0.5913\,x_5$$
$$\textit{puntuación factor}\ 2 = -0.3832\,x_1 - 0.7533\,x_2 + 0.1572\,x_3 + 0.0363\,x_4 + 0.8515\,x_5$$

Método de Máxima Verosimilitud

$$\textit{puntuación factor}\ 1 = 0.0222\,x_1 + 0.9347\,x_2 + 0.0024\,x_3 + 0.0024\,x_4 + 0.0773\,x_5$$
$$\textit{puntuación factor}\ 2 = 0.1081\,x_1 - 0.7448\,x_2 + 0.0977\,x_3 + 0.0497\,x_4 + 0.9976\,x_5$$

Dando las siguientes puntuaciones de los individuos para el primer factor

Sin rotar	*C. Principales*		*C. P. Iterados*		*M. Verosimilitud*	
Individuo	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*
1	-0.8760	0.1525	-1.0101	-0.0283	-0.6101	-0.6620
2	-0.7560	0.8487	-1.0794	0.0068	-1.1184	-0.4994
3	1.1741	0.9303	1.3308	1.2032	-0.1321	1.9625
4	-0.2954	-1.1085	0.0381	-0.6000	0.2120	-0.4456
5	-1.3331	0.1721	-1.3323	0.2784	-1.1292	-0.5663
6	0.6235	1.1826	0.4194	0.8105	-0.2192	0.9685
7	1.4732	0.4708	1.2497	0.3176	1.3406	0.7983
8	1.1009	-2.0436	1.0177	-1.9786	2.0153	-1.0408
9	-0.4334	-0.0258	-0.1492	0.5628	-0.2735	0.2829
10	-0.6778	-0.5792	-0.4806	-0.5725	-0.0853	-0.7975
$\overline{X}$	0.0	0.0	0.0	0.0	0.0	0.0
s^2	1.0	1.0	0.97	0.80	0.99	0.94

Estas puntuaciones o valores de los factores no son sino lo que en el apartado anterior se denominaron *componentes principales* pero con media cero y varianza uno, esto es, estos valores son las componentes principales tipificadas.

Y los coeficientes y puntuaciones con la rotación varimax son

Método de Componentes Principales

$$\textit{puntuación factor } 1 = 0.4229\, x_1 + 0.6121\, x_2 - 0.2842\, x_3 - 0.0671\, x_4 + 0.1270\, x_5$$

$$\textit{puntuación factor } 2 = -0.0097\, x_1 - 0.2911\, x_2 + 0.5919\, x_3 + 0.3842\, x_4 + 0.3032\, x_5$$

Método de Componentes Principales Iterados

$$\textit{puntuación factor } 1 = 0.5624\, x_1 + 0.5747\, x_2 - 0.1117\, x_3 - 0.0327\, x_4 - 0.1780\, x_5$$

$$\textit{puntuación factor } 2 = 0.0182\, x_1 - 0.4909\, x_2 + 0.1106\, x_3 + 0.0187\, x_4 + 0.0230\, x_5$$

Método de Máxima Verosimilitud

$$\textit{puntuación factor } 1 = 0.1103\, x_1 - 0.5658\, x_2 + 0.0966\, x_3 + 0.0493\, x_4 + 0.9946\, x_5$$

$$\textit{puntuación factor } 2 = 0.0025\, x_1 + 1.0528\, x_2 - 0.0151\, x_3 - 0.0065\, x_4 - 0.1061\, x_5$$

Varimax	*C. Principales*		*C. P. Iterados*		*M. Verosimilitud*	
Individuo	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*
1	-0.7347	-0.5008	-0.6971	-0.7317	-0.7603	-0.4820
2	-1.1336	0.0822	-0.7710	-0.7553	-0.6916	-1.0111
3	0.1942	1.4853	0.0971	1.7915	1.9073	-0.4807
4	0.5603	-1.0010	0.4496	-0.3990	-0.4005	0.2882
5	-1.0763	-0.8052	-1.1446	-0.7436	-0.7594	-1.0097
6	-0.3765	1.2828	-0.2732	0.8707	0.9138	-0.3887
7	0.7289	1.3640	0.6633	1.1057	1.0250	1.1764
8	2.2134	-0.6991	2.1161	-0.6876	-0.6639	2.1689
9	-0.2929	-0.3205	-0.5024	0.2944	0.2295	-0.3197
10	-0.0828	-0.8877	0.0621	-0.7449	-0.7999	0.0585
$\overline{X}$	0.0	0.0	0.0	0.0	0.0	0.0
s^2	1.0	1.0	0.87	0.90	0.94	0.99

Estas son las componentes principales tipificadas después de la rotación varimax.

Ejemplo.-

Se tiene el gasto anual medio que realizan 112 familias en siete productos o categorías de productos alimenticios. Se piensa que el gasto y el tipo de gasto viene influido por dos factores: un ***factor cultural*** que también determina el nivel laboral del padre y un ***factor socio-económico*** que también determina el número de hijos. Por lo que las familias están clasificadas según el nivel profesional del padre y según el número de hijos, habiendo doce tipos en total: **T2** trabajador manual con dos hijos, **O2** empleado de oficina con dos hijos, y **D2** directivo con dos hijos; y los mismos niveles profesionales del padre para 3, 4 y 5 hijos. Las categorías de productos alimenticios son: pan, legumbres, fruta, carne, pollo, leche y vino

Archivo de programa *SAS* (C18-3.SAS).-

```
title 'Análisis Factorial';
options ls=80 ps=60;
data vino;
infile 'c18-1.dat';
input  familia $ simfam $ pan legumbre fruta carne pollos leche vino @@;
proc factor c n=2 scree preplot score rotate=varimax plot rotate=promax;
run;
proc factor method=prinit n=2 scree preplot score rotate=varimax plot
            rotate=promax;
run;
proc factor method=ml n=2 scree preplot score rotate=varimax plot
            rotate=promax;
run;
```

Recuérdese que los doce tipos de familia se han simbolizado: con las vocales A, E, I y O los trabajadores de 2, 3, 4 y 5 hijos respectivamente; con las primeras consonantes del abecedario, B, C, D y F la de los oficinistas con 2, 3, 4 y 5,

respectivamente; y con las últimas consonantes del abecedario, V, W, X y Z la de los directivos con 2, 3, 4 y 5 hijos respectivamente

Las salidas que da este programa (**C18-3.LST**) son:

. La matriz de correlaciones (**Correlations**) (como en C18-2.lst) cuya traza es igual al número de variables, es decir, siete, que será la suma de los valores propios. Esta matriz está solo en la salida del primer Proc FACTOR, en donde se ha puesto la opción, **C**, para que saliera dicha matriz.

. Los valores propios (**Eigenvalues**). Para los tres métodos utilizados los dos primeros valores propios son los que valen más que la unidad. Para el método de componentes principales y el método de componentes principales iterado los valores son los mismos, estos son, el primer valore propio vale 3.7820 y absorbe el 54.03% de la varianza y el segundo valor propio vale 1.6017 y absorbe el 22.88% de la varianza, entre ambos absorben el 76.91% de la varianza. Para el método de máxima verosimilitud, el primer valor propio vale 22.4826 y absorbe el 89.38% de la varianza y el segundo valor propio vale 3.3463 y absorbe el 13.30% de la varianza, entre ambos suman mas del 100%.

```
Initial Factor Method: Principal Components
                    Prior Communality Estimates: ONE

        Eigenvalues of the Correlation Matrix:  Total = 7  Average = 1
                              1          2          3          4
         Eigenvalue        3.7820     1.6017     0.8184     0.3651
         Difference        2.1803     0.7834     0.4533     0.1224
         Proportion        0.5403     0.2288     0.1169     0.0522
         Cumulative        0.5403     0.7691     0.8860     0.9382

Initial Factor Method: Iterated Principal Factor Analysis

                    Prior Communality Estimates: ONE

              Preliminary Eigenvalues:  Total = 7  Average = 1
                              1          2          3          4
         Eigenvalue        3.7820     1.6017     0.8184     0.3651
         Difference        2.1803     0.7834     0.4533     0.1224
         Proportion        0.5403     0.2288     0.1169     0.0522
         Cumulative        0.5403     0.7691     0.8860     0.9382

Initial Factor Method: Maximum Likelihood
                    Prior Communality Estimates: SMC
          PAN  LEGUMBRE     FRUTA     CARNE    POLLOS     LECHE      VINO
     0.552809  0.739965  0.829909  0.852136  0.898430  0.586055  0.144109

      Preliminary Eigenvalues:  Total = 25.1535638  Average = 3.59336626
                              1          2          3          4
         Eigenvalue       22.4826     3.3463     0.2775     0.1656
         Difference       19.1363     3.0688     0.1119     0.3520
         Proportion        0.8938     0.1330     0.0110     0.0066
         Cumulative        0.8938     1.0268     1.0379     1.0445
```

. Los tres métodos presentan a continuación el **Screen** de los valores propios en el que se ve, en el método de máxima verosimilitud, que hay dos pendientes, una que unen el primer y segundo valor y otra que une los cinco restantes. En los otros dos métodos, además de la pendiente que unen el primero y segundo valor propio, hay más pendientes.

. En el método de los componentes iterados y en el de máxima verosimilitud, vienen a continuación las iteraciones de las comunalidades, y después viene la nueva estima de los valores propios. Estas nuevas estima son, en el método del factor principal iterado, de 3.5969 para el primer valor propio que absorbe el 73.66% y de 1.2866 que absorbe el 26.35%, entre ambos absorben el 100% de la varianza. Y en el método de máxima verosimilitud, a continuación de las iteraciones vienen las dos pruebas estadísticas, en la primera al ser significativa se acepta la hipótesis nula de que hay más de un factor común; la segunda prueba también es significativa por lo que habría que aceptar la hipótesis alternativa de que se necesitan más de dos factores para explicar los 120 datos. A continuación presenta las nuevas estimas de los valores propios, estas son, el valor del primero es de 29.3831 y absorbe el 86.65% y el valor del segundo es de 4.5631 y absorbe el 13.44%, entre ambos absorben el 100%, por lo que con dos factores es suficiente a pesar de lo que se desprenda de la segunda prueba de hipótesis planteada anteriormente.

```
          Eigenvalues of the Reduced Correlation Matrix:
              Total = 4.88317279  Average = 0.69759611

                        1          2          3          4
  Eigenvalue       3.5969     1.2866     0.1194     0.0615
  Difference       2.3103     1.1671     0.0580     0.0740
  Proportion       0.7366     0.2635     0.0245     0.0126
  Cumulative       0.7366     1.0001     1.0245     1.0371

Significance tests based on 120 observations:
   Test of H0: No common factors.
       vs HA: At least one common factor.
   Chi-square = 651.685    df = 21    Prob>chi**2 = 0.0001

   Test of H0: 2 Factors are sufficient.
       vs HA: More factors are needed.

   Chi-square = 37.175    df = 8    Prob>chi**2 = 0.0001
Chi-square without Bartlett's correction = 38.635709545
Akaike's Information Criterion = 22.635709545
Schwarz's Bayesian Criterion = 0.3357756032
Tucker and Lewis's Reliability Coefficient = 0.8785707825

    Eigenvalues of the Weighted Reduced Correlation Matrix:
              Total = 33.9461874  Average = 4.84945535
                        1          2          3          4
  Eigenvalue      29.3831    4.5631     0.5215     0.2619
  Difference      24.8200    4.0416     0.2596     0.3553
  Proportion       0.8656    0.1344     0.0154     0.0077
  Cumulative       0.8656    1.0000     1.0154     1.0231
```

. Después viene la matriz factorial, en la que se observa, en los tres métodos, que la variable más importante para el primer factor (*factor cultural*) es **POLLOS**, seguida muy de cerca por las variables, **FRUTA** y **CARNE**, mientras que la variable que tiene una menor correlación con el primer factor es **VINO**, si bien al tener una correlación negativa le da gran importancia pues significa que cuando aumenta el primer factor aumenta, prácticamente con la misma intensidad, el consumo de carne, pollos y fruta mientras que disminuye, aunque con menos intensidad, el consumo de vino. Para el segundo factor (*factor socioeconómico*), la variable más importante es el **PAN**, seguido muy de cerca por la leche y el vino también tiene una correlación significativamente positiva con este segundo factor, mientras que las variables carne, pollos y fruta tienen

correlaciones negativas con este factor, La variable **LEGUMBRE**, tiene una correlación significativa positiva con los dos factores, si bien es mayor la correlación con el segundo factor que con el primero.

. La siguiente salida es el de las comunalidades, se observa en los tres métodos que las comunalidades son altas en las seis variables de comida y no tanto en la variable vino.

. A continuación viene las salidas de la rotación VARIMAX, donde sigue siendo las variables más importantes para el primer factor los pollos, carne y fruta, pero donde ha disminuido drásticamente la correlación con el pan que incluso se ha hecho levemente negativa, mientras que la correlación negativa con el vino ha aumentado de valor. Con respecto al segundo factor, ahora es muy significativa la correlación con el pan y también con la leche y han desaparecido las correlaciones negativas, esto es, todas las variables tienen correlaciones positivas con el segundo factor.

. En la rotación oblicua ocurre lo mismo, con el primer factor esta muy correlacionado la variable *pollos*, *carne* y *leche*, y la variable *vino* esta significativamente correlacionada negativamente. El segundo factor esta correlacionado con *pan* y *leche*. La variable *legumbre* esta correlacionada con ambos factores.

. Tanto para la rotación ortogonal como para la oblicua da también los coeficientes para obtener la puntuación de cada individuo.

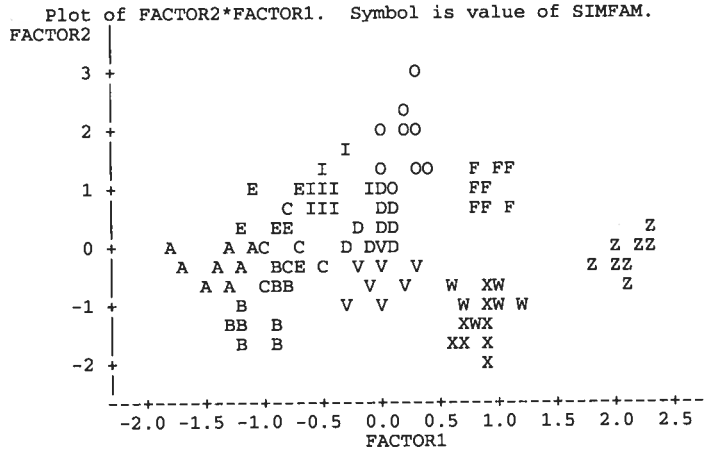
Si se desea la puntuación de cada familia para los dos primeros factores y su representación gráfica, tanto con los factores sin rotar como con los factores rotados, el programa sería

Archivo de programa *SAS* (C18-4.SAS).-

```
Title 'Análisis factorial; puntuación de cada familia';
Options ls=75 ps=30;
Data vino;
Infile 'c18-1.dat';
Input  familia $ simfam $ pan legumbre fruta carne pollos leche vino @@;
Proc factor data=vino n=2 score outstat=fact;
var pan legumbre fruta carne pollos leche vino;
run;
proc score data=vino score=fact out=scores;
var pan legumbre fruta carne pollos leche vino;
run;
proc plot;
plot factor2*factor1=simfam/vspace=3 hspace=5;
run;
proc factor data=vino n=2 rotate=varimax score outstat=fact;
var pan legumbre fruta carne pollos leche vino;
run;
proc score data=vino score=fact out=scores;
var pan legumbre fruta carne pollos leche vino;
run;
proc plot;
plot factor2*factor1=simfam/vspace=3 hspace=5;
run;
```

Archivo de resultados (C18-4.LST).-

```
Initial Factor Method: Principal Components
                Prior Communality Estimates: ONE
      Eigenvalues of the Correlation Matrix:  Total = 7  Average = 1
                        1          2          3          4
       Eigenvalue     3.7820     1.6017     0.8184     0.3651
       Difference     2.1803     0.7834     0.4533     0.1224
       Proportion     0.5403     0.2288     0.1169     0.0522
       Cumulative     0.5403     0.7691     0.8860     0.9382

          2 factors will be retained by the NFACTOR criterion.
                           Factor Pattern
                                FACTOR1    FACTOR2
                    PAN         0.47988    0.77024
                    LEGUMBRE    0.87298    0.22300
                    FRUTA       0.89600   -0.28393
                    CARNE       0.88917   -0.28421
                    POLLOS      0.90230   -0.29807
                    LECHE       0.54228    0.66952
                    VINO       -0.29658    0.51014

                    Variance explained by each factor
                          FACTOR1    FACTOR2
                        3.781998   1.601741

              Final Communality Estimates: Total = 5.383738
        PAN  LEGUMBRE      FRUTA      CARNE     POLLOS      LECHE       VINO
   0.823560  0.811834   0.883439   0.871396   0.902994   0.742313   0.348202

              Scoring Coefficients Estimated by Regression
       Squared Multiple Correlations of the Variables with each Factor
                          FACTOR1    FACTOR2
                         1.000000   1.000000

                    Standardized Scoring Coefficients
                                FACTOR1    FACTOR2
                    PAN         0.12688    0.48088
                    LEGUMBRE    0.23083    0.13923
                    FRUTA       0.23691   -0.17727
                    CARNE       0.23511   -0.17744
                    POLLOS      0.23858   -0.18609
                    LECHE       0.14338    0.41799
                    VINO       -0.07842    0.31849

          Plot of FACTOR2*FACTOR1.  Symbol is value of SIMFAM.
```




```
NOTE: 26 obs hidden.
Initial Factor Method: Principal Components
```

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1

	1	2	3	4
Eigenvalue	3.7820	1.6017	0.8184	0.3651
Difference	2.1803	0.7834	0.4533	0.1224
Proportion	0.5403	0.2288	0.1169	0.0522
Cumulative	0.5403	0.7691	0.8860	0.9382

2 factors will be retained by the NFACTOR criterion.

Factor Pattern

	FACTOR1	FACTOR2
PAN	0.47988	0.77024
LEGUMBRE	0.87298	0.22300
FRUTA	0.89600	-0.28393
CARNE	0.88917	-0.28421
POLLOS	0.90230	-0.29807
LECHE	0.54228	0.66952
VINO	-0.29658	0.51014

Variance explained by each factor

FACTOR1	FACTOR2
3.781998	1.601741

Final Communality Estimates: Total = 5.383738

PAN	LEGUMBRE	FRUTA	CARNE	POLLOS	LECHE	VINO
0.823560	0.811834	0.883439	0.871396	0.902994	0.742313	0.348202

Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.86828	0.49608
2	-0.49608	0.86828

Rotated Factor Pattern

	FACTOR1	FACTOR2
PAN	0.03456	0.90684
LEGUMBRE	0.64736	0.62670
FRUTA	0.91883	0.19796
CARNE	0.91304	0.19432
POLLOS	0.93131	0.18881
LECHE	0.13871	0.85034
VINO	-0.51058	0.29581

Variance explained by each factor

FACTOR1	FACTOR2
3.245446	2.138292

Rotation Method: Varimax

Final Communality Estimates: Total = 5.383738

PAN	LEGUMBRE	FRUTA	CARNE	POLLOS	LECHE	VINO
0.823560	0.811834	0.883439	0.871396	0.902994	0.742313	0.348202

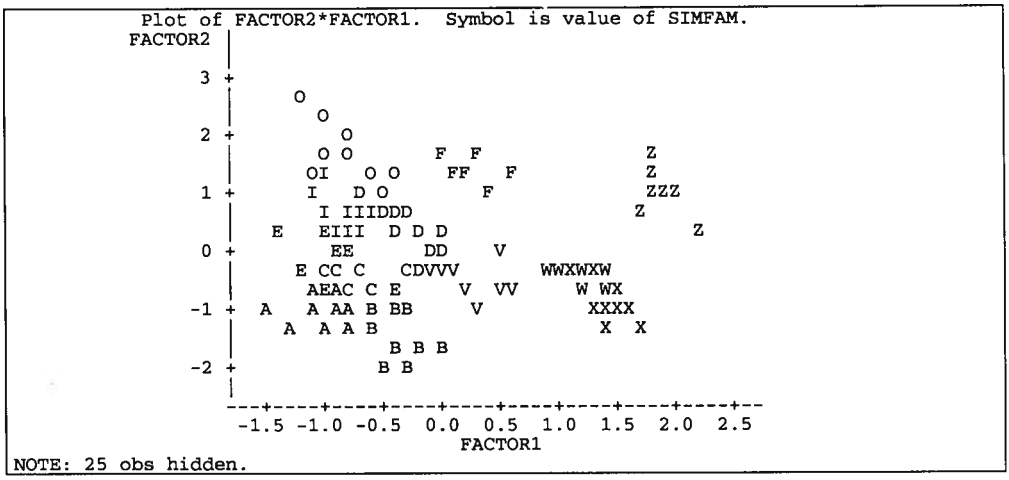
Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with each Factor

FACTOR1	FACTOR2
1.000000	1.000000

Standardized Scoring Coefficients

	FACTOR1	FACTOR2
PAN	-0.12838	0.48048
LEGUMBRE	0.13135	0.23540
FRUTA	0.29364	-0.03639
CARNE	0.29216	-0.03744
POLLOS	0.29947	-0.04322
LECHE	-0.08286	0.43406
VINO	-0.22609	0.23764



En el archivo de resultados (**C18-4.LST**) se observa en las salidas del procedimiento *plot* que el primer eje discrimina para el factor *cultural*, cuyo marcador es el nivel profesional del cabeza de familia, mientras que el segundo discrimina para el factor *socioeconómico*, cuyo marcador es el número de hijos. La discriminación del segundo factor se hace más clara con los componentes rotados.

Análisis de *tamaño* y *forma*.-

Una de las primeras aplicaciones del análisis de componentes principales en el estudio de morfométrico es el de establecer y detectar los conceptos de *tamaño* y el concepto de *forma* de los individuos.

La idea de tamaño se considera equivalente a la de crecimiento. Cuando los caracteres están representados por p variables aleatorias, parece razonable asociar el tamaño como la dirección de máxima variabilidad, es decir, como la primera componente principal.

Por otra parte, una variable biométrica, cuanto más variabilidad tiene, mejor expresa el concepto de tamaño. Por ejemplo, considérese un grupo de hombres de prácticamente el mismo peso pero con notable variación de altura; entonces, para ordenarlos de menor a mayor tamaño, se ordenan de menor a mayor altura. La variable peso no expresaría bien la noción de tamaño. La variable con mayor varianza será la que mejor expresará este concepto. Si esta variable (tamaño) puede ser una combinación lineal de las p variables originales, esta variable, *tamaño*, debe ser la primera componente principal.

Con respecto a la forma, este concepto es independiente del tamaño. Dos individuos pueden tener el mismo tamaño pero distinta forma y recíprocamente. Como la segunda, tercera, etc., componentes principales, están incorrelacionadas con la primera, parece también razonable interpretarlas como variables que expresen la forma de los individuos.

Para que las componentes representen adecuadamente tamaño y forma, deben

cumplirse tres condiciones:

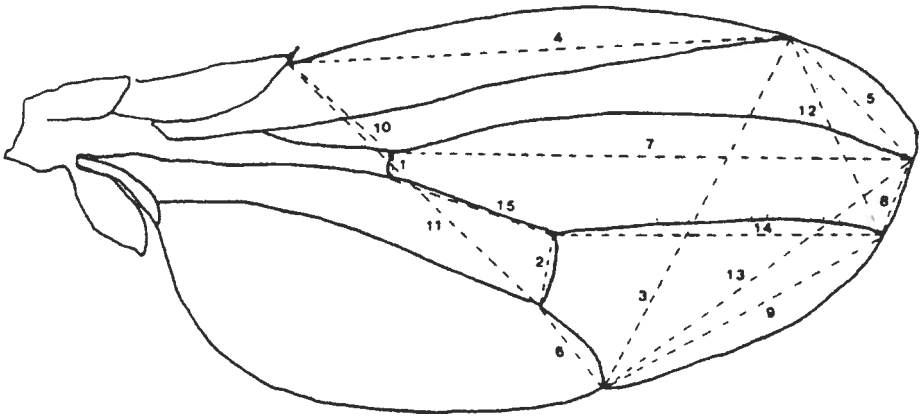
1) Todos los elementos del primer vector propio o del primer factor, deben ser positivos para que la primera componente principal se identifique como tamaño. En efecto, todo incremento positivo de las medidas biométricas redundarán en un incremento positivo de la primera componente principal, lo que viene a significar que si aumentan las medidas, aumenta el tamaño. Si esta condición no se verifica no se puede hablar de tamaño.

2) Para que una componente se identifique como forma, los elementos de su vector propio no debe tener todos el mismo signo, sino que algunos deben ser positivos y otros negativos. Un factor de forma debe ser tal que un incremento del factor resulte de un incremento de unas medidas y un decremento de otras medidas.

3) Si las componentes de forma se extraen de la matriz de covarianzas, es aconsejable considerar sólo aquellas cuyas varianzas superen a la menor de las varianzas de la diagonal de la matriz de varianzas/covarianzas.

Ejemplo.-

Se tienen las 15 posibles medidas del ala de *Drosophila melanogaster* que se muestran en el siguiente dibujo. Algunas de estas medidas son importantes para estudios de genética cuantitativa y para la diferenciación con especies gemelas. Se pretende saber cual o cuales de estas medidas pueden indicar el tamaño del ala y cual o cuales pueden indicar la forma del ala



Archivo de programa *SAS* (C18-5.SAS).-

```
title 'Análisis de tamaño y Forma';
Options ls=75 ps=60;
Data tf;
Infile 'c18-5.dat';
Input trata v1-v15;
Proc princomp cov std out=prefix;
var v1-v15;
run;
proc plot;
plot prin2*prin1=trata /vspace=3 hspace=5;
run;
proc factor data=tf scree n=2 preplot;
var v1-v15;
run;
```

Archivo de datos (C18-5.DAT).-

```
1 0.25 0.80 4.30 5.00 2.10 1.40 5.80 0.95 3.95 1.35 5.10 2.55 3.55 3.90 1.80
1 0.25 0.65 4.10 5.20 2.20 1.60 5.70 0.80 3.40 1.20 4.55 2.40 3.40 4.00 1.55
1 0.30 0.70 4.50 5.45 2.30 1.70 6.10 0.90 3.50 1.25 4.85 2.65 4.05 4.15 1.70
1 0.25 0.75 4.40 5.55 2.10 1.55 6.05 0.85 3.70 1.35 3.55 2.50 4.00 4.00 1.80
1 0.25 0.75 4.45 5.35 2.15 1.60 6.10 0.80 3.55 1.15 3.15 2.50 4.30 4.20 1.65
1 0.30 0.75 4.65 5.55 2.45 1.70 6.70 0.95 4.00 1.55 3.35 3.00 4.60 4.40 1.85
1 0.25 0.75 4.50 5.15 2.15 1.45 6.05 0.80 3.70 1.30 3.40 2.55 4.20 4.00 1.80
1 0.25 0.65 4.55 5.50 2.15 1.75 6.15 0.80 3.80 1.40 3.55 2.50 4.00 4.20 1.75
1 0.25 0.75 4.45 5.25 2.30 1.60 6.20 0.80 3.60 1.25 3.35 2.60 4.15 4.15 1.80
1 0.30 0.80 4.65 5.65 2.20 1.70 6.50 0.80 3.75 1.35 3.60 2.70 4.25 4.25 2.20
1 0.25 0.70 4.50 5.45 2.15 1.45 5.95 0.85 3.60 1.35 3.25 2.50 4.20 4.05 1.60
1 0.25 0.80 4.75 6.05 2.45 1.65 6.80 0.80 3.70 1.35 3.55 2.80 4.55 4.45 2.00
1 0.25 0.65 4.20 5.50 2.05 1.60 5.90 0.95 3.70 1.40 3.60 2.50 4.00 3.95 1.75
1 0.25 0.65 4.55 5.50 2.20 1.65 6.30 0.95 3.70 1.30 3.35 2.60 4.35 4.25 1.70
1 0.25 0.70 4.50 5.60 2.20 1.70 6.25 0.85 3.50 1.25 3.20 2.60 4.25 5.15 1.75
2 0.30 0.75 4.50 6.00 2.25 1.50 6.60 0.95 4.00 1.30 5.10 2.65 4.50 4.60 1.80
2 0.30 0.75 4.80 5.60 2.20 1.60 6.60 0.90 4.00 1.25 5.05 2.65 4.50 4.35 2.00
2 0.35 0.85 4.85 5.55 2.35 1.85 6.70 0.95 4.10 1.45 5.35 2.70 4.60 4.70 1.80
2 0.30 0.90 5.30 6.55 2.45 1.85 7.10 0.95 4.10 1.60 3.85 2.80 4.85 4.90 1.90
2 0.30 0.90 5.10 6.15 2.55 1.85 7.00 0.90 4.20 1.55 4.15 2.90 4.85 4.80 1.85
2 0.35 0.90 5.05 6.30 2.45 1.65 7.00 0.85 3.60 1.30 2.95 3.05 4.80 4.70 1.85
2 0.30 0.85 4.95 6.00 2.25 1.70 6.70 0.85 4.20 1.35 5.50 2.70 4.00 4.45 2.00
2 0.25 0.80 5.00 6.15 2.25 1.65 7.00 0.95 4.20 1.60 3.90 2.75 4.70 4.70 2.00
2 0.30 0.85 5.00 6.50 2.50 1.85 7.10 0.95 4.10 1.40 3.70 2.80 4.70 4.70 2.05
2 0.35 0.85 5.05 6.55 2.45 1.75 7.10 0.95 4.20 1.55 3.95 2.85 4.75 4.70 2.10
2 0.30 0.80 4.85 6.00 2.50 1.85 6.80 1.00 4.20 1.45 3.85 2.80 4.55 4.65 1.85
2 0.25 0.75 4.80 6.05 2.10 1.75 6.60 0.95 4.00 1.50 3.75 2.65 4.55 4.45 1.85
2 0.30 0.70 5.00 6.15 2.10 1.55 6.70 0.90 4.20 1.65 4.15 2.70 4.60 4.45 1.90
2 0.35 0.85 5.00 6.50 2.50 1.90 7.15 0.95 4.15 1.55 3.95 2.90 4.65 4.75 2.15
2 0.40 0.85 5.05 6.35 2.30 1.85 7.00 0.95 4.20 1.60 3.85 2.75 4.70 4.70 2.00
3 0.25 0.55 3.70 4.00 2.10 1.40 5.15 0.75 3.00 1.05 2.65 2.35 3.60 3.50 1.35
3 0.25 0.60 3.80 4.30 2.25 1.40 5.35 0.65 3.10 1.15 2.95 2.45 3.60 3.60 1.60
3 0.25 0.60 3.90 4.35 2.10 1.40 5.30 0.70 3.00 1.20 2.95 2.50 3.60 3.55 1.50
3 0.25 0.60 3.85 4.20 2.10 1.45 5.35 0.70 3.20 1.10 2.70 2.40 3.80 3.80 1.40
3 0.25 0.55 3.70 4.50 2.10 1.45 5.45 0.60 3.05 1.15 2.85 2.40 3.50 3.20 1.55
3 0.20 0.50 3.55 4.00 2.10 1.40 4.95 0.65 2.95 1.10 0.60 2.30 3.45 3.50 1.25
3 0.20 0.55 3.65 4.35 1.90 1.50 5.25 0.65 2.90 1.15 2.75 2.30 3.45 3.55 1.45
3 0.25 0.50 3.70 4.30 2.10 1.45 5.30 0.70 3.05 1.05 2.65 2.45 3.60 3.75 1.35
3 0.30 0.55 3.60 4.20 2.00 1.45 5.25 0.60 3.05 1.05 2.60 2.40 3.55 3.80 1.30
3 0.25 0.60 3.70 4.25 2.00 1.40 5.30 0.75 3.10 1.15 2.75 2.40 3.70 3.60 1.40
3 0.25 0.60 3.90 4.60 2.00 1.40 5.55 0.75 3.20 1.20 3.05 2.50 3.80 3.75 1.60
3 0.20 0.50 3.40 4.05 1.90 1.40 5.20 0.65 2.80 1.00 2.60 2.20 3.35 3.55 1.50
3 0.25 0.55 3.60 4.30 2.00 1.35 5.30 0.70 2.85 1.00 2.70 2.30 3.50 3.60 1.45
3 0.25 0.50 3.80 4.30 2.10 1.55 5.35 0.65 3.10 1.15 2.75 2.45 3.70 3.90 1.35
3 0.30 0.60 3.75 4.50 2.00 1.45 5.60 0.75 3.20 1.15 3.00 2.50 3.80 3.85 1.60
4 0.30 0.65 4.05 4.90 2.20 1.45 5.90 0.70 3.50 1.20 3.10 2.55 4.00 4.00 1.70
4 0.30 0.70 4.35 5.20 2.30 1.65 6.00 0.75 3.50 1.30 3.25 2.65 4.10 4.20 1.60
4 0.30 0.75 4.50 5.30 2.50 1.60 6.40 0.75 3.70 1.30 3.30 2.85 4.30 4.35 1.75
4 0.30 0.75 4.55 5.50 2.30 1.60 6.35 0.75 3.70 1.35 3.40 2.70 4.30 4.30 1.80
4 0.30 0.80 4.55 5.45 2.40 1.60 5.50 0.85 3.70 1.35 3.40 2.80 4.35 4.35 1.75
4 0.30 0.65 4.15 4.80 2.30 1.70 5.95 0.75 3.40 1.25 3.00 2.50 4.00 4.10 1.55
4 0.25 0.75 4.20 5.10 2.20 1.60 6.00 0.75 3.50 1.25 3.20 2.60 4.10 4.10 1.70
```

```
4 0.30 0.65 4.10 5.10 2.20 1.55 5.95 0.75 3.35 1.25 3.15 2.60 3.95 4.05 1.65
4 0.30 0.70 4.20 4.80 2.40 1.60 6.05 0.75 3.50 1.30 3.30 2.60 4.05 4.10 1.65
4 0.40 0.70 4.65 5.40 2.30 1.70 6.20 0.90 3.60 1.50 3.45 2.75 4.35 4.25 1.65
4 0.30 0.65 4.35 5.10 2.20 1.40 5.95 0.85 3.50 1.35 3.20 2.65 4.20 4.10 1.65
4 0.25 0.65 4.15 4.95 2.25 1.40 6.05 0.75 3.50 1.25 3.30 2.60 4.05 4.05 1.70
4 0.20 0.70 4.30 5.20 2.30 1.35 6.25 0.75 3.60 1.35 3.40 2.75 4.25 4.15 1.90
4 0.30 0.65 3.95 4.60 2.20 1.55 5.80 0.75 3.40 1.15 3.00 2.55 4.00 4.00 1.65
4 0.30 0.65 3.95 5.00 2.30 1.60 5.90 0.80 3.30 1.35 3.15 2.70 4.40 4.10 1.60
```

El primer dígito hace referencia a los cuatro tratamientos: 1 y 2 son machos y hembras, respectivamente, criados a 25°C; 3 y 4 lo mismo criados a 30°C.

Archivo de resultados (C18-5.LST).-

Principal Component Analysis

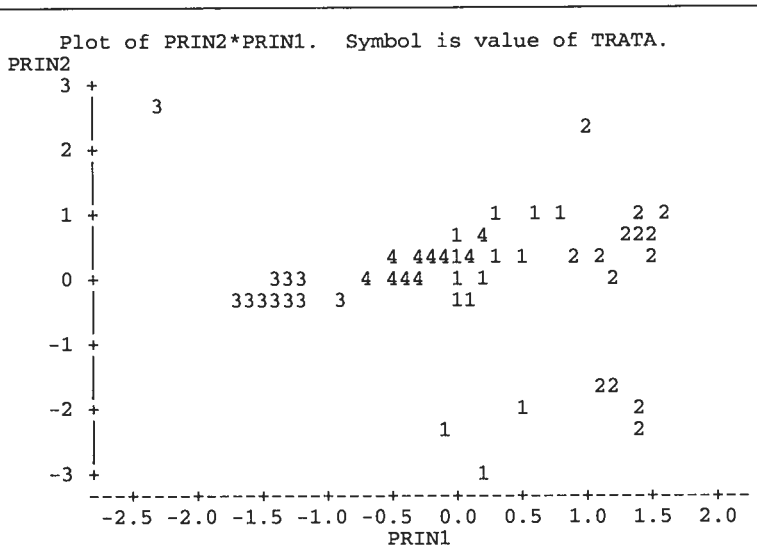
60 Observations
15 Variables

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	1.97980	1.64885	0.808087	0.80809
PRIN2	0.33095	0.29585	0.135081	0.94317
PRIN3	0.03509	0.01017	0.014323	0.95749
PRIN4	0.02493	0.00451	0.010174	0.96767
PRIN5	0.02041	0.00624	0.008332	0.97600
PRIN6	0.01418	0.00113	0.005786	0.98178
PRIN7	0.01304	0.00432	0.005325	0.98711
PRIN8	0.00873	0.00078	0.003563	0.99067
PRIN9	0.00795	0.00292	0.003246	0.99392
PRIN10	0.00503	0.00085	0.002053	0.99597
PRIN11	0.00418	0.00214	0.001706	0.99768
PRIN12	0.00204	0.00026	0.000831	0.99851
PRIN13	0.00177	0.00066	0.000723	0.99923
PRIN14	0.00111	0.00033	0.000452	0.99968
PRIN15	0.00078	.	0.000317	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
V1	0.017188	0.006120	0.063204	0.014314	-.002513
V2	0.071080	0.017188	0.017171	0.041870	0.007032
V3	0.332565	0.133334	-.079946	0.060841	0.333686
V4	0.498756	0.222249	-.588551	-.389415	-.076800
V5	0.078629	0.083676	0.281605	0.112259	-.119032
V6	0.083313	0.052243	0.077852	-.186366	-.000074
V7	0.413295	0.207273	0.073902	0.296961	-.609888
V8	0.065295	-.006557	-.024152	-.013287	0.187000
V9	0.277101	0.001634	-.100094	0.335608	0.524036
V10	0.097212	0.051511	-.087650	0.169740	0.284607
V11	0.437549	-.877942	0.123550	-.019643	-.053028
V12	0.104991	0.093106	0.216528	0.194952	-.061678
V13	0.264606	0.266252	0.423169	0.260055	0.172833
V14	0.263827	0.162119	0.506665	-.634299	0.081516
V15	0.136993	0.021886	-.189442	0.239162	-.250703



Initial Factor Method: Principal Components

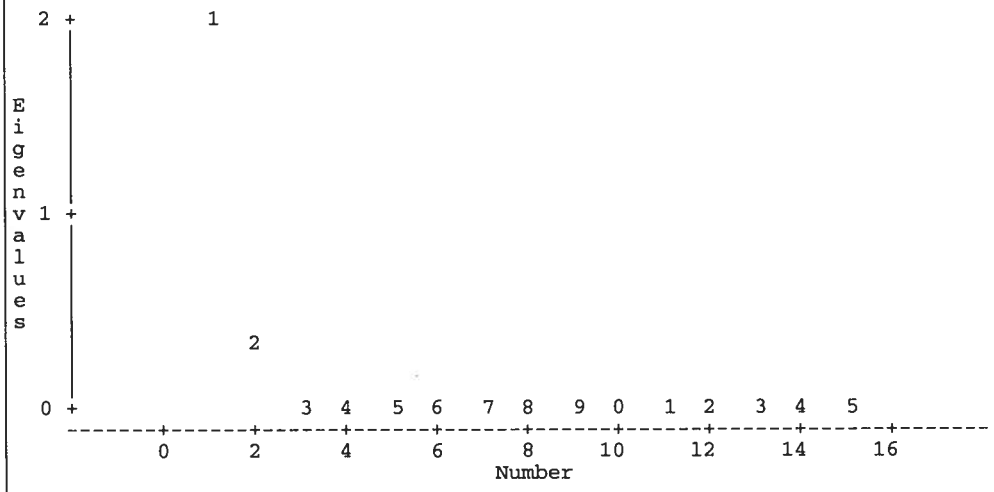
Prior Communality Estimates: ONE

Eigenvalues of the Covariance Matrix:
Total = 2.44998164 Average = 0.16333211

	1	2	3	4	5
Eigenvalue	1.9798	0.3309	0.0351	0.0249	0.0204
Difference	1.6489	0.2959	0.0102	0.0045	0.0062
Proportion	0.8081	0.1351	0.0143	0.0102	0.0083
Cumulative	0.8081	0.9432	0.9575	0.9677	0.9760

2 factors will be retained by the NFACTOR criterion.

Scree Plot of Eigenvalues



Factor Pattern

	FACTOR1	FACTOR2
V1	0.56834	0.08274
V2	0.91659	0.09062
V3	0.96781	0.15864
V4	0.96673	0.17613
V5	0.69678	0.30317
V6	0.77859	0.19961
V7	0.96172	0.19720
V8	0.86185	-0.03539
V9	0.95741	0.00231
V10	0.85431	0.18508
V11	0.77262	-0.63383
V12	0.82529	0.29923
V13	0.88393	0.36364
V14	0.90521	0.22742
V15	0.89313	0.05834

Variance explained by each factor

	FACTOR1	FACTOR2
Weighted	1.979799	0.330946
Unweighted	11.116703	0.956040

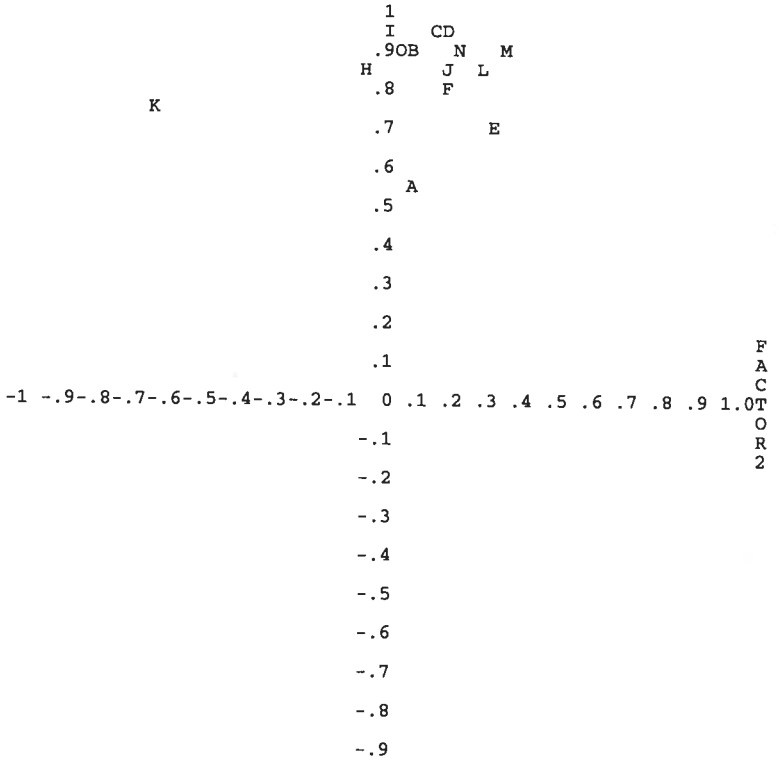
Initial Factor Method: Principal Components

Final Communality Estimates and Variable Weights

Total Communality: Weighted = 2.310745 Unweighted = 12.072743

	V1	V2	V3	V4	V5
Communality	0.329859	0.848342	0.961833	0.965582	0.577414
Weight	0.001811	0.011906	0.233770	0.526974	0.025211

Plot of Factor Pattern for FACTOR1 and FACTOR2



```
                                        -1

V1      =A      V2      =B      V3      =C      V4      =D      V5      =E
V6      =F      V7      =D      V8      =H      V9      =I      V10     =J
V11     =K      V12     =L      V13     =M      V14     =N      V15     =O
```

En el archivo de resultados, el **PROC PRINCOMP** da las salidas que ya se vio al estudiar el análisis de componentes principales. Lo interesante de este ejemplo es que el primer valor propio explica el 80.81% de la varianza y el segundo valor propio explica el 13.51% de la varianza, entre los dos explican el 94.32% de la varianza, lo que es razonablemente suficiente. En los vectores propios se observa que los coeficientes del primer vector propio son todos positivos, por lo que se cumple el primer requisito en el que se identifica la primera componente con el factor tamaño; mientras que los coeficientes del segundo vector propio son positivos y negativos por lo que se cumple el segundo requisito en el que se identifica la segunda componente con el factor forma

El **PROC FACTOR** da las salidas que se vio al estudiar el análisis factorial. Lógicamente los valores propios son los mismos del análisis de componentes principales. En el **scree** se observa que es buena la elección de dos factores pues hay un punto de inflexión entre la pendiente del primero y segundo factor y la pendiente del tercero y cuarto factor. En la matriz factorial se observa que las variables que tienen una mayor correlación con el factor tamaño es la medida 4 y la 7, que son las que tienen mayores comunalidades, por lo que la medida de estas dos variables nos puede servir para el estudio del tamaño del ala. Y las variables que tienen una mayor correlación con el factor forma son la medida 13 en sentido positivo y la medida 11 en sentido negativo, esto es, si aumenta el factor forma aumenta la medida 13 y disminuye la medida 11, y viceversa. Las dos medidas correlacionadas con el factor forma tienen una comunalidad muy alta, por lo que pueden servir para estudiar dicho factor.

El **PROC PLOT** es la representación de los individuos en las dos primeras componentes, como se ve, el primer factor discrimina tanto los sexos como la temperatura de desarrollo. Hacia el lado negativo (izquierda) de la primera componente están los individuos criados a 30°C y hacia el lado positivo los individuos criados a 25°C, esto es, que los individuos criados a 30°C son más pequeños que los criados a 25°C. Dentro de los individuos criados a una temperatura determinada están a la izquierda los machos y a la derecha las hembras, esto es, los machos son más pequeños que las hembras. Como se ve también, parte de los machos de 25°C se solapan con parte de las hembras de 30°C, esto es, parte de los mayores de los pequeños se solapan con parte de los pequeños de los mayores.

Con respecto a la segunda componente se observa que también discrimina tanto las temperaturas de desarrollo como los sexos. Hacia el lado negativo (abajo) de la segunda componente están los individuos criados a 30°C y hacia el lado positivo están los individuos criados a 25°C, esto es, los individuos criados a 30°C tienen una forma diferente a los criados a 25°C. Dentro de los individuos criados a una temperatura determinada están más abajo los machos y mas arriba las hembras, esto quiere decir que las alas de las hembras tienen formas diferentes que las alas de los machos.

También se observan en esta gráfica que hay tres machos y cuatro hembras de 25°C muy hacia abajo de la segunda componente principal, así como un macho de 30°C y una hembra de 25°C que está muy hacia arriba, esto puede ser debido a que existen datos erróneos (los dos que están hacia arriba) o a que se tienen individuos pertenecientes a otra población (los siete que están hacia abajo) como se va a estudiar en el siguiente epígrafe.

Detección de datos anómalos.-

Si se tipifican los componentes principales darán como resultado que se transforma la nube de dispersión, sensiblemente elipsoidal, en una nube esférica *p-dimensional* con el centro de coordenadas en el centro de la esfera.

Una nube de puntos transformada de esta manera, las distancias de todos los puntos al origen de coordenadas será 1, 2 y 2.6 con un 68%, 95% y 99%, respectivamente, de probabilidad. Los puntos que superen esta distancia podrán ser considerados puntos pertenecientes a otra distribución multinormal con un nivel de significación del 0.32, 0.05 y 0.001, respectivamente.

Ejemplo.-

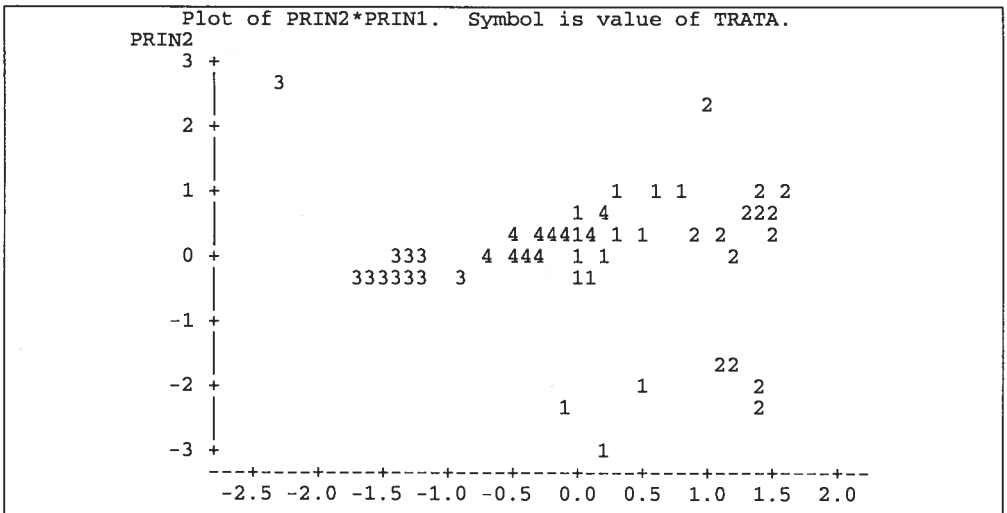
Se van a estudiar las dos bases de datos utilizadas en este capitulo, a saber, la del gasto en productos alimenticios de diferentes familias y la de las medidas del ala.

Con respecto a las alas:

Archivo de programa *SAS* (C18-6.SAS).-

```
title 'datos anómalos';
Options ls=75 ps=30;
data da;
infile 'C18-5.dat';
input trata v1-v15;
proc princomp noprint cov std out=prefix;
var v1-v15;
run;
proc plot;
plot prin2*prin1=trata /vspace=3 hspace=5;
run;
data da2;
set prefix;
dis=sqrt(prin1*prin1+prin2*prin2);
run;
proc print;
var trata prin1 prin2 dis;
run;
proc princomp noprint std data=da out=prefix;
var v1-v15;
run;
proc plot;
plot prin2*prin1=trata /vspace=3 hspace=5;
run;
data da3;
set prefix;
dis=sqrt(prin1*prin1+prin2*prin2);
run;
proc print;
var trata prin1 prin2 dis;
run;
```

Archivo de resultados (C18-6.LST).-

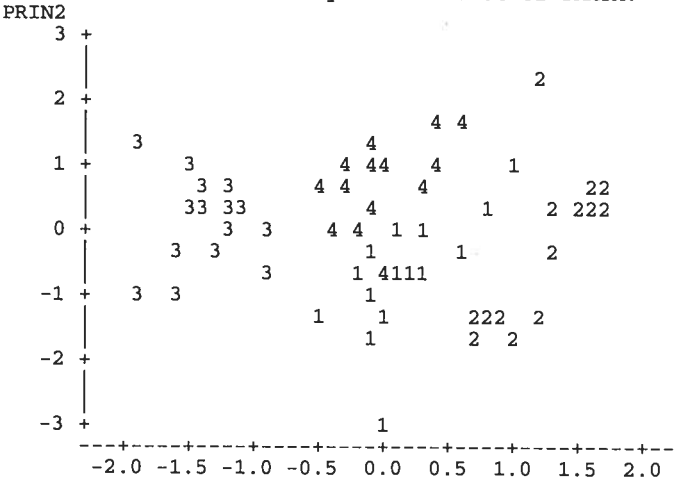


NOTE: 11 obs hidden.

OBS	TRATA	PRIN1	PRIN2	DIS
1	1	0.24748	-3.11308	3.12290
2	1	-0.09008	-2.33892	2.34066
3	1	0.52992	-2.04480	2.11235
4	1	0.10964	-0.18095	0.21157
5	1	-0.01975	0.56538	0.56573
6	1	0.63406	0.97234	1.16081
7	1	-0.02272	0.01155	0.02549
8	1	0.21978	-0.04723	0.22480
9	1	0.04599	0.22685	0.23146
10	1	0.52479	0.26611	0.58840
11	1	-0.02267	0.32086	0.32166
12	1	0.84524	0.86314	1.20808
13	1	0.00510	-0.39535	0.39538
14	1	0.25667	0.48964	0.55283
15	1	0.33179	0.93896	0.99586
16	2	1.22036	-1.71342	2.10359
17	2	1.10388	-1.78706	2.10050
18	2	1.35400	-2.00553	2.41981
19	2	1.58545	1.13939	1.95240
20	2	1.46366	0.44166	1.52885
21	2	0.96032	2.23662	2.43407
22	2	1.43662	-2.41418	2.80930
23	2	1.29024	0.62243	1.43253
24	2	1.38552	1.14725	1.79885
25	2	1.53276	0.82766	1.74194
26	2	1.09725	0.49389	1.20328
27	2	0.89495	0.43672	0.99582
28	2	1.18311	-0.02413	1.18335
29	2	1.51890	0.81332	1.72295
30	2	1.37640	0.81611	1.60016
31	3	-1.55566	-0.30751	1.58576

32	3	-1.19022	-0.46664	1.27843
33	3	-1.20123	-0.47016	1.28996
34	3	-1.22655	-0.00097	1.22655
35	3	-1.26721	-0.41125	1.33227
36	3	-2.34319	2.63633	3.52715
37	3	-1.41906	-0.37082	1.46671
38	3	-1.34296	-0.04717	1.34379
39	3	-1.44799	-0.08111	1.45026
40	3	-1.32219	-0.22885	1.34185
41	3	-0.88712	-0.29793	0.93581
42	3	-1.70573	-0.41861	1.75634
43	3	-1.45126	-0.28286	1.47857
44	3	-1.20625	-0.05129	1.20734
45	3	-0.94001	-0.24823	0.97223
46	4	-0.45498	0.11453	0.46917
47	4	-0.11850	0.26468	0.29000
48	4	0.23084	0.60542	0.64793
49	4	0.30654	0.46257	0.55492
50	4	0.07327	0.20341	0.21620
51	4	-0.47631	0.32495	0.57660
52	4	-0.22778	0.21252	0.31152
53	4	-0.36104	0.15479	0.39283
54	4	-0.28989	-0.03042	0.29148
55	4	0.25762	0.34829	0.43321
56	4	-0.19408	0.26894	0.33165
57	4	-0.27986	-0.04273	0.28310
58	4	0.04943	0.17364	0.18053
59	4	-0.66520	0.09388	0.67179
60	4	-0.34210	0.32735	0.47349

Plot of PRIN2*PRIN1. Symbol is value of TRATA.



NOTE: 2 obs hidden.

OBS	TRATA	PRIN1	PRIN2	DIS
1	1	-0.01753	-2.98306	2.98311
2	1	-0.50782	-1.35580	1.44778
3	1	0.28946	-0.56272	0.63281
4	1	-0.01717	-1.33153	1.33164
5	1	-0.12654	-0.24801	0.27842
6	1	1.04089	1.00603	1.44760
7	1	-0.08161	-0.83334	0.83732
8	1	0.08984	-0.81874	0.82365
9	1	0.07058	-0.03883	0.08056
10	1	0.62062	-0.26733	0.67574

11	1	-0.15333	-0.75543	0.77083
12	1	0.79886	0.41152	0.89863
13	1	-0.08088	-1.65321	1.65518
14	1	0.20851	-0.59701	0.63238
15	1	0.29467	0.01113	0.29488
16	2	0.73085	-1.46259	1.63502
17	2	0.72312	-1.51558	1.67925
18	2	1.27264	-0.46469	1.35483
19	2	1.66520	0.29209	1.69063
20	2	1.60216	0.76005	1.77330
21	2	1.21032	2.32653	2.62252
22	2	0.98413	-1.77575	2.03022
23	2	1.20354	-1.23273	1.72283
24	2	1.46965	0.46030	1.54005
25	2	1.63973	0.44350	1.69865
26	2	1.28750	0.36442	1.33808
27	2	0.77430	-1.38170	1.58387
28	2	0.92342	-1.39809	1.67552
29	2	1.74453	0.80219	1.92013
30	2	1.57949	0.57686	1.68153
31	3	-1.48234	0.17982	1.49321
32	3	-1.12937	0.47534	1.22532
33	3	-1.16205	0.08092	1.16486
34	3	-1.17493	0.30870	1.21481
35	3	-1.37623	0.17288	1.38704
36	3	-1.92955	1.34592	2.35258
37	3	-1.57921	-0.83583	1.78676
38	3	-1.37297	0.52196	1.46884
39	3	-1.46203	0.98356	1.76208
40	3	-1.28467	-0.23430	1.30586
41	3	-0.94580	-0.54045	1.08932
42	3	-1.89675	-0.88714	2.09396
43	3	-1.58198	-0.18782	1.59309
44	3	-1.22206	0.67632	1.39673
45	3	-0.89615	0.10499	0.90227
46	4	-0.45172	0.55758	0.71759
47	4	-0.01650	1.02488	1.02501
48	4	0.42698	1.66385	1.71777
49	4	0.34091	0.65039	0.73432
50	4	0.41120	0.92035	1.00803
51	4	-0.30167	1.15719	1.19587
52	4	-0.18167	0.07752	0.19751
53	4	-0.32935	0.64179	0.72137
54	4	-0.09222	1.13601	1.13975
55	4	0.59414	1.52531	1.63694
56	4	-0.12588	0.20431	0.23998
57	4	-0.36048	-0.03951	0.36264
58	4	-0.03853	-0.50404	0.50551
59	4	-0.50468	0.77683	0.92637
60	4	-0.11357	1.26416	1.26925

Efectivamente el macho y la hembra que están aislados encima de la gráfica quedan fuera del circulo de diámetro 2, para las componentes principales tipificadas tanto de los datos originales como de los datos tipificados (matriz de covarianza y matriz de correlación) por lo que se puede afirmar que pertenecen a otra población con un 95% de confianza. Con respecto a los siete individuos que quedaban en la parte inferior se observa que los siete quedan fuera del circulo de diámetro 2 con los componentes tipificados de la matriz de covarianza, pero con la matriz de correlación solo quedan fuera del circulo tres individuos. En total, con la matriz de covarianza, quedan nueve individuos fuera del circulo de diámetro dos y con la matriz de correlación quedan fuera cinco individuos.

Si se quiere saber cuales son estos individuos, se mira en el listado que se ha elaborado a continuación con las distancias (**dis**) al lado de las componentes

principales. Una vez localizados estos individuos se puede revisar si los valores de las medidas en el archivo de datos no son los correctos por un error en la informatización de los datos, en cuyo caso se corrigen. Si no es ese el caso, estos individuos pertenecen a otra población, con una significación del 0.05.

Con respecto a las familias:

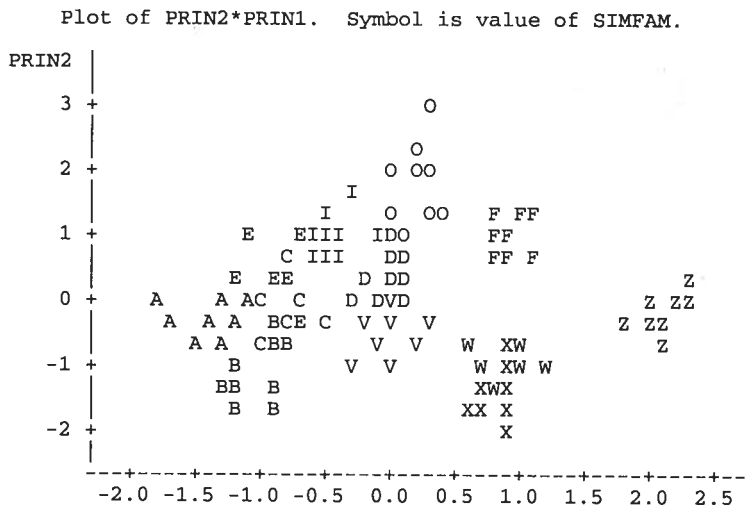
Archivo de programa *SAS* (C18-7.SAS).-

```
title 'datos anómalos';
Options ls=75 ps=30;
data da;
infile 'c18-1.dat';
input  familia $ simfam $ pan legumbre fruta carne pollos leche vino @@;
proc princomp noprint cov std out=prefix;
var pan legumbre fruta carne pollos leche vino;
run;
proc plot;
plot prin2*prin1=simfam /vspace=3 hspace=5;
run;
data da2;
set prefix;
dis=sqrt(prin1*prin1+prin2*prin2);
run;
proc print;
var simfam prin1 prin2 dis;
run;
proc princomp noprint std data=da out=prefix;
var pan legumbre fruta carne pollos leche vino;
run;
proc plot;
plot prin2*prin1=simfam /vspace=3 hspace=5;
run;
data da3;
set prefix;
dis=sqrt(prin1*prin1+prin2*prin2);
run;
proc print;
var simfam prin1 prin2 dis;
run;
```

Archivo de resultados (C18-7.LST).-

```
             Plot of PRIN2*PRIN1.  Symbol is value of SIMFAM.
          PRIN2 |
                |
              3 +
                |                O
                |                O   F F
              2 +               O        FFF
                |                     O  FF  F
                |          I O     OO                    Z   Z
              1 +     C     I                F             Z
                |   E       I I II DD    O                Z
                |           EIEI D      VO  W              ZZ  Z
              0 +     EAC C            D         W       ZZ Z
                |   B  ECBB BC    VDV VD         X
                |       AAB EC   E D V  VD   WW  X  W
             -1 +    ABAAB  BC             V VXW  XW W
                |    A  A A                      XXX
                |                                  XW
             -2 +                                  X
                |
                ---+----+----+----+----+----+----+----+----+--
                 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0  2.5
                                    PRIN1
NOTE: 19 obs hidden.
```

OBS	SIMFAM	PRIN1	PRIN2	DIS	OBS	SIMFAM	PRIN1	PRIN2	DIS
1	A	-1.12880	-1.35339	1.76234	61	E	-0.82925	0.27273	0.87295
2	I	-0.59663	0.62536	0.86431	62	O	0.02358	1.63267	1.63284
3	A	-1.42662	-1.12092	1.81431	63	E	-0.69230	-0.55558	0.88767
4	I	-0.78919	0.86163	1.16843	64	O	-0.48343	2.02120	2.07821
5	A	-1.12042	-1.12393	1.58700	65	E	-1.45542	0.50021	1.53898
6	I	-0.83844	0.80513	1.16242	66	O	0.41292	0.17857	0.44987
7	A	-1.10066	-0.94514	1.45077	67	E	-1.26477	0.01069	1.26481
8	I	-0.66563	0.44227	0.79916	68	O	-0.40474	2.17117	2.20857
9	A	-1.09273	-0.68555	1.28998	69	E	-1.10278	-0.53375	1.22515
10	I	-0.58756	0.71016	0.92171	70	O	-0.84439	1.49246	1.71477
11	A	-1.17480	-0.06692	1.17671	71	E	-0.58924	0.17413	0.61443
12	I	-1.03677	1.26628	1.63657	72	O	-0.33785	1.17897	1.22642
13	A	-1.37044	-1.32998	1.90970	73	E	-1.10087	-0.53089	1.22219
14	I	-0.45048	0.38424	0.59209	74	O	0.38667	0.82741	0.91330
15	A	-1.20026	-0.88672	1.49228	75	E	-1.17794	-0.26026	1.20635
16	I	-0.27563	0.50061	0.57147	76	O	-0.30017	1.22811	1.26426
17	A	-0.95791	-0.71512	1.19540	77	E	-0.27247	-0.69018	0.74202
18	I	-0.65061	0.44999	0.79107	78	O	-0.19382	1.17505	1.19093
19	A	-0.89153	-1.27633	1.55687	79	E	-1.31818	-0.01212	1.31823
20	I	-0.38021	0.77413	0.86247	80	O	-0.40084	2.74599	2.77510
21	B	-1.31110	-0.85054	1.56282	81	C	-0.92674	-0.77890	1.21060
22	D	-0.14969	0.53215	0.55280	82	F	0.46240	1.00796	1.10897
23	B	-0.69168	-0.42007	0.80925	83	C	-0.89066	0.02635	0.89105
24	D	-0.26753	0.75823	0.80404	84	F	0.41357	1.99004	2.03256
25	B	-0.91459	-0.43956	1.01474	85	C	-0.60255	-1.11988	1.27169
26	D	-0.06649	-0.25344	0.26202	86	F	0.48492	1.90686	1.96755
27	B	-0.88380	-0.76510	1.16896	87	C	-0.81970	0.31595	0.87848
28	D	0.11495	0.03375	0.11980	88	F	0.04221	2.28718	2.28757
29	B	-0.96766	-1.10752	1.47071	89	C	-0.57314	-0.16670	0.59689
30	D	0.04712	0.58758	0.58947	90	F	0.55970	1.88910	1.97027
31	B	-0.97815	-0.48816	1.09320	91	C	-1.13512	-0.16935	1.14769
32	D	0.31666	-0.47123	0.56774	92	F	0.15179	2.21728	2.22247
33	B	-1.33860	-0.93031	1.63013	93	C	-1.05383	0.05503	1.05527
34	D	-0.12006	-0.69251	0.70284	94	F	0.52620	0.92161	1.06125
35	B	-0.69141	-1.13974	1.33306	95	C	-0.64614	-0.65015	0.91662
36	D	0.14974	-0.05392	0.15915	96	F	0.31602	1.50396	1.53680
37	B	-1.47052	-0.44736	1.53706	97	C	-0.94175	-0.43164	1.03596
38	D	-0.32458	0.45965	0.56270	98	F	0.67580	1.63574	1.76985
39	B	-0.90698	-0.33820	0.96798	99	C	-1.29477	1.11589	1.70928
40	D	0.48678	-0.56000	0.74200	100	F	0.44397	1.62335	1.68297
41	V	0.54184	-1.14953	1.27083	101	W	0.68505	0.38786	0.78722
42	X	1.38189	-1.95342	2.39280	102	Z	1.92479	0.91032	2.12920
43	V	-0.05642	-0.76298	0.76506	103	W	1.19784	-1.07460	1.60922
44	X	1.15142	-1.18092	1.64935	104	Z	1.88286	0.57793	1.96956
45	V	0.36160	-0.76562	0.84671	105	W	0.93937	-0.98038	1.35778
46	X	1.36456	-1.31277	1.89351	106	Z	1.86446	0.37306	1.90142
47	V	0.68789	-0.85006	1.09353	107	W	1.45345	-0.81497	1.66634
48	X	1.28400	-1.22101	1.77187	108	Z	2.34965	0.23237	2.36111
49	V	-0.23248	-0.16795	0.28680	109	W	0.84710	-0.57032	1.02120
50	X	1.19384	-0.83427	1.45645	110	Z	1.78801	-0.01332	1.78806
51	V	-0.07987	-0.34283	0.35201	111	W	1.26489	-1.05778	1.64890
52	X	1.06065	-0.37904	1.12634	112	Z	1.99674	0.00929	1.99676
53	V	0.00118	-0.18252	0.18253	113	W	1.52989	-0.97166	1.81237
54	X	1.42842	-1.58226	2.13165	114	Z	2.04917	1.25027	2.40048
55	V	0.27321	0.36473	0.45571	115	W	1.02069	-0.05047	1.02194
56	X	1.24094	-1.42833	1.89211	116	Z	1.95552	0.49137	2.01631
57	V	0.14419	-0.76297	0.77647	117	W	0.92295	-0.56356	1.08140
58	X	0.81773	-0.96021	1.26123	118	Z	1.57161	1.18851	1.97041
59	V	0.24912	-0.34450	0.42513	119	W	1.47144	-1.66080	2.21887
60	X	1.23311	-0.79067	1.46483	120	Z	1.69371	0.00827	1.69373



NOTE: 26 obs hidden.

OBS	SIMFAM	PRIN1	PRIN2	DIS	OBS	SIMFAM	PRIN1	PRIN2	DIS
1	A	-1.50014	-0.60683	1.61822	61	E	-0.75952	0.31931	0.82391
2	I	-0.36189	0.86717	0.93966	62	O	0.34810	2.16125	2.18910
3	A	-1.70815	-0.46583	1.77053	63	E	-0.86991	0.38065	0.94955
4	I	-0.50186	1.10835	1.21668	64	O	0.17436	1.98484	1.99249
5	A	-1.43885	-0.35646	1.48234	65	E	-1.14453	0.91897	1.46781
6	I	-0.46167	0.63672	0.78648	66	O	0.26764	1.32976	1.35643
7	A	-1.29142	-0.63635	1.43969	67	E	-1.29477	0.03529	1.29525
8	I	-0.58797	0.73830	0.94382	68	O	0.20576	2.40237	2.41116
9	A	-1.19504	-0.27825	1.22700	69	E	-1.19674	0.27731	1.22845
10	I	-0.43113	0.91808	1.01427	70	O	-0.33186	1.68128	1.71372
11	A	-1.10385	-0.11220	1.10954	71	E	-0.68151	0.86407	1.10049
12	I	-0.52761	1.35965	1.45844	72	O	-0.01044	2.01042	2.01044
13	A	-1.75560	-0.08266	1.75755	73	E	-1.18085	-0.33536	1.22755
14	I	-0.44810	0.77562	0.89575	74	O	0.40575	1.43694	1.49313
15	A	-1.28838	-0.07095	1.29033	75	E	-1.20574	0.36417	1.25954
16	I	-0.14399	1.00703	1.01727	76	O	0.10534	1.15434	1.15913
17	A	-1.18511	-0.49043	1.28258	77	E	-0.68033	-0.34539	0.76298
18	I	-0.56079	0.85955	1.02631	78	O	0.04132	1.34162	1.34225
19	A	-1.24706	-0.34983	1.29520	79	E	-1.11689	0.04530	1.11781
20	I	-0.30310	1.55216	1.58147	80	O	0.26883	2.87726	2.88979
21	B	-1.23315	-1.64155	2.05313	81	C	-1.01917	-0.51061	1.13992
22	D	-0.12957	1.12310	1.13055	82	F	0.88298	0.66327	1.10435
23	B	-0.79973	-0.78011	1.11720	83	C	-0.73712	0.11866	0.74661
24	D	-0.00510	0.73521	0.73522	84	F	1.01846	1.26330	1.62271
25	B	-0.92964	-1.27390	1.57703	85	C	-1.09954	0.09149	1.10334
26	D	-0.06177	0.06783	0.09174	86	F	1.14646	0.76273	1.37700
27	B	-0.92781	-1.21847	1.53150	87	C	-0.46598	-0.18451	0.50118
28	D	-0.03962	0.84054	0.84147	88	F	0.84726	1.01755	1.32411
29	B	-1.19590	-1.19971	1.69395	89	C	-0.79161	0.41442	0.89352
30	D	0.06903	0.80824	0.81118	90	F	1.10376	1.39035	1.77520
31	B	-0.93471	-0.44174	1.03383	91	C	-1.10409	-0.06377	1.10593
32	D	-0.00365	0.41745	0.41747	92	F	0.86322	1.10934	1.40562
33	B	-1.32145	-1.35732	1.89435	93	C	-1.01658	0.04650	1.01764
34	D	-0.27599	-0.09422	0.29163	94	F	0.80162	0.64173	1.02685

35	B	-0.91506	-1.51865	1.77303	95	C	-0.84389	-0.30065	0.89585
36	D	0.10984	0.23190	0.25660	96	F	0.83783	1.41686	1.64605
37	B	-1.23686	-0.99526	1.58756	97	C	-0.96231	-0.04703	0.96346
38	D	-0.19880	0.40330	0.44964	98	F	1.09030	0.73966	1.31751
39	B	-0.87807	-0.80196	1.18918	99	C	-0.75480	0.74153	1.05811
40	D	0.06006	0.09057	0.10868	100	F	0.78049	1.21539	1.44442
41	V	0.02131	-0.96624	0.96648	101	W	0.88138	-0.86931	1.23796
42	X	0.70220	-1.59834	1.74579	102	Z	2.15156	0.16104	2.15758
43	V	-0.26325	-1.14507	1.17494	103	W	0.81509	-1.32742	1.55770
44	X	0.68723	-1.40961	1.56821	104	Z	2.04469	-0.02804	2.04488
45	V	-0.18509	-0.44130	0.47855	105	W	0.65278	-0.91655	1.12525
46	X	0.88083	-1.47442	1.71749	106	Z	2.04416	-0.20391	2.05431
47	V	0.21320	-0.80733	0.83501	107	W	0.98092	-0.98585	1.39072
48	X	0.87300	-1.45456	1.69643	108	Z	2.28661	-0.03101	2.28682
49	V	-0.22593	-0.23987	0.32952	109	W	0.72241	-1.17951	1.38316
50	X	0.89143	-1.60502	1.83596	110	Z	1.84755	-0.23994	1.86306
51	V	-0.20318	-0.42790	0.47368	111	W	0.94139	-1.40175	1.68853
52	X	0.91923	-0.98382	1.34643	112	Z	2.05773	-0.75210	2.19087
53	V	-0.03176	-0.29025	0.29198	113	W	1.16182	-0.86426	1.44803
54	X	0.92365	-1.87481	2.08998	114	Z	2.28676	0.42601	2.32610
55	V	0.34314	-0.31554	0.46616	115	W	0.96422	-0.80804	1.25804
56	X	0.74034	-1.67578	1.83203	116	Z	2.10314	-0.22440	2.11508
57	V	-0.10471	-0.52400	0.53436	117	W	0.64068	-0.65460	0.91596
58	X	0.61373	-1.76842	1.87189	118	Z	2.15327	0.00161	2.15327
59	V	-0.03163	-0.06957	0.07642	119	W	0.80585	-1.37981	1.59790
60	X	0.85185	-0.68081	1.09048	120	Z	1.81080	-0.16820	1.81860

Se observa que en las componentes tipificadas de la matriz de covarianza hay 7 familias de directivos fuera del circulo de diámetro 2 y que todas las familias de directivos de 5 hijos (las **Z**) que no están fuera del circulo están próximas a este. También hay familias de trabajadores con cinco hijos (tres **O**) y familias de oficinistas con cinco hijos (tres **F**) que están fuera del circulo de diámetro dos; así como dos familias de directivos con tres hijos (**X**) y una familia de directivos con dos hijos (**W**). En total quedan fuera 13 familias.

Con la matriz de correlaciones quedan 9 familias de directivos fuera del circulo de diámetro dos, ocho de ellas son de cinco hijos, solo una de estas queda dentro del circulo. También quedan fuera cuatro familias de trabajadores con cinco hijos (**O**). En total quedan fuera 14 familias.

Como en el ejemplo anterior, para saber cuales son estos individuos, se mira en el listado que se ha elaborado a continuación con las distancias (**dis**) al lado de las componentes principales. Una vez localizados estos individuos se puede revisar si los valores de las medidas en el archivo de datos no son los correctos por un error en la informatización de los datos, en cuyo caso se corrigen. Si no es ese el caso, estos individuos pertenecen a otra población, con una significación del 0.05.

Puede ser de interés el realizar este análisis con los factores rotados para comprobar si se mantienen los datos anómalos detectados, en ese caso, como el procedimiento que rota los factores es el procedimiento *FACTOR* el programa sería, para los datos de las alas

Archivo de programa *SAS* (C18-8.SAS).-

```
title 'datos anómalos con factores rotados';
Options ls=75 ps=30;
Data da;
Infile 'c18-5.dat';
Input indi v1-v15;
proc factor cov data=da n=2 rotate=varimax score outstat=fact;
var v1-v15;
run;
proc score data=da score=fact out=scores;
var v1-v15;
run;
proc plot;
plot factor2*factor1=indi/vspace=3 hspace=5;
run;
data da2;
set scores;
dis=sqrt(factor1*factor1+factor2*factor2);
run;
proc print;
var indi factor1 factor2 dis;
run;
proc factor data=da n=2 rotate=varimax score outstat=fact;
var v1-v15;
run;
proc score data=da score=fact out=scores;
var v1-v15;
run;
proc plot;
plot factor2*factor1=indi/vspace=3 hspace=5;
run;
data da4;
set scores;
dis=sqrt(factor1*factor1+factor2*factor2);
run;
proc print;
var indi factor1 factor2 dis;
run;
```

Este programa calcula las componentes principales tipificadas de los factores rotados tanto con la matriz de covarianzas como con la matriz de correlaciones. Aunque el valor de las componentes principales no son los mismos, las distancias si son las mismas, por lo que los resultados de exclusión son exactamente los mismos que los obtenidos anteriormente.

Para los datos de las familias ocurre exactamente lo mismo.

Análisis de Varianza de las Componentes Principales.-

Habitualmente, cuando se miden unas variables se hace respondiendo a varios objetivos, por ello, como se observa en los ejemplos vistos hasta el momento, además de las variables *independientes* o *explicativas* también se han medido variables *dependientes* que habitualmente son nominales o de clase y que pueden ser factores de clasificación y/o factores ambientales y/o tratamientos experimentales.

Se han medido estos factores o fuentes de variación porque se quiere saber si son significativos para la variación de los caracteres medidos. En los capítulos 4 al 10, cuando se estudió el Análisis Multivariante de la Varianza, se realizó este estudio.

Pero ¿son significativas estas fuentes de variación en la variabilidad observada en las componentes principales? o ¿ se ven afectadas las variables que no han podido ser medidas directamente, a las que se han denominado **factores,** por estas fuentes de variación?

Dado que las componentes principales son variables (función lineal de las variables originales) que incluyen gran parte de la información de las variables originales, el análisis de la varianza de las componentes principales nos informará del grado de influencia de las diferentes fuentes de variación o factores del ANOVA en los *factores* o *componentes principales.*

Ejemplo.-

En el análisis de ***tamaño*** y ***forma*** se vio la influencia de estos dos factores en las 15 medidas del ala de *Drosophila.* Pero recuérdese que también se ha medido el sexo del individuo y la temperatura en la que se ha criado porque estas dos fuentes de variación pueden influir en la variabilidad de las 15 medidas. Ahora la pregunta puede ser: ¿influye el sexo y la temperatura en el factor **tamaño** y/o en el factor **forma**? y ¿de que manera influyen?

Archivo de programa *SAS* (C18-9.SAS).-

```
title 'ANOVA de las componentes principales';
Options ls=75 ps=60;
data alas;
infile 'c18-5.dat';
input trata v1-v15;
if trata=1 then sexo="M";
if trata=3 then sexo="M";
if trata=2 then sexo="H";
if trata=4 then sexo="H";
if trata=1 then tempe="25º";
if trata=3 then tempe="30º";
if trata=2 then tempe="25º";
if trata=4 then tempe="30º";
proc princomp out=scores noprint;
var v1-v15;
run;
proc anova data=scores;
class sexo tempe;
model prin1-prin2 = sexo tempe sexo*tempe;
means sexo tempe/ duncan;
run;
proc factor data=alas n=4 rotate=varimax score outstat=fact;
var v1-v15;
run;
proc score data=alas score=fact out=scores;
var v1-v15;
run;
proc anova data=scores;
class sexo tempe;
model factor1-factor2 = sexo tempe sexo*tempe;
means sexo tempe / duncan;
run;
proc glm data= alas;
class sexo tempe ;
model v1-v15 = sexo tempe sexo*tempe / nouni ;
manova h=sexo tempe sexo*tempe;
run;
```

Archivo de resultados (C18-9.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: PRIN1
                                  Sum of          Mean
Source                     DF        Squares        Square  F Value     Pr > F
Model                       3     599.084206    199.694735   139.23     0.0001
Error                      56      80.317161      1.434235
Corrected Total            59     679.401366

                   R-Square          C.V.       Root MSE          PRIN1 Mean
                   0.881782       9999.99        1.19760             0.00000

Source                     DF       Anova SS   Mean Square  F Value     Pr > F
SEXO                        1     250.599992    250.599992   174.73     0.0001
TEMPE                       1     346.303299    346.303299   241.46     0.0001
SEXO*TEMPE                  1       2.180915      2.180915     1.52     0.2227

                    Analysis of Variance Procedure
Dependent Variable: PRIN2
                                  Sum of          Mean
Source                     DF        Squares        Square  F Value     Pr > F
Model                       3      15.5603204     5.1867735     6.85     0.0005
Error                      56      42.4136606     0.7573868
Corrected Total            59      57.9739810

                   R-Square          C.V.       Root MSE          PRIN2 Mean
                   0.268402      -9999.99        0.87028            -0.00000

Source                     DF       Anova SS   Mean Square  F Value     Pr > F
SEXO                        1       4.0382075     4.0382075     5.33     0.0247
TEMPE                       1      11.4512184    11.4512184    15.12     0.0003
SEXO*TEMPE                  1       0.0708945     0.0708945     0.09     0.7608

        Duncan's Multiple Range Test for variable: PRIN1
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 1.434235
                   Number of Means      2
                   Critical Range   .6194
Means with the same letter are not significantly different.

      Duncan Grouping              Mean      N  SEXO
                   A            2.0437     30  H
                   B           -2.0437     30  M

        Duncan's Multiple Range Test for variable: PRIN2
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 0.757387
                   Number of Means      2
                   Critical Range   .4501
Means with the same letter are not significantly different.

      Duncan Grouping              Mean      N  SEXO
                   A            0.2594     30  H
                   B           -0.2594     30  M
```

Duncan's Multiple Range Test for variable: PRIN1

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 56 MSE= 1.434235
Number of Means 2
Critical Range .6194

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	TEMPE
A	2.4024	30	25º
B	-2.4024	30	30º

Duncan's Multiple Range Test for variable: PRIN2

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 56 MSE= 0.757387
Number of Means 2
Critical Range .4501

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	TEMPE
A	0.4369	30	30º
B	-0.4369	30	25º

Initial Factor Method: Principal Components

Analysis of Variance Procedure
Dependent Variable: FACTOR1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	31.6596973	10.5532324	21.62	0.0001
Error	56	27.3403027	0.4882197		
Corrected Total	59	59.0000000			

R-Square	C.V.	Root MSE	FACTOR1 Mean
0.536605	9999.99	0.69873	0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
SEXO	1	3.5664454	3.5664454	7.31	0.0091
TEMPE	1	27.8760460	27.8760460	57.10	0.0001
SEXO*TEMPE	1	0.2172059	0.2172059	0.44	0.5075

Analysis of Variance Procedure
Dependent Variable: FACTOR2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	17.2895247	5.7631749	7.74	0.0002
Error	56	41.7104753	0.7448299		
Corrected Total	59	59.0000000			

R-Square	C.V.	Root MSE	FACTOR2 Mean
0.293043	-9999.99	0.86304	-0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
SEXO	1	13.9153372	13.9153372	18.68	0.0001
TEMPE	1	1.1953903	1.1953903	1.60	0.2105
SEXO*TEMPE	1	2.1787972	2.1787972	2.93	0.0927

```
    Duncan's Multiple Range Test for variable: FACTOR1
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 0.48822
                  Number of Means      2
                  Critical Range   .3614
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  SEXO
                  A            0.2438      30  H
                  B           -0.2438      30  M

    Duncan's Multiple Range Test for variable: FACTOR2
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 0.74483
                  Number of Means      2
                  Critical Range   .4464
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  SEXO
                  A            0.4816      30  H
                  B           -0.4816      30  M

    Duncan's Multiple Range Test for variable: FACTOR1
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 0.48822
                  Number of Means      2
                  Critical Range   .3614
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  TEMPE
                  A            0.6816      30  25ª
                  B           -0.6816      30  30ª

    Duncan's Multiple Range Test for variable: FACTOR2
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
            Alpha= 0.05  df= 56  MSE= 0.74483
                  Number of Means      2
                  Critical Range   .4464
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  TEMPE
                  A            0.1411      30  25ª
                  A           -0.1411      30  30ª
             General Linear Models Procedure
            Multivariate Analysis of Variance

     Manova Test Criteria and Exact F Statistics for
         the Hypothesis of no Overall SEXO Effect
H = Type III SS&CP Matrix for SEXO   E = Error SS&CP Matrix
                  S=1    M=6.5    N=20

Statistic                   Value          F       Num DF     Den DF  Pr > F
Wilks' Lambda           0.13375510    18.1338           15         42  0.0001
Pillai's Trace          0.86624490    18.1338           15         42  0.0001
Hotelling-Lawley Trace  6.47635046    18.1338           15         42  0.0001
Roy's Greatest Root     6.47635046    18.1338           15         42  0.0001
```

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall TEMPE Effect
H = Type III SS&CP Matrix for TEMPE E = Error SS&CP Matrix
S=1 M=6.5 N=20

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.05621256	47.0109	15	42	0.0001
Pillai's Trace	0.94378744	47.0109	15	42	0.0001
Hotelling-Lawley Trace	16.78961746	47.0109	15	42	0.0001
Roy's Greatest Root	16.78961746	47.0109	15	42	0.0001

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall SEXO*TEMPE Effect
H = Type III SS&CP Matrix for SEXO*TEMPE E = Error SS&CP Matrix
S=1 M=6.5 N=20

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.82287951	0.6027	15	42	0.8559
Pillai's Trace	0.17712049	0.6027	15	42	0.8559
Hotelling-Lawley Trace	0.21524474	0.6027	15	42	0.8559
Roy's Greatest Root	0.21524474	0.6027	15	42	0.8559

En el archivo de resultados, la primera salida es el ANOVA de las dos primeras componentes principales que se corresponden con el factor **tamaño** y el factor **forma**, respectivamente. En el **PROC PRINCOMP** se ha especificado la opción **NOPRINT** para que no se impriman los resultados, ya conocidos, del análisis de componentes principales.

Como es de esperar, el coeficiente de determinación de la primera componente es mayor (88%) que el de la segunda componente (27%.) Tanto el factor tamaño como el factor forma están influenciados significativamente por el **sexo** y por la **temperatura,** no habiendo interacción entre éstos. Pero cuando se observa el sentido de esta influencia con las salidas del análisis *Duncan* se observa que el factor temperatura tiene una influencia inversa en el tamaño (son mayores las de 25ºC) que en la forma (son mayores las de 30ºC), por lo que sería conveniente rotar las componentes con objeto de clarificar estos resultados.

En la segunda parte del programa se ha utilizado el **PROC FACTOR** con objeto de poder hacer la rotación, en este caso la **VARIMAX**. El procedimiento FACTOR no tiene la opción NOPRINT, por lo que no podemos evitar que salgan los resultados ya conocidos de este análisis.

Si nos vamos directamente a la salida del **PROC ANOVA** de los factores rotados se observa que son significativos tanto el sexo como la temperatura para el factor tamaño mientras que para el factor forma solo es significativo el sexo. Por lo que la conclusión es que, en el tamaño, influyen tanto la temperatura como el sexo mientras que en la forma influye solo el sexo.

También se ha incluido un análisis multivariante de la varianza en el que se ha especificado la opción **NOUNI** para que no den los resultados de los análisis univariantes. En el archivo se salida se observa que las lambdas de Wilk valen: para el factor sexo, 0.1337 cuyo valor *F* es significativo al 0.001; para el factor temperatura, 0.0562 cuyo valor de *F* es significativo al 0.001; y para la interacción sexo*temperatura,

0.8229, cuyo valor de *F* es no significativo. por lo que se concluye que tanto el sexo como la temperatura son significativos en el conjunto de las 15 variables, y no existe interacción entre ellos.

Aunque no hay interacción entre el sexo y la temperatura, puede ser de interés el analizar los cuatro tratamiento directamente, estos son: tratamiento 1, machos criados a 25ºC; tratamiento 2, hembras criadas a 25ºC; tratamiento 3, machos criados a 30ºC; tratamiento 4, hembras criadas a 30ºC. En ese caso el programa sería

Archivo de programa *SAS* (C18-10.SAS).-

```
title 'ANOVA de las componentes principales';
Options ls=75 ps=60;
data alas;
infile 'c18-5.dat';
input trata v1-v15;
proc princomp out=scores noprint;
var v1-v15;
run;
proc anova data=scores;
class trata ;
model prin1-prin2 = trata  ;
means trata / duncan;
run;
proc factor data=alas n=4 rotate=varimax score outstat=fact;
var v1-v15;
run;
proc score data=alas score=fact out=scores;
var v1-v15;
run;
proc anova data=scores;
class trata ;
model factor1-factor2 = trata  ;
means trata  / duncan;
run;
proc glm data= alas;
class trata ;
model v1-v15 = trata / nouni ;
manova h=trata;
run;
```

Archivo de resultados (C18-10.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: PRIN1
                                   Sum of          Mean
Source                    DF      Squares        Square  F Value     Pr > F
Model                      3   599.084206   199.694735   139.23     0.0001
Error                     56    80.317161     1.434235
Corrected Total           59   679.401366

                  R-Square          C.V.      Root MSE          PRIN1 Mean
                  0.881782       9999.99       1.19760             0.00000

Source                    DF       Anova SS  Mean Square  F Value     Pr > F
TRATA                      3   599.084206   199.694735   139.23     0.0001

                    Analysis of Variance Procedure
Dependent Variable: PRIN2
                                   Sum of          Mean
Source                    DF      Squares        Square  F Value     Pr > F
Model                      3   15.5603204    5.1867735     6.85     0.0005
```

```
Error                      56      42.4136606       0.7573868
Corrected Total            59      57.9739810

                  R-Square          C.V.       Root MSE          PRIN2 Mean
                  0.268402      -9999.99        0.87028            -0.00000


Source                     DF       Anova SS   Mean Square  F Value  Pr > F
TRATA                       3     15.5603204     5.1867735     6.85    0.0005

          Duncan's Multiple Range Test for variable: PRIN1
        NOTE: This test controls the type I comparisonwise error rate, not
              the experimentwise error rate
                   Alpha= 0.05  df= 56  MSE= 1.434235
                    Number of Means      2      3      4
                     Critical Range   .8760  .9215  .9514
        Means with the same letter are not significantly different.

              Duncan Grouping              Mean      N  TRATA
                           A             4.2555     15  2
                           B             0.5494     15  1
                           B            -0.1681     15  4
                           C            -4.6368     15  3

          Duncan's Multiple Range Test for variable: PRIN2
        NOTE: This test controls the type I comparisonwise error rate, not
              the experimentwise error rate
                    Alpha= 0.05  df= 56  MSE= 0.757387
                     Number of Means      2      3      4
                     Critical Range   .6366  .6696  .6914
        Means with the same letter are not significantly different.

              Duncan Grouping              Mean      N  TRATA
                           A             0.7307     15  4
                    B      A             0.1431     15  3
                    B      C            -0.2118     15  2
                           C            -0.6619     15  1

Initial Factor Method: Principal Components
                        Analysis of Variance Procedure
Dependent Variable: FACTOR1
                                    Sum of          Mean
Source                     DF       Squares        Square  F Value     Pr > F
Model                       3     31.6596973    10.5532324    21.62     0.0001
Error                      56     27.3403027     0.4882197
Corrected Total            59     59.0000000

                  R-Square          C.V.       Root MSE        FACTOR1 Mean
                  0.536605       9999.99        0.69873             0.00000
Source                     DF       Anova SS   Mean Square  F Value     Pr > F
TRATA                       3     31.6596973    10.5532324    21.62     0.0001

                        Analysis of Variance Procedure
Dependent Variable: FACTOR2
                                    Sum of          Mean
Source                     DF       Squares        Square  F Value     Pr > F
Model                       3     17.2895247     5.7631749     7.74     0.0002
Error                      56     41.7104753     0.7448299
Corrected Total            59     59.0000000

                  R-Square          C.V.       Root MSE        FACTOR2 Mean
                  0.293043      -9999.99        0.86304            -0.00000
Source                     DF       Anova SS   Mean Square  F Value     Pr > F
TRATA                       3     17.2895247     5.7631749     7.74     0.0002
```

```
        Duncan's Multiple Range Test for variable: FACTOR1
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
              Alpha= 0.05  df= 56  MSE= 0.48822
              Number of Means      2      3      4
              Critical Range   .5111  .5376  .5551
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  TRATA
                   A            0.9856      15  2
                   B            0.3776      15  1
                   C           -0.4980      15  4
                   C           -0.8653      15  3

                Analysis of Variance Procedure
        Duncan's Multiple Range Test for variable: FACTOR2
NOTE: This test controls the type I comparisonwise error rate, not
      the experimentwise error rate
              Alpha= 0.05  df= 56  MSE= 0.74483
              Number of Means      2      3      4
              Critical Range   .6313  .6641  .6856
Means with the same letter are not significantly different.

      Duncan Grouping              Mean       N  TRATA
                   A            0.5310      15  4
             B     A            0.4322      15  2
             B                 -0.1499      15  1
                   C           -0.8133      15  3

                 General Linear Models Procedure
                Multivariate Analysis of Variance

          Manova Test Criteria and F Approximations for
             the Hypothesis of no Overall TRATA Effect
    H = Type III SS&CP Matrix for TRATA    E = Error SS&CP Matrix

                        S=3     M=5.5     N=20

Statistic                    Value          F      Num DF     Den DF  Pr > F
Wilks' Lambda           0.01769970     8.0585          45  125.5519  0.0001
Pillai's Trace          1.63050767     3.4924          45       132  0.0001
Hotelling-Lawley Trace 23.48121267    21.2201          45       122  0.0001
Roy's Greatest Root    22.25903934    65.2932          15        44  0.0001

      NOTE: F Statistic for Roy's Greatest Root is an upper bound.
```

En el archivo de resultados, la primera salida es el ANOVA de las dos primeras componentes principales que se corresponden con el factor **tamaño** y el factor **forma**, respectivamente. En el **PROC PRINCOMP** se ha especificado la opción **NOPRINT** para que no se impriman los resultados, ya conocidos, del análisis de componentes principales.

Como es de esperar, el coeficiente de determinación de la primera componente es mayor (88%) que el de la segunda componente (27%). Tanto el tamaño como la forma están influenciados significativamente por los cuatro tratamientos. Pero cuando se observa el sentido de esta influencia con las salidas del análisis *Duncan* se observa que, para el tamaño, los cuatro tratamientos están claramente diferenciados mientras que para la forma lo claramente diferenciados son las hembras a 30ºC de los machos a

25ºC, por lo que sería conveniente rotar las componentes con objeto de clarificar estos resultados.

En la segunda parte del programa se ha utilizado el **PROC FACTOR** con objeto de poder hacer la rotación, en este caso la **VARIMAX**. El procedimiento FACTOR no tiene la opción NOPRINT, por lo que no podemos evitar que salgan los resultados ya conocidos de este análisis.

Si nos vamos directamente a la salida del **PROC ANOVA** de los factores rotados se observa que los cuatro tratamientos son significativos tanto para el factor tamaño como para el factor forma, pero cuando se observa el sentido de esta influencia con las salidas del análisis *Duncan* se observa que, para el tamaño, los cuatro tratamientos están claramente diferenciados mientras que para la forma lo claramente diferenciados son las hembras (a cualquier temperatura) de los machos (también a cualquier temperatura). Por lo que la conclusión es que en el factor tamaño influyen tanto la temperatura como el sexo mientras que para el factor forma influye solo el sexo.

También se ha incluido un análisis multivariante de la varianza en el que se ha especificado la opción **NOUNI** para que no den los resultados de los análisis univariantes,. En el archivo se salida se observa que la lambda de Wilk vale 0.0177 cuyo valor *F* es significativo al 0.001, por lo que se concluye que los tratamientos son significativos en el conjunto de las 15 variables.

Ejemplo.-

En el análisis del ***factor cultural*** y el ***factor socioeconómico*** en el consumo de siete productos o categorías de productos alimenticios se vio la influencia de estos dos factores en las siete medidas. Pero recuérdese que también se ha medido el nivel profesional del cabeza de familia y el número de hijos de cada familia porque estas dos fuentes de variación pueden influir en la variabilidad de las siete medidas. Ahora la pregunta puede ser: ¿influye el nivel profesional del cabeza de familia y el número de hijos factor **cultural** y/o en el factor **socioeconómico**? y ¿de que manera influyen?

Archivo de programa *SAS* (C18-11.SAS).-

```
title 'ANOVA de las componentes principales';
Options ls=75 ps=60;
Data vino;
Infile 'c18-1.dat';
Input  famil $ simb $ pan legumbre fruta carne pollos leche vino @@;
if famil="t2" or famil="o2" or famil="d2" then nhijos=2;
if famil="t3" or famil="o3" or famil="d3" then nhijos=3;
if famil="t4" or famil="o4" or famil="d4" then nhijos=4;
if famil="t5" or famil="o5" or famil="d5" then nhijos=5;
if famil="t2" or famil="t3" or famil="t4" or famil="t5" then prof="trab";
if famil="o2" or famil="o3" or famil="o4" or famil="o5" then prof="ofic";
if famil="d2" or famil="d3" or famil="d4" or famil="d5" then prof="dire";
proc princomp out=scores noprint;
var pan legumbre fruta carne pollos leche vino;
run;
proc anova data=scores;
class prof nhijos;
model prin1-prin2 = prof nhijos prof*nhijos;
means prof nhijos / duncan;
run;
proc factor data=vino n=4 rotate=varimax score outstat=fact;
var pan legumbre fruta carne pollos leche vino;
run;
proc score data=vino score=fact out=scores;
var pan legumbre fruta carne pollos leche vino;
run;
proc anova data=scores;
class prof nhijos;
model factor1-factor2 = prof nhijos prof*nhijos;
means prof nhijos / duncan;
run;
```

Archivo de resultados (C18-11.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: PRIN1
                                   Sum of          Mean
Source                     DF      Squares        Square    F Value    Pr > F
Model                      11   437.057740     39.732522     330.09    0.0001
Error                     108    12.999964      0.120370
Corrected Total           119   450.057703

                  R-Square          C.V.      Root MSE          PRIN1 Mean
                  0.971115       9999.99       0.34694             0.00000

Source                     DF       Anova SS  Mean Square  F Value    Pr > F
PROF                        2   206.583582    103.291791     858.12    0.0001
NHIJOS                      3   217.041326     72.347109     601.04    0.0001
PROF*NHIJOS                 6    13.432831      2.238805      18.60    0.0001

                    Analysis of Variance Procedure
Dependent Variable: PRIN2
                                   Sum of          Mean
Source                     DF      Squares        Square    F Value    Pr > F
Model                      11   168.490648     15.317332      74.80    0.0001
Error                     108    22.116511      0.204783
Corrected Total           119   190.607158

                  R-Square          C.V.      Root MSE          PRIN2 Mean
                  0.883968       9999.99       0.45253             0.00000
```

Source	DF	Anova SS	Mean Square	F Value	Pr > F
PROF	2	69.3868785	34.6934392	169.42	0.0001
NHIJOS	3	64.7030641	21.5676880	105.32	0.0001
PROF*NHIJOS	6	34.4007052	5.7334509	28.00	0.0001

Duncan's Multiple Range Test for variable: PRIN1

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.12037

Number of Means 2 3

Critical Range .1538 .1618

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	PROF
A	1.79734	40	dire
B	-0.49934	40	ofic
C	-1.29800	40	trab

Duncan's Multiple Range Test for variable: PRIN2

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.204783

Number of Means 2 3

Critical Range .2006 .2111

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	PROF
A	0.8631	40	trab
B	0.1241	40	ofic
C	-0.9871	40	dire

Duncan's Multiple Range Test for variable: PRIN1

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.12037

Number of Means 2 3 4

Critical Range .1776 .1869 .1931

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	NHIJOS
A	2.05062	30	5
B	0.21261	30	4
C	-0.67154	30	3
D	-1.59169	30	2

Duncan's Multiple Range Test for variable: PRIN2

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.204783

Number of Means 2 3 4

Critical Range .2316 .2437 .2518

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	NHIJOS
A	1.1619	30	5
B	-0.0033	30	4
C	-0.3188	30	3
D	-0.8397	30	2

Analysis of Variance Procedure

Dependent Variable: FACTOR1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	113.167345	10.287940	190.50	0.0001
Error	108	5.832655	0.054006		
Corrected Total	119	119.000000			

R-Square	C.V.	Root MSE	FACTOR1 Mean

```
                0.950986        9999.99        0.23239              0.00000

Source                  DF       Anova SS    Mean Square  F Value     Pr > F
PROF                     2     97.0598843     48.5299422   898.60     0.0001
NHIJOS                   3     11.3296853      3.7765618    69.93     0.0001
PROF*NHIJOS              6      4.7777751      0.7962959    14.74     0.0001
                    Analysis of Variance Procedure
Dependent Variable: FACTOR2
                                   Sum of          Mean
Source                  DF       Squares        Square  F Value     Pr > F
Model                   11    79.7348772     7.2486252    19.94     0.0001
Error                  108    39.2651228     0.3635660
Corrected Total        119   119.0000000

                R-Square          C.V.       Root MSE          FACTOR2 Mean
                0.670041       -9999.99        0.60296              -0.00000

Source                  DF       Anova SS    Mean Square  F Value     Pr > F
PROF                     2      2.2032828      1.1016414     3.03     0.0524
NHIJOS                   3     51.9685403     17.3228468    47.65     0.0001
PROF*NHIJOS              6     25.5630541      4.2605090    11.72     0.0001

          Duncan's Multiple Range Test for variable: FACTOR1
       NOTE: This test controls the type I comparisonwise error rate, not
             the experimentwise error rate
                  Alpha= 0.05  df= 108  MSE= 0.054006
                      Number of Means      2      3
                      Critical Range   .1030  .1084
       Means with the same letter are not significantly different.

             Duncan Grouping              Mean       N  PROF
                         A           1.25518      40  dire
                         B          -0.44973      40  ofic
                         C          -0.80546      40  trab

          Duncan's Multiple Range Test for variable: FACTOR2

       NOTE: This test controls the type I comparisonwise error rate, not
             the experimentwise error rate
                  Alpha= 0.05  df= 108  MSE= 0.363566
                      Number of Means      2      3
                      Critical Range   .2673  .2813
       Means with the same letter are not significantly different.

             Duncan Grouping              Mean       N  PROF
                         A           0.1217      40  trab
                    B    A           0.0673      40  ofic
                    B               -0.1890      40  dire

          Duncan's Multiple Range Test for variable: FACTOR1
       NOTE: This test controls the type I comparisonwise error rate, not
             the experimentwise error rate
                  Alpha= 0.05  df= 108  MSE= 0.054006
                   Number of Means      2      3      4
                   Critical Range   .1189  .1252  .1293
       Means with the same letter are not significantly different.

            Duncan Grouping              Mean       N  NHIJOS
                         A           0.43908      30  5
                         B           0.11165      30  4
                         C          -0.17351      30  3
                         D          -0.37722      30  2
```

```
        Duncan's Multiple Range Test for variable: FACTOR2
     NOTE: This test controls the type I comparisonwise error rate, not
           the experimentwise error rate
                Alpha= 0.05  df= 108  MSE= 0.363566
                 Number of Means      2      3      4
                 Critical Range   .3086  .3248  .3355

          Duncan Grouping              Mean       N  NHIJOS
                        A            0.9704      30  5
                        B            0.1425      30  4
                        C           -0.2814      30  3
                        D           -0.8314      30  2
```

En el archivo de resultados se observa que existe una interacción muy grande entre las dos fuentes de variación, por lo que es más conveniente hacer el estudio para los 12 tratamientos o 12 tipos de familias.

Archivo de programa *SAS* (C18-12.SAS).-

```
title 'ANOVA de las componentes principales';
Options ls=75 ps=60;
Data vino;
Infile 'c18-1.dat';
input  familia $ simfam $ pan legumbre fruta carne pollos leche vino
@@;
proc princomp cov data=vino out=scores noprint;
var pan legumbre fruta carne pollos leche vino;
run;
proc anova data=scores;
class familia;
model prin1-prin2 = familia;
means familia / duncan;
run;
proc factor data=vino cov n=4 rotate=varimax score outstat=fact;
var pan legumbre fruta carne pollos leche vino;
run;
proc score data=vino score=fact out=scores;
var pan legumbre fruta carne pollos leche vino;
run;
proc anova data=scores;
class familia;
model factor1-factor2 = familia;
means familia / duncan;
run;
```

Archivo de resultados (c18-12.LST).-

```
                    Analysis of Variance Procedure
Dependent Variable: PRIN1
                                  Sum of          Mean
Source                     DF        Squares        Square  F Value     Pr > F
Model                      11     31424751.5     2856795.6   143.60     0.0001
Error                     108      2148561.6       19894.1
Corrected Total           119     33573313.1

                   R-Square           C.V.       Root MSE          PRIN1 Mean
                   0.936004        9999.99        141.046             0.00000

Source                     DF       Anova SS   Mean Square  F Value     Pr > F
FAMILIA                    11     31424751.5     2856795.6   143.60     0.0001
```

Dependent Variable: PRIN2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	3076038.24	279639.84	35.74	0.0001
Error	108	845127.07	7825.25		
Corrected Total	119	3921165.31			

R-Square	C.V.	Root MSE	PRIN2 Mean
0.784470	9999.99	88.4604	0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
FAMILIA	11	3076038.24	279639.84	35.74	0.0001

Duncan's Multiple Range Test for variable: PRIN1

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 19894.09

Number of Means	2	3	4	5	6	7
Critical Range	125.0	131.6	135.9	139.1	141.6	143.6

Number of Means	8	9	10	11	12
Critical Range	145.3	146.7	147.9	148.9	149.9

Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	FAMILIA
	A	1013.26	10	d5
	B	645.71	10	d4
	B	601.94	10	d3
	C	216.53	10	o5
D	C	100.40	10	d2
D	E	9.93	10	o4
	E	-113.78	10	t5
	F	-333.10	10	t4
	G	-471.90	10	o3
H	G	-520.71	10	t3
H	G	-539.36	10	o2
H		-608.93	10	t2

Duncan's Multiple Range Test for variable: PRIN2

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 7825.251

Number of Means	2	3	4	5	6	7
Critical Range	78.42	82.53	85.26	87.26	88.81	90.07

Number of Means	8	9	10	11	12
Critical Range	91.11	91.99	92.75	93.41	93.99

Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	FAMILIA
	A	308.28	10	o5
	A	265.96	10	t5
	B	123.80	10	t4
	B	91.27	10	d5
	C	6.18	10	o4
D	C	-29.50	10	t3
D	C	-32.74	10	o3
D	E	-90.11	10	d2
	E	-125.73	10	o2
F	E	-133.54	10	d3
F	E	-172.52	10	t2
F		-211.35	10	d4

Analysis of Variance Procedure

Dependent Variable: FACTOR1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	110.970027	10.088184	135.68	0.0001

Error	108	8.029973	0.074352	
Corrected Total	119	119.000000		

R-Square	C.V.	Root MSE	FACTOR1 Mean
0.932521	9999.99	0.27267	0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
FAMILIA	11	110.970027	10.088184	135.68	0.0001

Dependent Variable: FACTOR2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	104.142021	9.467456	68.82	0.0001
Error	108	14.857979	0.137574		
Corrected Total	119	119.000000			

R-Square	C.V.	Root MSE	FACTOR2 Mean
0.875143	9999.99	0.37091	0.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
FAMILIA	11	104.142021	9.467456	68.82	0.0001

Duncan's Multiple Range Test for variable: FACTOR1

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.074352

Number of Means	2	3	4	5	6	7
Critical Range	.2417	.2544	.2628	.2690	.2738	.2776

Number of Means	8	9	10	11	12
Critical Range	.2808	.2836	.2859	.2879	.2897

Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	FAMILIA
	A	1.5298	10	d5
	A	1.4384	10	d4
	A	1.4148	10	d3
	B	0.5359	10	d2
	C	0.0486	10	o4
	C	-0.0116	10	o5
	D	-0.3705	10	t5
	E	-0.7252	10	t4
F	E	-0.9136	10	o3
F	E	-0.9481	10	t2
F	E	-0.9614	10	t3
F		-1.0371	10	o2

Duncan's Multiple Range Test for variable: FACTOR2

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate

Alpha= 0.05 df= 108 MSE= 0.137574

Number of Means	2	3	4	5	6	7
Critical Range	.3288	.3460	.3575	.3659	.3724	.3776

Number of Means	8	9	10	11	12
Critical Range	.3820	.3857	.3889	.3917	.3941

Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	FAMILIA
	A	1.5046	10	o5
	A	1.3465	10	t5
	A	1.2090	10	d5
	B	0.6468	10	t4
	C	0.2621	10	o4
	D	-0.1573	10	t3
	D	-0.1656	10	o3
	E	-0.8553	10	d3
	E	-0.8954	10	d4
	E	-0.9123	10	t2

```
E       -0.9421     10  o2
E       -1.0408     10  d2
```

En el archivo de resultados, la primera salida es el ANOVA de las dos primeras componentes principales que se corresponden con el factor **cultural** y el factor **socioeconómico**, respectivamente. En el **PROC PRINCOMP** se ha especificado la opción **NOPRINT** para que no se impriman los resultados, ya conocidos, del análisis de componentes principales.

Como es de esperar, el coeficiente de determinación de la primera componente es mayor (94%) que el de la segunda componente (78%). Tanto el factor cultural como el factor socioeconómico están influenciados significativamente por los 12 tipos de familias. Pero cuando se observa el sentido de esta influencia con las salidas del análisis *Duncan* se observa que, para el factor cultural, las primeras familias son las correspondientes al nivel profesional **directivo** del cabeza de familia y dentro de ellas están ordenadas de mayor a menor por el número de hijos, mientras que para el factor socioeconómico las primeras familias son las correspondientes al número elevado de hijos y dentro de ellas ordenadas por el nivel profesional del cabeza de familia pero en un sentido diferente al del primer factor, este es el de **oficinistas**, **trabajadores**, **directivos**. Como tal vez la evidencia no sea determinante conviene rotar las componentes con objeto de clarificar estos resultados.

En la segunda parte del programa se ha utilizado el **PROC FACTOR** con objeto de poder hacer la rotación, en este caso la **VARIMAX**. El procedimiento FACTOR no tiene la opción NOPRINT, por lo que no podemos evitar que salgan los resultados ya conocidos de este análisis.

Si nos vamos directamente a la salida del **PROC ANOVA** de los factores rotados se observa que los doce tipos de familias son significativos tanto para el factor cultural como para el factor socioeconómico, pero cuando se observa el sentido de esta influencia con las salidas del análisis *Duncan* se observa que, para el factor cultural, las cuatro primeras familias son las correspondientes al nivel profesional de **directivo** del cabeza de familia y dentro de ellas están perfectamente ordenadas de **5 hijos** a **2 hijos**, mientras que para el factor socioeconómico las tres primeras familias son las correspondientes a **5 hijos** y los tres últimos son los correspondientes a **2 hijos** y dentro de ellas la última es el nivel profesional de **directivo**, mientras que las dos primeras se alternan según el número de hijos Por lo que la conclusión es que en el factor cultural influyen primeramente en el nivel profesional del cabeza de familia y también secundariamente en el número de hijos mientras que en el factor socioeconómico influye primero el número de hijos y después el nivel profesional del padre, interaccionando estos dos factores.

Componentes Principales de variables cualitativas.-

El análisis de componentes principales y el factorial son métodos indicados en variables de proporción aunque pueden ser utilizados con cierta precaución en variables de intervalo. ¿Y que ocurre si el conjunto de variables medidas (o gran parte de ellas) son nominales u ordinales? En este caso hay que realizar una transformación previa de las variables. Esta transformación puede ser lineal o no lineal y tiene como

objetivo optimizar las propiedades de la matriz de covarianzas o correlaciones de las variables transformadas.

Esta metodología, que el *SAS* presenta como el **PROC PRINQUAL** (componentes principales de variables cualitativas), se puede usar en las siguientes circunstancias

1) Análisis de Componentes Principales Ordinarios generalizado para datos que no son cuantitativos.
2) Realizar análisis de preferencia multidimensional de variables métricas y no métricas (ver más adelante)
3) Preprocesado de datos, transformación de las variables antes de su uso en otros análisis
4) Estima de valores perdidos en datos multivariantes antes de analizarlos
5) Resumir datos cualitativos y cuantitativos y detectar relaciones no lineales

El procedimiento **PRINQUAL** provee tres de métodos para transformar un conjunto de variables cualitativas y cuantitativas para optimizar la matriz de covarianzas o la de correlación de las variables transformadas. Estos métodos son

MTV o método de la varianza total máxima

MGV o método la varianza generalizada mínima

MAC o método de la correlación promedio máxima.

Estos tres métodos intentan encontrar la transformación que disminuya el rango de la matriz de covarianzas calculada a partir de las variables transformadas. La transformación de las variables para maximizar la varianza contabilizada para unas combinaciones lineales (usando el método MTV) localiza las observaciones en un espacio de dimensión aproxima al del número constatable de combinaciones lineales tanto como es posible dadas las limitaciones de la transformación. La transformación de los variables para minimizar su varianza generalizada o maximizar la suma de las correlaciones también reduce la dimensionalidad. Las transformaciones de las variables cualitativas (nominal y ordinal) pueden serlo pensando en un análisis de variables cuantitativas. Los datos se cuantifican para que la proporción de la varianza acumulada por un número determinado de componentes principales sea máxima, a varianza generalizada de las variables es localmente mínima, o el promedio de las correlaciones es localmente máxima.

Los datos pueden contener variables cuya escala de medida sea nominal, ordinal, de intervalo, y de proporción. Cualquier mezcla se permite con todos los métodos:

1) Las variables nominales pueden transformarse puntuando las categorías (**transform opscore**) para optimizar la matriz de covarianzas.
2) Las variables ordinales pueden transformarse monotónicamente (**transform monotone**) puntuando las categorías ordenadas para que el orden se conserve débilmente (las categorías adyacentes pueden combinarse) y se optimice la matriz de

covarianzas. Las variables ordinales pueden también transformarse a rangos (**transform rank**).

Ejemplo.-

En un estudio sobre los linfocitos en humanos se han hecho 20 pruebas con anticuerpos a una muestra de 50 individuos. Se asume que la membrana celular de los linfocitos muestreados tiene algunos de los antígenos que pueden ser reconocidos por los anticuerpos. El resultado de la prueba será negativo (1), positivo (2) o reacción desconocida (0)

Archivo de programa *SAS* (C18-13.SAS).-

```
title 'Componentes principales de datos cualitativos';
Options ls=75 ps=30;
data comprin;
infile 'c18-13.dat';
input indi $ p1-p20 @@;
proc princomp out=prefix;
run;
proc plot;
plot prin2*prin1=indi / vspace=3 hspace=5;
run;
proc prinqual data=comprin out=results method=mgv n=3 replace;
transform opscore(p1-p20);
id indi;
run;
proc princomp data=results out=prefix;
var p1-p20;
run;
proc plot;
plot prin2*prin1=indi / vspace=3 hspace=5;
run;
proc factor data=results n=3 preplot rotate=varimax plot;
var p1-p20;
run;
```

Archivo de datos (C18-13.DAT).-

```
A 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 0 1 2 1
B 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
C 2 2 1 1 1 1 1 1 2 1 2 1 2 2 2 1 2 1 1
D 1 1 1 1 1 2 1 1 2 1 2 1 1 1 1 1 2 1 2 1
E 2 0 2 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1
F 1 1 1 1 1 1 1 2 2 2 2 1 1 2 1 1 1 2 1 1
G 2 0 2 1 2 0 1 1 0 1 1 2 1 1 2 0 1 2 0 1
H 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1
I 1 1 1 1 1 1 1 2 2 2 2 1 1 2 1 1 1 1 1 1
J 2 2 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1
K 2 1 2 1 2 1 1 2 2 2 2 1 1 2 1 1 1 2 1 1
L 1 1 0 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1
M 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
N 1 1 1 1 1 1 1 1 2 2 2 2 1 1 2 1 1 1 1 1
O 1 1 1 1 1 1 1 1 1 2 1 2 1 2 2 2 1 1 1 1
P 2 2 1 2 1 1 2 1 1 1 1 2 1 0 2 2 1 2 1 1
Q 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1
R 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 2 1 1 1 1
S 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1
T 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 1 1 2
U 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1
V 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 2 1 1
W 1 1 1 2 1 1 2 2 2 1 2 1 1 1 1 1 2 1 1 1
X 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Y 1 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 2
Z 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
a 2 2 1 1 1 1 1 1 1 1 1 2 1 1 2 2 1 2 1 1
b 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
c 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
d 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
e 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
f 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
g 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
h 1 2 2 1 2 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1
i 2 2 1 1 1 1 1 1 1 2 1 2 1 2 2 2 1 2 1 1
j 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2
k 1 1 2 1 2 1 1 2 2 2 2 1 1 2 1 1 1 1 1 1
l 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
m 2 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1
n 1 1 1 1 1 1 1 2 2 2 2 2 1 2 2 2 1 1 1 1
o 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1 2 1 2 1
p 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2
q 1 1 1 1 1 1 1 1 1 2 1 2 1 0 1 1 1 2 1 1
r 1 1 1 1 1 1 1 1 2 2 2 2 2 1 2 2 1 1 1 1
s 2 1 2 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 1
t 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 1 1 1 1
u 1 1 1 1 1 1 0 1 2 2 2 2 1 1 2 1 1 1 1 1
v 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
w 2 2 2 1 2 1 1 1 1 1 1 2 1 1 2 2 1 2 1 1
x 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1
```

Archivo de resultados (C18-13.LST).-

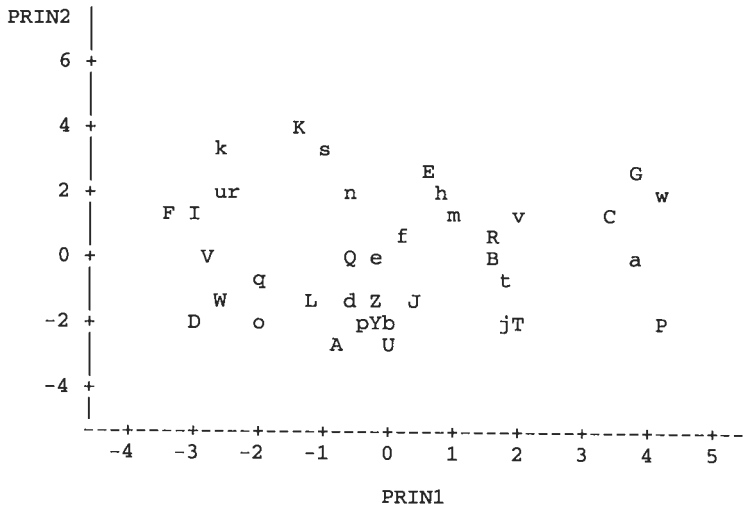
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.01692	0.859175	0.200846	0.20085
PRIN2	3.15774	0.584823	0.157887	0.35873
PRIN3	2.57292	0.393772	0.128646	0.48738
PRIN4	2.17915	0.329734	0.108957	0.59634
PRIN5	1.84941	0.348518	0.092471	0.68881
PRIN6	1.50089	0.347579	0.075045	0.76385
PRIN7	1.15332	0.182231	0.057666	0.82152
PRIN8	0.97108	0.204542	0.048554	0.87007
PRIN9	0.76654	0.145878	0.038327	0.90840
PRIN10	0.62066	0.151192	0.031033	0.93943
PRIN11	0.46947	0.237286	0.023474	0.96291
PRIN12	0.23219	0.067141	0.011609	0.97452
PRIN13	0.16505	0.035851	0.008252	0.98277
PRIN14	0.12920	0.037400	0.006460	0.98923
PRIN15	0.09180	0.016055	0.004590	0.99382
PRIN16	0.07574	0.033147	0.003787	0.99760
PRIN17	0.04259	0.037268	0.002130	0.99973
PRIN18	0.00533	0.005326	0.000266	1.00000
PRIN19	0.00000	0.000000	0.000000	1.00000
PRIN20	0.00000	.	0.000000	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
P1	0.296267	0.245463	-.073468	0.049007	0.310799	-.041910	-.159082
P2	0.220368	0.002845	0.138285	-.056185	0.431602	-.044103	-.324791
P3	0.090525	0.348365	-.324537	0.050087	0.149608	-.038631	0.404816
P4	0.027412	-.190288	0.075744	0.539139	0.189582	0.309633	0.081972
P5	0.082662	0.357919	-.338011	0.033723	0.125771	-.025727	0.396652
P6	-.111337	-.239090	0.096101	-.327404	0.412483	0.085046	0.318979
P7	0.027412	-.190288	0.075744	0.539139	0.189582	0.309633	0.081972
P8	-.302456	0.264447	0.206302	0.231317	0.006033	-.123099	-.081009
P9	-.363729	0.184374	0.222666	0.132070	0.196314	-.213300	0.006466
P10	-.121766	0.317966	0.281049	-.092617	-.052318	0.348895	0.066585
P11	-.346916	0.241656	0.169283	0.189790	0.120007	-.247682	0.009407
P12	0.362689	0.054839	0.330215	0.039146	-.128162	-.203559	0.162042
P13	0.001558	0.076951	0.138017	-.095256	-.225758	0.443070	0.128657
P14	-.143091	0.310543	0.259744	-.178587	-.143578	0.299451	0.059122
P15	0.362689	0.054839	0.330215	0.039146	-.128162	-.203559	0.162042
P16	0.247858	-.008382	0.455807	-.063161	0.108338	-.127777	0.087689
P17	-.172427	-.138548	0.025027	0.097633	-.021129	-.370298	0.243443
P18	0.260949	0.289586	-.001553	0.043298	0.222325	0.139907	-.014578
P19	-.171813	-.218455	0.134797	-.325373	0.405079	0.078154	0.251179
P20	0.058635	-.195050	0.065191	0.146345	-.225943	-.058806	0.467826

Plot of PRIN2*PRIN1. Symbol is value of INDI.



NOTE: 11 obs hidden.

PRINQUAL MGV Iteration History

Iteration Number	Average Change	Maximum Change	Mean Squared Multiple R	R Square Change
1	0.05693	2.78880	0.81451	.
2	0.02168	2.09375	0.85955	0.04504
3	0.01372	0.98857	0.87501	0.01546
4	0.00820	0.53085	0.87991	0.00490
5	0.00429	0.30919	0.88116	0.00125
6	0.00645	0.58536	0.87542	-.00574

NOTE: Algorithm converged.

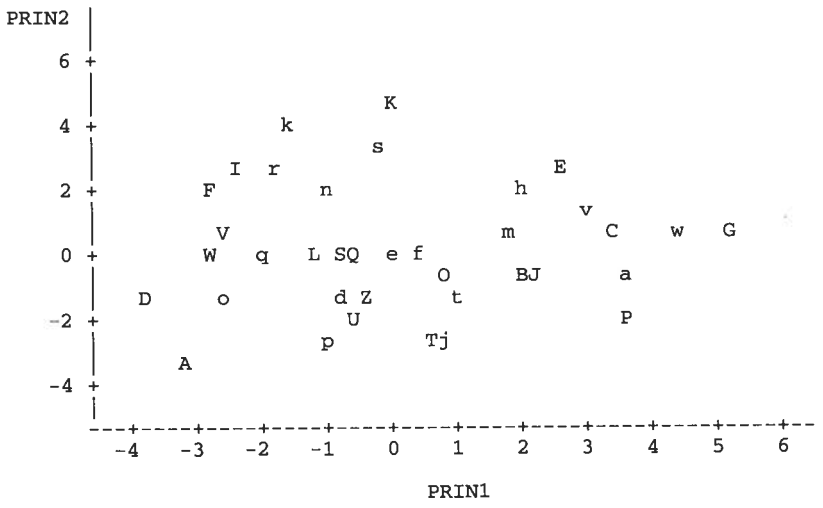
Componentes principales de datos cualitativos
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.47374	1.16534	0.223687	0.22369
PRIN2	3.30840	0.65304	0.165420	0.38911
PRIN3	2.65536	0.49112	0.132768	0.52188
PRIN4	2.16425	0.33790	0.108212	0.63009
PRIN5	1.82634	0.37285	0.091317	0.72140
PRIN6	1.45349	0.19518	0.072675	0.79408
PRIN7	1.25831	0.38555	0.062916	0.85700
PRIN8	0.87276	0.19521	0.043638	0.90063
PRIN9	0.67756	0.14728	0.033878	0.93451
PRIN10	0.53027	0.16017	0.026514	0.96102
PRIN11	0.37011	0.17261	0.018505	0.97953
PRIN12	0.19749	0.09977	0.009875	0.98940
PRIN13	0.09772	0.02778	0.004886	0.99429
PRIN14	0.06994	0.03716	0.003497	0.99779
PRIN15	0.03277	0.02130	0.001639	0.99943
PRIN16	0.01147	0.01145	0.000573	1.00000
PRIN17	0.00002	0.00002	0.000001	1.00000
PRIN18	0.00000	0.00000	0.000000	1.00000
PRIN19	0.00000	0.00000	0.000000	1.00000
PRIN20	0.00000	.	0.000000	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
P1	0.360051	0.151965	-.074696	0.018896	0.307423	-.005942	-.117078
P2	0.390173	0.069341	-.068855	0.038468	0.285632	0.040343	-.196677
P3	0.179426	0.285019	-.310532	-.024528	-.007709	0.068015	0.477919
P4	-.008366	-.137134	0.153735	-.526694	0.284742	0.268371	0.087523
P5	0.179424	0.285019	-.310534	-.024525	-.007715	0.068024	0.477916
P6	-.183426	-.160517	-.122715	0.348427	0.374325	0.201927	0.163433
P7	-.008366	-.137134	0.153735	-.526694	0.284742	0.268371	0.087523
P8	-.223135	0.355782	0.164462	-.144515	0.096631	-.212985	0.051934
P9	-.292268	0.320663	0.118747	-.065401	0.224776	-.254625	0.040185
P10	-.057425	0.354288	0.253253	0.151375	-.050879	0.394231	-.088914
P11	-.264637	0.348044	0.087335	-.094310	0.204674	-.291427	0.077098
P12	0.281802	-.084665	0.376736	0.129722	0.030926	-.219538	0.258534
P13	-.018974	0.038478	0.172555	0.064268	-.254396	0.392155	0.296805
P14	-.047224	0.338150	0.290618	0.169279	-.067522	0.374668	-.076204
P15	0.281802	-.084665	0.376736	0.129722	0.030926	-.219538	0.258534
P16	0.106516	-.077438	0.429026	0.196444	0.162261	-.079038	0.064796
P17	-.232203	-.192765	-.135237	0.162890	0.278607	-.089724	0.245442
P18	0.355519	0.169254	-.063525	0.024622	0.292984	0.042487	-.170641
P19	-.241520	-.152206	-.020364	0.338473	0.373033	0.214005	0.059295
P20	-.003490	-.212271	0.129155	-.130537	-.130518	-.035525	0.325506

Plot of PRIN2*PRIN1. Symbol is value of INDI.



NOTE: 13 obs hidden.

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 20 Average = 1

	1	2	3	4	5
Eigenvalue	4.4737	3.3084	2.6554	2.1642	1.8263
Difference	1.1653	0.6530	0.4911	0.3379	0.3728
Proportion	0.2237	0.1654	0.1328	0.1082	0.0913
Cumulative	0.2237	0.3891	0.5219	0.6301	0.7214

3 factors will be retained by the NFACTOR criterion.

Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
P1	0.76155	0.27641	-0.12172
P2	0.82526	0.12613	-0.11220
P3	0.37951	0.51842	-0.50602
P4	-0.01769	-0.24943	0.25052
P5	0.37950	0.51842	-0.50602
P6	-0.38797	-0.29196	-0.19997
P7	-0.01769	-0.24943	0.25052
P8	-0.47196	0.64713	0.26800
P9	-0.61818	0.58325	0.19350
P10	-0.12146	0.64441	0.41268
P11	-0.55974	0.63306	0.14232
P12	0.59605	-0.15400	0.61390
P13	-0.04013	0.06999	0.28118
P14	-0.09988	0.61506	0.47357
P15	0.59605	-0.15400	0.61390
P16	0.22530	-0.14085	0.69911
P17	-0.49114	-0.35062	-0.22037
P18	0.75197	0.30786	-0.10352
P19	-0.51084	-0.27685	-0.03318
P20	-0.00738	-0.38610	0.21046

Componentes principales de datos cualitativos
Initial Factor Method: Principal Components

Variance explained by each factor

FACTOR1	FACTOR2	FACTOR3
4.473740	3.308403	2.655361

Final Communality Estimates: Total = 10.437505

P1	P2	P3	P4	P5	P6	P7
0.671179	0.709556	0.668842	0.125289	0.668844	0.275749	0.125289

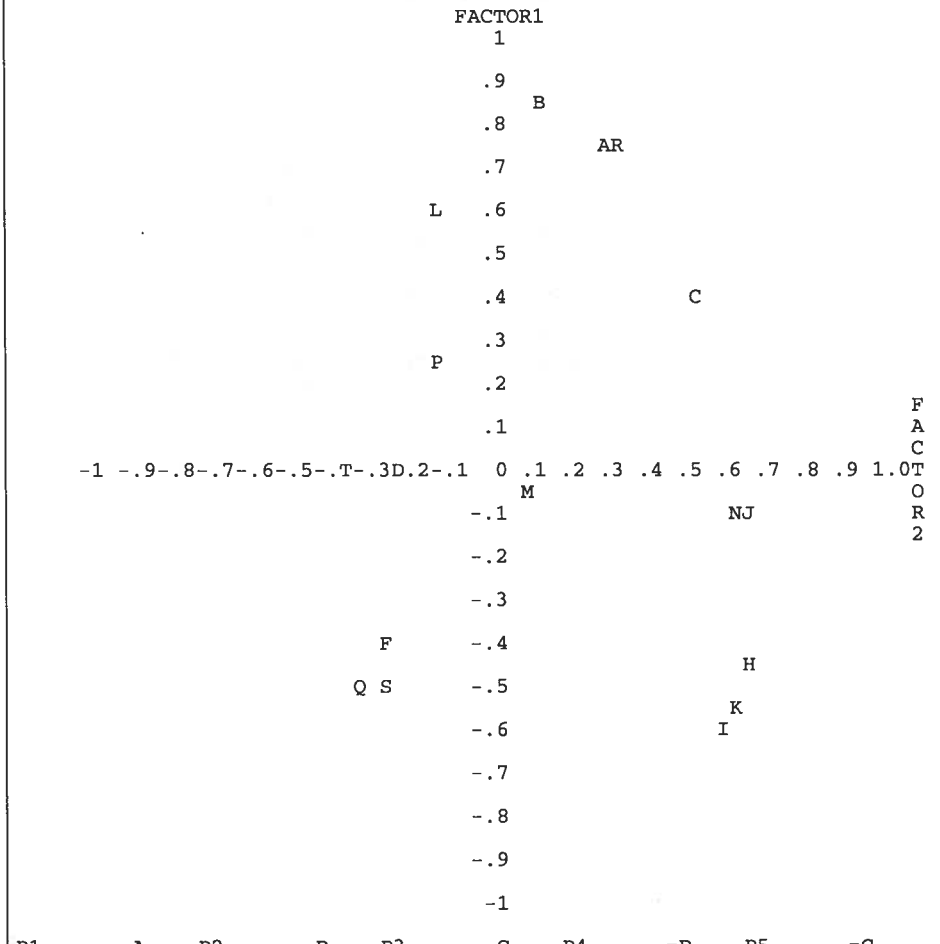
P8	P9	P10	P11	P12	P13	P14
0.713346	0.759778	0.600330	0.734323	0.755861	0.085573	0.612545

P15	P16	P17	P18	P19	P20
0.755861	0.559352	0.412714	0.670945	0.338708	0.193421

Componentes principales de datos cualitativos

Initial Factor Method: Principal Components

Plot of Factor Pattern for FACTOR1 and FACTOR2



P1	=A	P2	=B	P3	=C	P4	=D	P5	=C
P6	=F	P7	=D	P8	=H	P9	=I	P10	=J
P11	=K	P12	=L	P13	=M	P14	=N	P15	=L
P16	=P	P17	=Q	P18	=R	P19	=S	P20	=T

Componentes principales de datos cualitativos

Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	0.91010	-0.40967	0.06234
2	0.37980	0.76450	-0.52086
3	0.16572	0.49771	0.85136

Rotation Method: Varimax

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
P1	0.77790	-0.16125	-0.20013
P2	0.78038	-0.29750	-0.10977
P3	0.45843	-0.01099	-0.67718
P4	-0.06932	-0.05876	0.34210
P5	0.45842	-0.01099	-0.67718
P6	-0.49712	-0.16380	-0.04235
P7	-0.06932	-0.05876	0.34210
P8	-0.13934	0.82146	-0.13833
P9	-0.30902	0.79545	-0.17759
P10	0.20260	0.74781	0.00812
P11	-0.24540	0.78411	-0.24347
P12	0.58571	-0.05636	0.64002
P13	0.03665	0.20990	0.20043
P14	0.22118	0.74683	0.07659
P15	0.58571	-0.05636	0.64002
P16	0.26741	0.14798	0.68260
P17	-0.61667	-0.17653	-0.03561
P18	0.78414	-0.12422	-0.20161
P19	-0.57557	-0.01889	0.08410
P20	-0.11848	-0.18740	0.37982

Rotation Method: Varimax

Variance explained by each factor

FACTOR1	FACTOR2	FACTOR3
4.255703	3.342205	2.839596

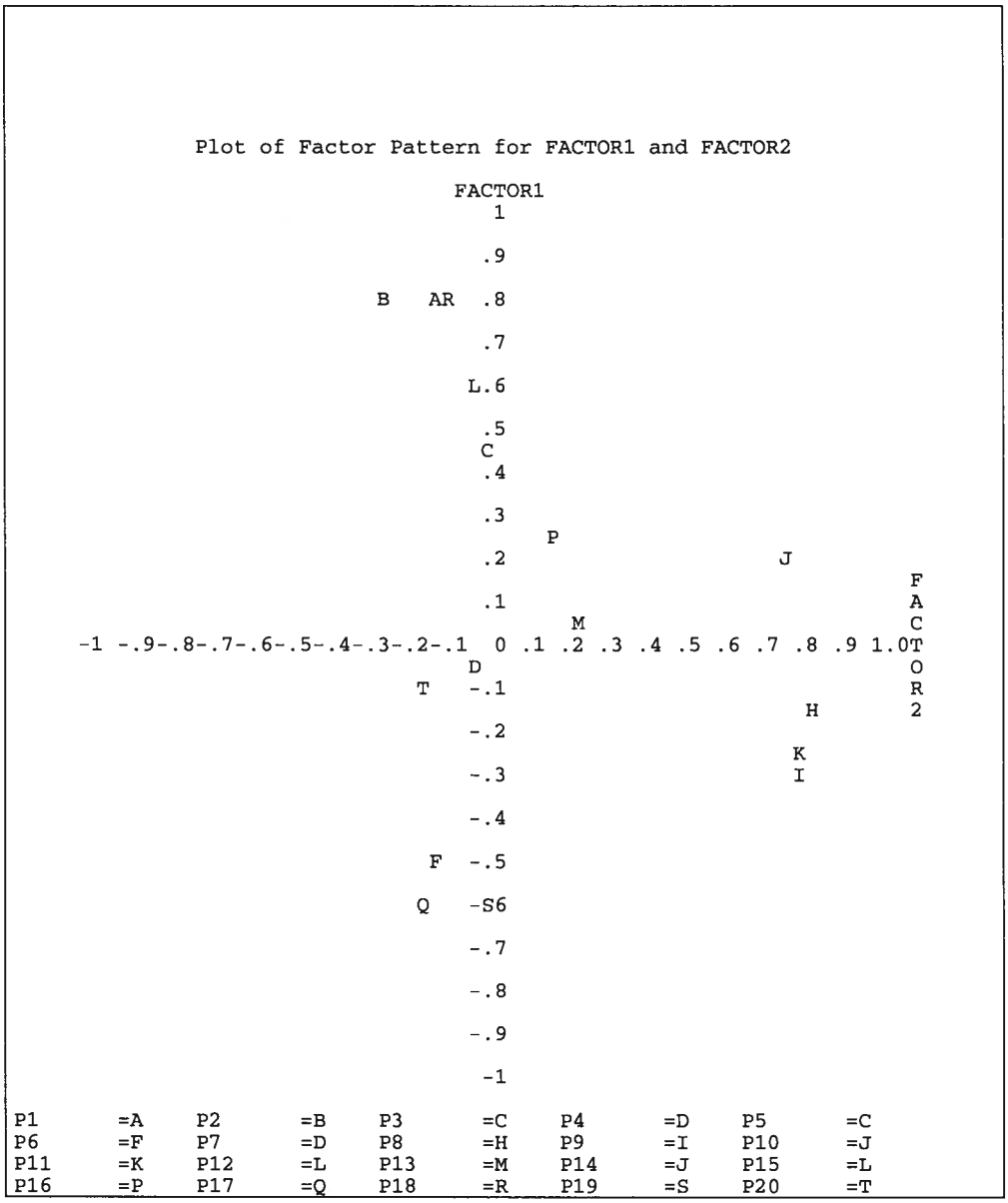
Final Communality Estimates: Total = 10.437505

P1	P2	P3	P4	P5	P6	P7
0.671179	0.709556	0.668842	0.125289	0.668844	0.275749	0.125289

P8	P9	P10	P11	P12	P13	P14
0.713346	0.759778	0.600330	0.734323	0.755861	0.085573	0.612545

P15	P16	P17	P18	P19	P20
0.755861	0.559352	0.412714	0.670945	0.338708	0.193421

Componentes principales de datos cualitativos
Rotation Method: Varimax



El primer procedimiento es el de componentes principales (**princomp**), se ha hecho el análisis de componentes principales con los datos originales, el segundo procedimiento es el de componentes principales para datos cualitativos (**prinqual**), al ser datos nominales, se han transformado con **opscore** y por último se han analizado los datos transformados con el procedimiento **princomp** y el procedimiento **factor** ya visto en los epígrafes anteriores.

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

CAPÍTULO 19

Análisis discriminante y Cluster

CAPÍTULO 19

Análisis discriminante y Cluster

Análisis Discriminante.-

El análisis discriminante es una técnica de clasificación y asignación de un individuo o grupo de individuos a uno de varios grupos en base a sus características que habitualmente serán una o más variables cuantitativas. Si se asume que la distribución de las variables dentro de cada grupo es normal se usan métodos paramétricos. Si no se puede realizar dicho supuesto se pueden usar métodos no paramétricos.

El análisis discriminante puede plantearse con dos objetivos:

a) Objetivo descriptivo o explicatorio cuando se realiza la representación de un conjunto de observaciones con objeto de verificar si pertenecen a grupos diferentes. O bien se pretende encontrar la variable o variables que mejor discriminan a grupos preestablecidos.

b) Objetivo de predecir o tomar decisiones sobre la pertenencia de un individuo o grupo de individuos a uno de los grupos definidos a priori.

Para el primer objetivo se tiene el *ANÁLISIS FACTORIAL DISCRIMINANTE* o, más comúnmente, *ANÁLISIS CANÓNICO DISCRIMINANTE*, y para el segundo objetivo se tiene la *FUNCIÓN DISCRIMINANTE*.

El análisis factorial/canónico discriminante es un método factorial porque obtiene factores o variables sintéticas que permiten discriminar los grupos. Para ello en primer lugar se determina si los grupos quedan perfectamente discriminados en función de las variables originales disponibles, para después analizar cuáles son las variables que contribuyen más a discriminar entre los grupos que se han formado. Para ello lo que se hace es reducir las variables que mejor discriminan a otras nuevas variables denominadas *variables canónicas*. Generalmente una sola variable canónica es la que

aporta más información. Las variables canónicas son combinación lineal de las variables originales y vienen expresadas por medio de una *función discriminante lineal.*

Por lo que el análisis factorial/canónico discriminante es una técnica de reducción de dimensiones relacionada con el análisis de correlaciones canónicas y el análisis de componentes principales. Dada una variable de clasificación y varias variables cuantitativas, este análisis halla una variables canónicas (que son combinaciones lineales de las variables cuantitativas originales) que recoge la variación entre clases de la misma manera que los componentes principales recogían la variación total.

La función discriminante es una ecuación lineal con una variable dependiente que representa la pertenencia de un individuo a un grupo, por lo que la función discriminante pone énfasis en la predicción. Las funciones discriminantes son combinaciones lineales de las variables, interveniendo cada una de ellas con una ponderación diferente, de forma que esas ponderaciones indican cuales son las variables mejor discriminantes.

Si se tienen dos grupos, la función discriminante no es más que una ecuación de regresión múltiple en la que la variable dependiente es una variable nominal con espacio muestral (0, 1), que representa la pertenencia a uno u otro grupo.

En general, cuando se dispone de *k* grupos, se pueden calcular *k*-1 funciones discriminantes incorrelacionadas.

Ejemplo ilustrativo.-

Para ilustrar las ideas que se van ha exponer pongamos un ejemplo simple. El análisis discriminante tiene su origen en le trabajo, ya clásico, de *Fisher* que trata de la clasificación de tres especies del género *Iris* en base a cuatro medidas: longitud y anchura del sépalo y del pétalo. Para hacerlo más simple supóngase que se han medido solo 10 individuos de dos de estas especies siendo los datos

Versicolor				*Virginica*			
l.s.	*a.s.*	*l.p.*	*a.p.*	*l.s.*	*a.s.*	*l.p.*	*a.p.*
6.5	2.8	4.6	1.5	6.4	2.8	5.6	2.2
6.2	2.2	4.5	1.5	6.7	3.1	5.6	2.4
5.9	3.2	4.8	1.8	6.3	2.8	5.1	1.5
6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
6.0	2.7	5.1	1.6	6.5	3.0	5.2	2.0
5.6	2.5	3.9	1.1	6.5	3.0	5.5	1.8
5.7	2.8	4.5	1.3	5.8	2.7	5.1	1.9
6.3	3.3	4.7	1.6	6.2	3.4	5.4	2.3
7.0	3.2	4.7	1.4	6.8	3.2	5.9	2.3
6.4	3.2	4.5	1.5	6.7	3.3	5.7	2.5

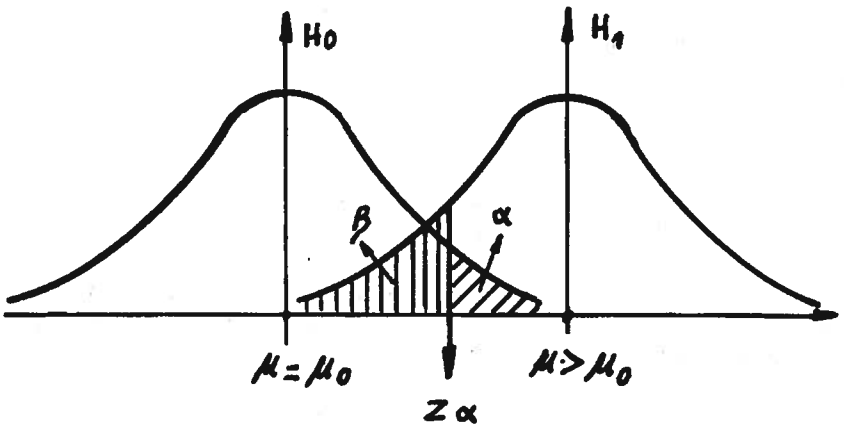
Los estadísticos descriptivos son

	Versicolor(n=10)		*Virginica*(n=10)	
	$\overline{X}$	S	$\overline{X}$	S
LONGSEP	6.170	0.4111	6.480	0.3259
ANCHSEP	2.890	0.3573	3.040	0.2270
LONGPET	4.590	0.3035	5.420	0.2860
ANCHPET	1.470	0.1888	2.120	0.3120

Supóngase que se quiere saber si un individuo es de la especie versicolor o de la especie virginica. Mirando la tabla anterior se observa que la especie versicolor tiene los sépalos menos anchos, por lo que, intuitivamente, se pueden clasificar como versicolor los individuos que tengan un valor bajo en la anchura del sépalo. Pero si se usa la información las cuatro variables conjuntamente, la clasificación será más efectiva que si se usa solo una variable.

Conceptos básico de clasificación.-

Supóngase que se tiene un individuo que puede pertenecer a una de estas dos poblaciones. Comencemos considerando como un individuo puede ser clasificado en una de estas poblaciones en base a la medida de un solo carácter, que podemos denominar *X*. Supóngase que se tiene una muestra representativa de cada una de las poblaciones que nos posibilita estimar las distribuciones de *X* y sus medias. Estas distribuciones pueden representarse de la siguiente manera



Viendo esta figura es intuitivamente obvio que un valor bajo de *X* puede propiciar que se clasifique al individuo como de la población versicolor y un valor alto de la variable *X* propicia que clasifiquemos al individuo en la población virginica. Para definir que significa valor *bajo* o *alto*, debemos seleccionar un punto de separación. Si simbolizamos este punto de separación como *C*, entonces se clasificará un individuo como de la población virginica si $X \geq C$. Para cualquier valor de *C* se puede incurrir en cierto porcentaje de error. Si un individuo procede de la población virginica pero la medida de *X* es menor que *C*, se clasifica incorrectamente como un individuo de la población versicolor, y viceversa. Estos dos tipos de error son los ilustrados en la figura

anterior. Si se asume que las dos poblaciones tienen la misma varianza, entonces el valor de C es

$$C = \frac{\overline{X}_{versi} + \overline{X}_{virgi}}{2}$$

Este valor asegura que las dos probabilidades de error son iguales.

La situación ilustrada en la figura anterior es ideal y raramente se encuentra en la practica. En situaciones reales el grado de solapamiento de las dos distribuciones es frecuentemente mayor y la varianza raramente es exactamente igual. Por ejemplo, con los datos de la anchura del sépalo de las dos especies del género *Iris* se tiene que el punto de separación C es

$$C = \frac{2.89+3.04}{2} = 2.965$$

Como se puede comprobar el porcentaje de error cometido es bastante alto, pues hay cinco (50%) individuos de la especie versicolor que tienen una anchura del sépalo superior a este punto, por lo que serían clasificados, incorrectamente, como de la especie virginica, y hay tres (30%) individuos de la especie virginica con una anchura del sépalo inferior al punto de separación por lo que serían clasificados, incorrectamente, como versicolor. Por lo que de los 20 individuos estudiados se han clasificado correctamente 5+7=12 individuos, esto es, el 60% de los individuos, por lo que esta variable no es util para identificar cual individuos es versicolor o virginica.

Si se combina dos o más variables se puede conseguir una mejor clasificación. El número de variables usadas debe ser menor que N_I más N_{II} menos uno. Si se tiene en cuenta la variable longitud del sépalo, se tiene que para esta variable, ignorando las demás

$$C_2 = \frac{6.17+6.48}{2} = 6.325$$

se clasifican correctamente 14 individuos, esto es, el 70%. Si se juntan los resultados de las dos C se clasifican correctamente 10 individuos, el 50%.

Para usar ambas variables simultáneamente *Fisher* representó la línea de separación, $Z=C$ como una ecuación donde Z es una combinación lineal de X_1 y X_2 y C es una constante definida como

$$C = \frac{\overline{Z}_{versi} + \overline{Z}_{virgi}}{2}$$

siendo $\overline{Z}_{versi}$ la media de los valores de Z en la población virginica y $\overline{Z}_{virgi}$ la media de los valores de Z de la población versicolor. Y siendo Z la *función discriminante de Fisher*, definida como

$$Z = a_1 X_1 + a_2 X_2$$

para el caso de dos variables. Más adelante se verá la forma en que se pueden calcular estos dos coeficientes, si bien no es preciso su cálculo como se verá, también,

dentro de un momento.

Una vez que se tiene esta función discriminante, se puede calcular el valor de Z para todos los individuos pudiéndose representar los dos poblaciones a modo de histograma o distribución de frecuencias, con lo que se tendría que se ha reducido un problema de clasificación bivariante (X_1 y X_2) a un problema de clasificación univariante (Z).

Con los datos del ejemplo se tienen que la función discriminante es

$$Z = 2.05526 \times longsep + 0.55899 \times anchsep$$

La media de Z para cada especie sería

$$\overline{Z}_i = 2.05526 \times \overline{longsep}_i + 0.55899 \times \overline{anchsep}_i$$

por lo que la media de Z para la especie versicolor es

$$\overline{Z}_{versicolor} = 2.05526 \times 6.17 + 0.55899 \times 2.89 = 14.2964$$

y para la especie virginica es

$$\overline{Z}_{virginca} = 2.05526 \times 6.48 + 0.55899 \times 3.04 = 15.0174$$

La media de estos dos puntos es

$$C = \frac{14.2964 + 15.0174}{2} = 14.6529$$

Por lo que un individuos se clasificará como versicolor si su Z es menor que este valor y se clasificará como virgínica si su Z es mayor que este valor.
Si calculamos las Z para los diez individuos se tiene

Versicolor			*Virginica*		
l.s.	*a.s.*	*Z*	*l.s.*	*a.s.*	*Z*
6.5	2.8	14.9244	6.4	2.8	14.7188
6.2	2.2	13.9724	6.7	3.1	15.5031
5.9	3.2	13.9148	6.3	2.8	14.5133
6.1	3.0	14.2141	6.9	3.1	15.9142
6.0	2.7	13.8408	6.5	3.0	15.0362
5.6	2.5	12.9069	6.5	3.0	15.0362
5.7	2.8	13.2802	5.8	2.7	13.4298
6.3	3.3	14.7928	6.2	3.4	15.7645
7.0	3.2	16.1756	6.8	3.2	14.6432
6.4	3.2	14.9424	6.7	3.3	15.6149

lo que se han clasificado correctamente seis individuos versicolor y siete individuos virginica, en total se ha clasificado correctamente el 65% de los individuos (recuérdese

que este es un ejemplo con pocos individuos y pocas variables).

Interpretación de los coeficientes.-

Además de su uso para clasificación, la función discriminante de *Fisher* es útil para indicar la dirección y grado en la que cada variable contribuye a la clasificación. Lo primero a examinar es el signo de cada coeficiente: si es positivo, los individuos con valores elevados de la correspondiente variable tienden a ser de la población de media mayor, y viceversa. En el ejemplo, ambos coeficientes son positivos, indicando que altos valores de ambas variables están asociados con la especie virginica. En ejemplos mas completos, comparar las variables que tienen coeficientes positivos con la variables que los tiene negativos puede ser revelador. Para cuantificar la magnitud de la distribución, los coeficientes estandarizados puedes ser de mucha utilidad, como se explicará más adelante,

Base teórica.-

En la deducción de la función discriminante no se asumió una distribución para las variables. *Fisher* define la función discriminante como

$$Z = a_1 X_1 + a_2 X_2 + ... + a_p X_p$$

Se puede simbolizar los dos valores medios de Z como $\overline{Z}_1$ y $\overline{Z}_2$. También se puede simbolizar la varianza conjunta de Z como S^2_Z

$$S^2_Z = \frac{(n_1 - 1) S^2_{Z_1} + (n_2 - 1) S^2_{Z_2}}{(n_1 - 1) + (n_2 - 1)}$$

Para medir el grado de separación de dos grupos de individuos en términos de valores Z, se calcula

$$D^2 = \frac{(\overline{Z}_1 - \overline{Z}_2)^2}{S^2_Z}$$

Fisher selecciona los coeficientes a_1, a_2,.., a_p de manera que D^2 tenga el máximo valor posible.

Para el ejemplo que se está desarrollando se tiene

	$\overline{Z}$	S^2_Z
Versicolor	14.2964	0.8909
Virgincia	15.0174	0.5511

siendo

$$S^2_Z = \frac{(9)0.8909 + (9)0.5511}{18} = 0.7210$$

y

$$D^2 = \frac{(14.2964 - 15.0174)^2}{0.7210} = 0.7210$$

El término D^2 puede interpretarse como el cuadrado de las distancias entre las medias de los valores estandarizados de Z. Un valor grande de D^2 indica que es fácil discriminar entre los dos grupos. La cantidad D^2, que no es sino una distancia euclídea, es conocida como *distancia de Mahalanobis*. Tanto los coeficientes, a_i, como la distancia de *Mahalanobis*, D^2, son función de las medias, de las varianzas conjuntas y covarianzas de las variables (ver más adelante en los análisis *CLUSTER*).

Si se hacen algunos supuestos de la distribución de las variables es posible desarrollar algunos procedimientos estadísticos relacionados con problemas de clasificación. Estos procedimientos incluyen pruebas de hipótesis para la más o menos utilidad de las variables y métodos para estimar los errores de clasificación.

Si definimos las variables utilizadas para la clasificación como X_1, X_2,..., X_p, el modelo estándar realiza el supuesto de que estas variables tienen una distribución normal multivariante en cada una de las poblaciones. A veces se asume que la matriz de covarianzas es la misma en las diferentes poblaciones. Pero el valor medio de una variable determinada puede ser diferente en las diferentes poblaciones. Se asume también que se tiene una muestra aleatoria de cada una de las poblaciones.

Alternativamente, se puede pensar que las diferentes poblaciones son subpoblaciones de una población.

Analogía con la regresión.-

Existe una conexión útil entre la regresión y el análisis discriminante. Para la regresión hay que tomar las variables de clasificación X_1, X_2,..., X_p, como variables independientes mientras que la variable dependiente será una variable de diseño que indicará la población de la que procede cada observación. En el caso de dos poblaciones puede ser

$$Y = \frac{N_1}{N_1 + N_2}$$

si la observación es de la población 1, y

$$Y = -\frac{N_1}{N_1 + N_2}$$

si la observación es de la población 2 Por ejemplo para los datos que se están desarrollando Y=0.5 si el individuo es versicolor e Y=-0.5 si el individuo es virginica.

Si se realiza un análisis de regresión múltiple, los coeficientes de regresión resultantes serán proporcionales a los coeficientes de la función discriminante, a_1, a_2,...,a_p, y el valor del coeficiente de determinación múltiple esta relacionado con el valor de D^2 de *Mahalanobis* de la siguiente manera

$$D^2 = \frac{R^2}{1-R^2} \frac{[N_1+N_2][N_1+N_2-2]}{N_1 \times N_2}$$

Por lo que realizando la regresión múltiple es posible obtener los coeficientes de la función discriminante y el valor de D^2. Los valores de $\overline{Z}$ de cada grupo puede obtenerse multiplicando cada coeficiente por la media de la correspondiente variable. El punto de separación, C, se puede calcular igual que antes.

Para el ejemplo que se está desarrollando, los coeficientes de regresión múltiple son

$$b_{LONGSEP} = 0.475648$$
$$b_{ANCHSEP} = 0.129366$$

Efectivamente

$$\frac{0.4756}{0.1293} = \frac{2.0553}{0.5590}$$

teniendo en cuenta los errores de redondeo.

El valor del coeficiente de determinación de este análisis es R^2=0.1669, lo que da

$$D^2 = \frac{0.1669^2}{1-0.1669} \frac{[10+10][10+10-2]}{10 \times 10} = 0.7212$$

que es el mismo resultado anterior.

Los valores medios de Z son

$$\overline{Z}_{versicolor} = 0.4756 \times 6.17 + 0.1294 \times 2.89 = 3.3086$$
$$\overline{Z}_{virginica} = 0.4756 \times 6.48 + 0.1294 \times 3.04 = 3.4754$$

por lo que el valor del punto de separación es

$$C = \frac{3.3086+3.4754}{2} = 3.3920$$

Por lo que un individuos se clasificará como versicolor si su Z (calculada con los coeficientes de regresión múltiple) es menor que este valor y se clasificará como virgínica si su Z es mayor que este valor.

Si calculamos las Z para los diez individuos se tiene

Versicolor			*Virginica*		
l.s.	*a.s.*	*Z*	*l.s.*	*a.s.*	*Z*
6.5	2.8	3.45390	6.4	2.8	3.40634
6.2	2.2	3.23359	6.7	3.1	3.58784
5.9	3.2	3.22026	6.3	2.8	3.35878
6.1	3.0	3.28952	6.9	3.1	3.68297
6.0	2.7	3.20315	6.5	3.0	3.47978
5.6	2.5	2.98702	6.5	3.0	3.47978
5.7	2.8	3.07339	5.8	2.7	3.10802
6.3	3.3	3.42346	6.2	3.4	3.64834
7.0	3.2	3.74347	6.8	3.2	3.38883
6.4	3.2	3.45809	6.7	3.3	3.61372

Lo que da el mismo resultado obtenido anteriormente, este es, se han clasificado correctamente seis individuos versicolor y siete individuos virginica, en total se ha clasificado correctamente el 65% de los individuos (recuerdese que este es un ejemplo con pocos individuos y pocas variables).

Cálculo de la *función discriminante de Fisher*.-

Los paquetes estadísticos ofrecen alguna función relacionada con esta, por ejemplo, el *SAS* ofrece la denominada *función lineal discriminante*, consistente en los coeficientes de la función discriminante pero para cada población, pudiéndose, en ese caso, obtener la función discriminante de *Fisher* restando los coeficientes dados por el paquete estadístico.

Para el caso del ejemplo que se está desarrollando el *SAS* da como parte de su salida del **Proc Discrim** lo siguiente

Discriminant Analysis Linear Discriminant Function

$$\text{Constant} = -.5\ \bar{X}_j'\ COV^{-1}\ \bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1}\ \bar{X}_j$$

	VARIEDAD	
	Versicol	virginic
CONSTANT	-141.80445	-156.46138
LONGSEP	41.37200	43.42726
ANCHSEP	9.80750	10.36649

Siendo, por tanto, el coeficiente a_1 para la longitud del sépalo: 43.42726-41.372=2.05526. El coeficiente a_2 para la anchura del sépalo: 10.36649-9.8075=0.55899. Y el punto de separación *C* se obtienen haciendo la resta en orden contrario, esta es, -141.80445-(-156.46138)=14.6569. Lo que da los mismos valores que los obtenidos anteriormente. Para más de dos variables el proceso es el mismo.

Coeficientes tipificados.-

Los valores de los coeficientes de la función discriminante no son directamente comparables, de la misma manera que no lo son los coeficientes de regresión. Pero puede obtenerse una impresión del efecto relativo de cada variable en la función discriminante si se obtienen los coeficientes discriminantes tipificados. Esta técnica implica el uso de la matriz de covarianzas conjunta. Para el ejemplo que se está desarrollando se tiene que

	longsep	*anchsep*
Versicolor	0.1690	0.1277
Virgincia	0.1062	0.0515

siendo

$$S^2_{longsep} = \frac{(9)0.1690+(9)0.1062}{18} = 0.1376$$

y

$$S^2_{anchsep} = \frac{(9)0.1277+(9)0.0515}{18} = 0.0896$$

Por tanto, la desviación típica conjunta es 0.3709 para la longitud del sépalo y 0.2993 para la anchura del sépalo. Los coeficientes típicos se obtienen multiplicando los coeficientes por su correspondiente desviación típica, por lo que los coeficientes discriminantes típicos son: 2.0553*0.3708=0.7624 para la longitud del sépalo y 0.5589*0.2993=0.1673 para la anchura del sépalo. Con estos coeficientes se observa que la longitud del sépalo tiene un mayor efecto en la función discriminante que la anchura del sépalo.

Probabilidades a posteriori.-

Una vez asignado un individuo a un grupo o a otro en el proceso de clasificación, siempre existe la posibilidad de haber hecho una clasificación errónea, por lo que se puede calcular la probabilidad de que un individuo pertenezca a un grupo o a otro. Esta probabilidad se puede calcular bajo el supuesto de distribución normal multivariante. La probabilidad de que un individuo pertenezca a la población 1 es

$$P_{(1)} = \frac{1}{1+e^{(-Z+C)}}$$

Y la probabilidad de pertenecer a la población 2 es $1-P_{(1)}$.

Por ejemplo, supóngase que un individuo cuyo sépalo mide de longitud 5.6 y de anchura 2.7; la función discriminante de este individuo es

$$Z = 2.05526 \times 5.6 + 0.55894 \times 2.7 = 13.0186$$

como C=14.6569 y el valor de Z es menor, este individuo se clasifica como perteneciente a la población 1, es decir, como versicolor. Para determinar la probabilidad de que este individuo sea virginica, se tiene

$$P_{(I)} = \frac{1}{1+e^{(-13.0186+14.6569)}} = 0.1627$$

y la probabilidad de que sea versicolor es 1-0.1627=0.8373, por lo que es mucho más probable que sea versicolor a que sea virginica.

El *SAS* estima estas probabilidades, denominándolas *Posterior Probability of Membership in ...*, como se verá en el ejemplo que se desarrollará mas adelante.

Estas probabilidades posteriores sirven para interpretar los resultados de la clasificación. En investigador puede aceptar la clasificación de los individuos cuya probabilidad favorece claramente la pertenencia a un grupo y juzgar si es conveniente el eliminar los individuos cuya probabilidad de pertenencia a uno de los grupos esta próxima al 0.5. Más adelante se verá la probabilidad a priori que puede posibilitar la modificación del punto de separación.

Hay que notar que la función discriminante presentada aquí es una estima muestral de la función discriminante poblacional. Más adelante se verá si estos valores son buena estimas de los parámetros. Si ambas poblaciones son normales y tienen la misma matriz de covarianzas entonces puede considerarse como óptima la clasificación de la función discriminante.

Incorporando las probabilidades a priori en la elección del punto *C*.-

El punto C se ha usado como el punto de separación que produce el mismo porcentaje de error en ambos tipos, esto es, la probabilidad de clasificar erróneamente un individuo de la población 1 en la población 2 o viceversa. Pero la elección del valor C puede realizarse de manera que produzca una proporción determinada de estas probabilidades de error. Para explicar como puede realizarse esta elección hay que introducir el concepto de *probabilidad a priori*. Puesto que las dos poblaciones constituyen una población general, es interesante examinar su tamaño relativo. La probabilidad a priori de la población 1 es la probabilidad de que un individuo seleccionado al azar proceda efectivamente de la población 1. En otras palabras, es la proporción de individuos de la población general que caen el la población 1. Esta proporción se denota como q_I.

Como en el ejemplo que se está desarrollando existe el mismo número versicolor que de virginica, la probabilidad a priori de ambas es 0.5. Si no se conoce ninguna característica de un individuo dado, se puede clasificar como versicolor con una probabilidad del 50%. En este caso puede ser correcto el 50% de las veces. Este es un ejemplo intuitivo de la interpretación de la probabilidad a priori.

La elección teórica del punto de separación C se realiza de manera que sea mínima la probabilidad total de clasificaciones erróneas. Esta probabilidad total se

define como q_I por la probabilidad de clasificar erróneamente un individuo de la población 2 en la población 1 más q_{II} por la probabilidad de clasificar erróneamente un individuo de la población 1 en la población 2, o

$$q_1 \times P_{(2\ siendo\ 1)} + q_2 \times P_{(1\ siendo\ 2)}$$

Bajo el modelo normal multivariante la elección del punto C es

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln \frac{q_2}{q_1}$$

nótese que si $q_1=q_2=1/2$, entonces $q_2/q_1=1$ y $\ln(q_2/q_1)=0$. En este caso C es

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$$

Por lo que en los epígrafes previos se ha asumido que $q_1=q_2=1/2$.

Supóngase que en el ejemplo que se esta desarrollando la muestra hubiera sido de 40 versicolor (se repiten 10 veces los valores de versicolor de la tabla de datos) y 10 virginica, dando, por lo demás, exactamente todos los resultados obtenidos hasta el momento. En ese caso se tendría que $q_1=0.8$ y $q_2=0.2$. por lo tanto el punto de separación sería

$$C = 14.6569 + \ln(0.25) = 14.6569 + 1.2840 = 15.9409$$

Observando los datos de la tabla con los valores Z se tiene

Versicolor			*Virginica*		
l.s.	*a.s.*	*Z*	*l.s.*	*a.s.*	*Z*
6.5	2.8	14.9244	6.4	2.8	14.7188
6.2	2.2	13.9724	6.7	3.1	15.5031
5.9	3.2	13.9148	6.3	2.8	14.5133
6.1	3.0	14.2141	6.9	3.1	15.9142
6.0	2.7	13.8408	6.5	3.0	15.0362
5.6	2.5	12.9069	6.5	3.0	15.0362
5.7	2.8	13.2802	5.8	2.7	13.4298
6.3	3.3	14.7928	6.2	3.4	15.7645
7.0	3.2	16.1756	6.8	3.2	14.6432
6.4	3.2	14.9424	6.7	3.3	15.6149

lo que se han clasificado correctamente 36 individuos versicolor (9x4), esto es, el 90% de los individuos, mientras que de los virginicas no se han clasificado correctamente ninguno. Por lo tanto la probabilidad de clasificar un individuo versicolor como virginica es próxima a cero (0.1), pero la probabilidad de clasificar un individuo virginica como versicolor es 1, siendo la probabilidad total de clasificación errónea: 0.8x0.1+0.2+1=0.28. Si se utiliza el valor original de C=14.6569, la probabilidad de clasificar erróneamente es de 0.4 y 0.3, respectivamente, por lo que la probabilidad total de clasificación errónea es 0.8x0.4+0.2x0.3=0.38. Este resultado verifica que el

punto de separación teórico produce una pequeña disminución en la probabilidad total de clasificación errónea.

Por lo que en la práctica se debería de elegir varios valores de *C* y para cada uno de ellos determinar la dos probabilidades de clasificación errónea y elegir el valor de *C* que equilibre ambos valores.

Bondad de la función discriminante.-

Una medida de la bondad del procedimiento de clasificación consiste en las dos probabilidades de clasificación errónea: probabilidad de clasificarlo en 2 siendo del 1 y probabilidad de clasificarlo en 1 siendo del 2. Existen varios métodos para estimar estas probabilidades. Uno es el denominado *método empírico*, que es el que se ha visto anteriormente y que se han denominado probabilidades a posteriori. Estas probabilidades es una manera de validar la función discriminante. Aunque este método es intuitivo produce estimas sesgadas, de hecho subestima la verdadera probabilidad de clasificación errónea porque se utiliza la misma muestra para determinar y validar la función discriminante.

Idealmente se debería utilizar una muestra para estimar la función discriminante y aplicar esta función discriminante en otra muestra para estimar la proporción de clasificaciones erróneas. Este procedimiento se denomina de *validación cruzada*, y produce estimas insesgadas. Se puede realizar validación cruzada dividiendo la muestra aleatoriamente en dos submuestras una para derivar la función discriminante y la otra para validar esta función.

Si la muestra es pequeña es difícil poder dividirla para la validación cruzada. A veces se usa un método alternativo que imita la división de la muestra, se denomina *procedimiento navaja*, consistente en excluir una observación del primer grupo, calcular la función discriminante con el resto de las observaciones y clasificar con esta la observación del primer grupo excluida. Este procedimiento se repite para todas las observaciones del primer grupo. La proporción de individuos clasificados erroneamente por el método de la navaja es *Prob*(2 siendo de 1). Un procedimiento similar se usa para estimar la *Prob*(1 siendo de 2). Este método produce estimaciones próximamente insesgadas. Los paquetes estadísticos traen algunas de estas validaciones, por ejemplo el *SAS* tiene la **CROSSVALIDATE** que es un tipo de validación cruzada consistente en clasificar cada observación con la función discriminante calculada con todas las demás observaciones.

Si se realiza la valoración cruzada del *SAS* de los datos del ejemplo, da como resultado que se clasifican erróneamente dos individuos más que los que daban con el método empírico, esto es, se clasifican erróneamente además de los individuos anteriormente detectados con la función discriminante calculada con todos ellos, el primer individuos virginica y el segundo individuo versicolor.

Prueba de la contribución de las variables de clasificación.-

Se puede plantear si la clasificación de los individuos realizada usando las variables disponibles es mejor que la realizada aleatoriamente. Esta cuestión es un problema de prueba de hipótesis. La hipótesis nula es que ninguna de las variables mejora la clasificación basada en el azar. Una hipótesis nula equivalente es que las medias de las variables son las mismas en las diferentes poblaciones, o que la D^2 poblacional es cero. Por lo que la prueba estadística para esta hipótesis nula es la T^2 de *Hotelling* en el caso de dos poblaciones o el análisis multivariante de la varianza para el caso de más de dos poblaciones.

Ejemplo.-

Analicemos los datos completos de *Fisher* que trata de la clasificación de tres especies del género *Iris*.

Archivo de programa *SAS* (C19-1.SAS).-

```
title 'Análisis discriminante';
options ls =80 ps=60;
data ad;
infile 'c19-1.dat';
input longsep anchsep longpet anchpet variedad $ simb $ @@;
proc discrim distance list crosslist;
class variedad;
var longsep anchsep longpet anchpet;
run;
```

Archivo de datos C19-1.DAT.-

```
5.0 3.3 1.4 0.2 setosa     s  4.8 3.0 1.4 0.3 setosa     s  4.7 3.2 1.3 0.2 setosa     s
6.4 2.8 5.6 2.2 virginica  a  5.1 3.8 1.6 0.2 setosa     s  4.6 3.1 1.5 0.2 setosa     s
6.5 2.8 4.6 1.5 versicolor v  6.1 3.0 4.9 1.8 virginica  a  6.9 3.2 5.7 2.3 virginica  a
6.7 3.1 5.6 2.4 virginica  a  4.8 3.4 1.9 0.2 setosa     s  6.2 2.9 4.3 1.3 versicolor v
6.3 2.8 5.1 1.5 virginica  a  5.0 3.0 1.6 0.2 setosa     s  7.4 2.8 6.1 1.9 virginica  a
4.6 3.4 1.4 0.3 setosa     s  5.0 3.2 1.2 0.2 setosa     s  5.9 3.0 4.2 1.5 versicolor v
6.9 3.1 5.1 2.3 virginica  a  6.1 2.6 5.6 1.4 virginica  a  5.1 3.4 1.5 0.2 setosa     s
6.2 2.2 4.5 1.5 versicolor v  6.4 2.8 5.6 2.1 virginica  a  5.0 3.5 1.3 0.3 setosa     s
5.9 3.2 4.8 1.8 versicolor v  4.3 3.0 1.1 0.1 setosa     s  5.6 2.8 4.9 2.0 virginica  a
4.6 3.6 1.0 0.2 setosa     s  5.8 4.0 1.2 0.2 setosa     s  6.0 2.2 4.0 1.0 versicolor v
6.1 3.0 4.6 1.4 versicolor v  5.1 3.8 1.9 0.4 setosa     s  7.3 2.9 6.3 1.8 virginica  a
6.0 2.7 5.1 1.6 versicolor v  6.7 3.1 4.4 1.4 versicolor v  6.7 2.5 5.8 1.8 virginica  a
6.5 3.0 5.2 2.0 virginica  a  6.2 2.8 4.8 1.8 virginica  a  4.9 3.1 1.5 0.1 setosa     s
5.6 2.5 3.9 1.1 versicolor v  4.9 3.0 1.4 0.2 setosa     s  6.7 3.1 4.7 1.5 versicolor v
6.5 3.0 5.5 1.8 virginica  a  5.1 3.5 1.4 0.2 setosa     s  6.3 2.3 4.4 1.3 versicolor v
5.8 2.7 5.1 1.9 virginica  a  5.6 3.0 4.5 1.5 versicolor v  5.4 3.7 1.5 0.2 setosa     s
6.8 3.2 5.9 2.3 virginica  a  5.8 2.7 4.1 1.0 versicolor v  5.6 3.0 4.1 1.3 versicolor v
5.1 3.3 1.7 0.5 setosa     s  5.0 3.4 1.6 0.4 setosa     s  6.3 2.5 4.9 1.5 versicolor v
5.7 2.8 4.5 1.3 versicolor v  4.6 3.2 1.4 0.2 setosa     s  6.1 2.8 4.7 1.2 versicolor v
6.2 3.4 5.4 2.3 virginica  a  6.0 2.9 4.5 1.5 versicolor v  6.4 2.9 4.3 1.3 versicolor v
7.7 3.8 6.7 2.2 virginica  a  5.7 2.6 3.5 1.0 versicolor v  5.1 2.5 3.0 1.1 versicolor v
6.3 3.3 4.7 1.6 versicolor v  5.7 4.4 1.5 0.4 setosa     s  5.7 2.8 4.1 1.3 versicolor v
6.7 3.3 5.7 2.5 virginica  a  5.0 3.6 1.4 0.2 setosa     s  6.5 3.0 5.8 2.2 virginica  a
7.6 3.0 6.6 2.1 virginica  a  7.7 3.0 6.1 2.3 virginica  a  6.9 3.1 5.4 2.1 virginica  a
4.9 2.5 4.5 1.7 virginica  a  6.3 3.4 5.6 2.4 virginica  a  5.4 3.9 1.3 0.4 setosa     s
5.5 3.5 1.3 0.2 setosa     s  5.8 2.7 5.1 1.9 virginica  a  5.1 3.5 1.4 0.3 setosa     s
6.7 3.0 5.2 2.3 virginica  a  5.7 2.9 4.2 1.3 versicolor v  7.2 3.6 6.1 2.5 virginica  a
7.0 3.2 4.7 1.4 versicolor v  7.2 3.0 5.8 1.6 virginica  a  6.5 3.2 5.1 2.0 virginica  a
6.4 3.2 4.5 1.5 versicolor v  5.4 3.4 1.5 0.4 setosa     s  6.1 2.9 4.7 1.4 versicolor v
6.1 2.8 4.0 1.3 versicolor v  5.2 4.1 1.5 0.1 setosa     s  5.6 2.9 3.6 1.3 versicolor v
4.8 3.1 1.6 0.2 setosa     s  7.1 3.0 5.9 2.1 virginica  a  6.9 3.1 4.9 1.5 versicolor v
5.9 3.0 5.1 1.8 virginica  a  6.4 3.1 5.5 1.8 virginica  a  6.4 2.7 5.3 1.9 virginica  a
5.5 2.4 3.8 1.1 versicolor v  6.0 3.0 4.8 1.8 virginica  a  6.8 3.0 5.5 2.1 virginica  a
6.3 2.5 5.0 1.9 virginica  a  6.3 2.9 5.6 1.8 virginica  a  5.5 2.5 4.0 1.3 versicolor v
6.4 3.2 5.3 2.3 virginica  a  4.9 2.4 3.3 1.0 versicolor v  4.8 3.4 1.6 0.2 setosa     s
5.2 3.4 1.4 0.2 setosa     s  5.6 2.7 4.2 1.3 versicolor v  4.8 3.0 1.4 0.1 setosa     s
4.9 3.6 1.4 0.1 setosa     s  5.7 3.0 4.2 1.2 versicolor v  4.5 2.3 1.3 0.3 setosa     s
5.4 3.0 4.5 1.5 versicolor v  5.5 4.2 1.4 0.2 setosa     s  5.7 2.5 5.0 2.0 virginica  a
7.9 3.8 6.4 2.0 virginica  a  4.9 3.1 1.5 0.2 setosa     s  5.7 3.8 1.7 0.3 setosa     s
4.4 3.2 1.3 0.2 setosa     s  7.7 2.6 6.9 2.3 virginica  a  5.1 3.8 1.5 0.3 setosa     s
6.7 3.3 5.7 2.1 virginica  a  6.0 2.2 5.0 1.5 virginica  a  5.5 2.3 4.0 1.3 versicolor v
```

5.0	3.5	1.6	0.6	setosa	s	5.4	3.9	1.7	0.4	setosa	s	6.6	3.0	4.4	1.4	versicolor	v
5.8	2.6	4.0	1.2	versicolor	v	6.6	2.9	4.6	1.3	versicolor	v	6.8	2.8	4.8	1.4	versicolor	v
4.4	3.0	1.3	0.2	setosa	s	5.2	2.7	3.9	1.4	versicolor	v	5.4	3.4	1.7	0.2	setosa	s
7.7	2.8	6.7	2.0	virginica	a	6.0	3.4	4.5	1.6	versicolor	v	5.1	3.7	1.5	0.4	setosa	s
6.3	2.7	4.9	1.8	virginica	a	5.0	3.4	1.5	0.2	setosa	s	5.2	3.5	1.5	0.2	setosa	s
4.7	3.2	1.6	0.2	setosa	s	4.4	2.9	1.4	0.2	setosa	s	5.8	2.8	5.1	2.4	virginica	a
5.5	2.6	4.4	1.2	versicolor	v	5.0	2.0	3.5	1.0	versicolor	v	6.7	3.0	5.0	1.7	versicolor	v
5.0	2.3	3.3	1.0	versicolor	v	5.5	2.4	3.7	1.0	versicolor	v	6.3	3.3	6.0	2.5	virginica	a
7.2	3.2	6.0	1.8	virginica	a	5.8	2.7	3.9	1.2	versicolor	v	5.3	3.7	1.5	0.2	setosa	s

Archivo de resultados (C19-1.LST).-

```
                          Discriminant Analysis
               150 Observations           149 DF Total
                 4 Variables              147 DF Within Classes
                 3 Classes                  2 DF Between Classes

                         Class Level Information
                                                                    Prior
      VARIEDAD      Frequency        Weight      Proportion     Probability
      setosa               50       50.0000        0.333333        0.333333
      versicol             50       50.0000        0.333333        0.333333
      virginic             50       50.0000        0.333333        0.333333

         Discriminant Analysis       Pooled Covariance Matrix Information
                Covariance         Natural Log of the Determinant
                Matrix Rank        of the Covariance Matrix
                     4                     -9.9585388

      Discriminant Analysis       Pairwise Squared Distances Between Groups
                            2                         -1
                           D (i|j) = (X - X )' COV   (X - X )
                                       i   j           i   j

                         Squared Distance to VARIEDAD
             From
             VARIEDAD          setosa        versicol        virginic
             setosa                 0        89.86419       179.38471
             versicol        89.86419               0        17.20107
             virginic       179.38471        17.20107               0

                         F Statistics, NDF=4, DDF=144 for
                         Squared Distance to VARIEDAD
             From
             VARIEDAD          setosa        versicol        virginic
             setosa                 0       550.18889            1098
             versicol       550.18889               0       105.31265
             virginic            1098       105.31265               0

                         Prob > Mahalanobis Distance for
                         Squared Distance to VARIEDAD
             From
             VARIEDAD          setosa        versicol        virginic
             setosa            1.0000          0.0001          0.0001
             versicol          0.0001          1.0000          0.0001
             virginic          0.0001          0.0001          1.0000
Discriminant Analysis       Pairwise Generalized Squared Distances Between Groups
                            2                         -1
                           D (i|j) = (X - X )' COV   (X - X )
                                       i   j           i   j
                         Generalized Squared Distance to VARIEDAD
             From
             VARIEDAD          setosa        versicol        virginic
             setosa                 0        89.86419       179.38471
             versicol        89.86419               0        17.20107
             virginic       179.38471        17.20107               0

             Discriminant Analysis      Linear Discriminant Function
                                 -1                                 -1
           Constant = -.5 X' COV   X       Coefficient Vector = COV   X
                           j         j                                  j
                                    VARIEDAD
                          setosa        versicol        virginic
             CONSTANT    -85.20986       -71.75400      -103.26971
             LONGSEP      23.54417        15.69821        12.44585
             ANCHSEP      23.58787         7.07251         3.68528
             LONGPET     -16.43064         5.21145        12.76654
             ANCHPET     -17.39841         6.43423        21.07911
```

			Posterior Probability of Membership in VARIEDAD:		
Obs	From VARIEDAD	Classified into VARIEDAD	setosa	versicol	virginic
1	setosa	setosa	1.0000	0.0000	0.0000
2	setosa	setosa	1.0000	0.0000	0.0000
3	setosa	setosa	1.0000	0.0000	0.0000
4	virginic	virginic	0.0000	0.0000	1.0000
5	setosa	setosa	1.0000	0.0000	0.0000
6	setosa	setosa	1.0000	0.0000	0.0000
7	versicol	versicol	0.0000	0.9956	0.0044
8	virginic	virginic	0.0000	0.1342	0.8658
9	virginic	virginic	0.0000	0.0000	1.0000
10	virginic	virginic	0.0000	0.0000	1.0000
11	setosa	setosa	1.0000	0.0000	0.0000
12	versicol	versicol	0.0000	1.0000	0.0000
13	virginic	versicol *	0.0000	0.7294	0.2706
14	setosa	setosa	1.0000	0.0000	0.0000
15	virginic	virginic	0.0000	0.0001	0.9999
16	setosa	setosa	1.0000	0.0000	0.0000
17	setosa	setosa	1.0000	0.0000	0.0000
18	versicol	versicol	0.0000	0.9992	0.0008
19	virginic	virginic	0.0000	0.0004	0.9996
20	virginic	virginic	0.0000	0.0660	0.9340
21	setosa	setosa	1.0000	0.0000	0.0000
22	versicol	versicol	0.0000	0.9596	0.0404
23	virginic	virginic	0.0000	0.0000	1.0000
24	setosa	setosa	1.0000	0.0000	0.0000
25	versicol	virginic *	0.0000	0.2532	0.7468
26	setosa	setosa	1.0000	0.0000	0.0000
27	virginic	virginic	0.0000	0.0008	0.9992
28	setosa	setosa	1.0000	0.0000	0.0000
29	setosa	setosa	1.0000	0.0000	0.0000
30	versicol	versicol	0.0000	1.0000	0.0000
31	versicol	versicol	0.0000	0.9981	0.0019
32	setosa	setosa	1.0000	0.0000	0.0000
33	virginic	virginic	0.0000	0.0001	0.9999
34	versicol	virginic *	0.0000	0.1434	0.8566
35	versicol	versicol	0.0000	1.0000	0.0000
36	virginic	virginic	0.0000	0.0002	0.9998
37	virginic	virginic	0.0000	0.0031	0.9969
38	virginic	virginic	0.0000	0.1884	0.8116
39	setosa	setosa	1.0000	0.0000	0.0000
40	versicol	versicol	0.0000	1.0000	0.0000
41	setosa	setosa	1.0000	0.0000	0.0000
42	versicol	versicol	0.0000	0.9982	0.0018
43	virginic	virginic	0.0000	0.0061	0.9939
44	setosa	setosa	1.0000	0.0000	0.0000
45	versicol	versicol	0.0000	0.9995	0.0005
46	virginic	virginic	0.0000	0.0011	0.9989
47	versicol	versicol	0.0000	0.9806	0.0194
48	setosa	setosa	1.0000	0.0000	0.0000
49	virginic	virginic	0.0000	0.0000	1.0000
50	versicol	versicol	0.0000	1.0000	0.0000
51	versicol	versicol	0.0000	0.9999	0.0001
52	setosa	setosa	1.0000	0.0000	0.0000
53	setosa	setosa	1.0000	0.0000	0.0000
54	versicol	versicol	0.0000	0.8155	0.1845
55	versicol	versicol	0.0000	0.9985	0.0015
56	setosa	setosa	1.0000	0.0000	0.0000
57	versicol	versicol	0.0000	0.9996	0.0004
58	virginic	virginic	0.0000	0.0000	1.0000
59	versicol	versicol	0.0000	0.9925	0.0075
60	versicol	versicol	0.0000	1.0000	0.0000
61	virginic	virginic	0.0000	0.0000	1.0000
62	versicol	versicol	0.0000	1.0000	0.0000
63	versicol	versicol	0.0000	1.0000	0.0000
64	versicol	versicol	0.0000	0.9858	0.0142
65	setosa	setosa	1.0000	0.0000	0.0000
66	versicol	versicol	0.0000	0.9999	0.0001
67	virginic	virginic	0.0000	0.0000	1.0000
68	setosa	setosa	1.0000	0.0000	0.0000
69	virginic	virginic	0.0000	0.0000	1.0000
70	virginic	virginic	0.0000	0.0000	1.0000
71	virginic	virginic	0.0000	0.0000	1.0000
72	virginic	virginic	0.0000	0.0008	0.9992
73	virginic	virginic	0.0000	0.0486	0.9514
74	virginic	virginic	0.0000	0.0000	1.0000
75	setosa	setosa	1.0000	0.0000	0.0000
76	setosa	setosa	1.0000	0.0000	0.0000
77	virginic	virginic	0.0000	0.0011	0.9989
78	setosa	setosa	1.0000	0.0000	0.0000
79	virginic	virginic	0.0000	0.0001	0.9999
80	versicol	versicol	0.0000	0.9999	0.0001
81	virginic	virginic	0.0000	0.0000	1.0000
82	versicol	versicol	0.0000	0.9999	0.0001
83	virginic	virginic	0.0000	0.1037	0.8963
84	virginic	virginic	0.0000	0.0131	0.9869

85	versicol	versicol	0.0000	0.9993	0.0007
86	setosa	setosa	1.0000	0.0000	0.0000
87	versicol	versicol	0.0000	0.9943	0.0057
88	versicol	versicol	0.0000	1.0000	0.0000
89	setosa	setosa	1.0000	0.0000	0.0000
90	versicol	versicol	0.0000	1.0000	0.0000
91	setosa	setosa	1.0000	0.0000	0.0000
92	virginic	virginic	0.0000	0.0000	1.0000
93	versicol	versicol	0.0000	0.9958	0.0042
94	virginic	virginic	0.0000	0.0175	0.9825
95	virginic	virginic	0.0000	0.0062	0.9938
96	virginic	virginic	0.0000	0.0017	0.9983
97	versicol	versicol	0.0000	1.0000	0.0000
98	virginic	virginic	0.0000	0.1925	0.8075
99	virginic	virginic	0.0000	0.0002	0.9998
100	virginic	virginic	0.0000	0.0059	0.9941
101	virginic	virginic	0.0000	0.0011	0.9989
102	versicol	versicol	0.0000	0.9998	0.0002
103	virginic	virginic	0.0000	0.0000	1.0000
104	versicol	versicol	0.0000	1.0000	0.0000
105	setosa	setosa	1.0000	0.0000	0.0000
106	setosa	setosa	1.0000	0.0000	0.0000
107	versicol	versicol	0.0000	0.9997	0.0003
108	setosa	setosa	1.0000	0.0000	0.0000
109	setosa	setosa	1.0000	0.0000	0.0000
110	versicol	versicol	0.0000	1.0000	0.0000
111	setosa	setosa	1.0000	0.0000	0.0000
112	versicol	versicol	0.0000	0.9636	0.0364
113	setosa	setosa	1.0000	0.0000	0.0000
114	virginic	virginic	0.0000	0.0002	0.9998
115	virginic	virginic	0.0000	0.0005	0.9995
116	setosa	setosa	1.0000	0.0000	0.0000
117	setosa	setosa	1.0000	0.0000	0.0000
118	setosa	setosa	1.0000	0.0000	0.0000
119	virginic	virginic	0.0000	0.0000	1.0000
120	setosa	setosa	1.0000	0.0000	0.0000
121	virginic	virginic	0.0000	0.0001	0.9999
122	virginic	virginic	0.0000	0.2208	0.7792
123	versicol	versicol	0.0000	0.9996	0.0004
124	setosa	setosa	1.0000	0.0000	0.0000
125	setosa	setosa	1.0000	0.0000	0.0000
126	versicol	versicol	0.0000	0.9999	0.0001
127	versicol	versicol	0.0000	1.0000	0.0000
128	versicol	versicol	0.0000	0.9999	0.0001
129	versicol	versicol	0.0000	0.9983	0.0017
130	setosa	setosa	1.0000	0.0000	0.0000
131	versicol	versicol	0.0000	0.9995	0.0005
132	setosa	setosa	1.0000	0.0000	0.0000
133	virginic	virginic	0.0000	0.0000	1.0000
134	versicol	versicol	0.0000	0.9940	0.0060
135	setosa	setosa	1.0000	0.0000	0.0000
136	virginic	virginic	0.0000	0.0971	0.9029
137	setosa	setosa	1.0000	0.0000	0.0000
138	setosa	setosa	1.0000	0.0000	0.0000
139	setosa	setosa	1.0000	0.0000	0.0000
140	setosa	setosa	1.0000	0.0000	0.0000
141	virginic	virginic	0.0000	0.0000	1.0000
142	versicol	versicol	0.0000	0.9994	0.0006
143	versicol	versicol	0.0000	1.0000	0.0000
144	versicol	versicol	0.0000	0.6892	0.3108
145	versicol	versicol	0.0000	1.0000	0.0000
146	versicol	versicol	0.0000	1.0000	0.0000
147	virginic	virginic	0.0000	0.0000	1.0000
148	virginic	virginic	0.0000	0.0027	0.9973
149	versicol	versicol	0.0000	1.0000	0.0000
150	setosa	setosa	1.0000	0.0000	0.0000

* Misclassified observation

Discriminant Analysis Classification Summary for Calibration Data: WORK.AD
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \; COV^{-1} \; (X-\bar{X}_j)$$

Posterior Probability of Membership in each VARIEDAD:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \; / \; SUM_k \exp(-.5\, D_k^2(X))$$

Number of Observations and Percent Classified into VARIEDAD:

From VARIEDAD	setosa	versicol	virginic	Total

setosa	50	0	0	50
	100.00	0.00	0.00	100.00
versicol	0	48	2	50
	0.00	96.00	4.00	100.00
virginic	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
Percent	33.33	32.67	34.00	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for VARIEDAD:

	setosa	versicol	virginic	Total
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

Posterior Probability of Membership in VARIEDAD:

Obs	From VARIEDAD	Classified into VARIEDAD	setosa	versicol	virginic
1	setosa	setosa	1.0000	0.0000	0.0000
2	setosa	setosa	1.0000	0.0000	0.0000
3	setosa	setosa	1.0000	0.0000	0.0000
4	virginic	virginic	0.0000	0.0000	1.0000
5	setosa	setosa	1.0000	0.0000	0.0000
6	setosa	setosa	1.0000	0.0000	0.0000
7	versicol	versicol	0.0000	0.9951	0.0049
8	virginic	virginic	0.0000	0.1438	0.8562
9	virginic	virginic	0.0000	0.0000	1.0000
10	virginic	virginic	0.0000	0.0000	1.0000
11	setosa	setosa	1.0000	0.0000	0.0000
12	versicol	versicol	0.0000	1.0000	0.0000
13	virginic	versicol *	0.0000	0.7876	0.2124
14	setosa	setosa	1.0000	0.0000	0.0000
15	virginic	virginic	0.0000	0.0002	0.9998
16	setosa	setosa	1.0000	0.0000	0.0000
17	setosa	setosa	1.0000	0.0000	0.0000
18	versicol	versicol	0.0000	0.9992	0.0008
19	virginic	virginic	0.0000	0.0006	0.9994
20	virginic	virginic	0.0000	0.1578	0.8422
21	setosa	setosa	1.0000	0.0000	0.0000
22	versicol	versicol	0.0000	0.9390	0.0610
23	virginic	virginic	0.0000	0.0000	1.0000
24	setosa	setosa	1.0000	0.0000	0.0000
25	versicol	virginic *	0.0000	0.1773	0.8227
26	setosa	setosa	1.0000	0.0000	0.0000
27	virginic	virginic	0.0000	0.0010	0.9990
28	setosa	setosa	1.0000	0.0000	0.0000
29	setosa	setosa	1.0000	0.0000	0.0000
30	versicol	versicol	0.0000	1.0000	0.0000
31	versicol	versicol	0.0000	0.9980	0.0020
32	setosa	setosa	1.0000	0.0000	0.0000
33	virginic	virginic	0.0000	0.0002	0.9998
34	versicol	virginic *	0.0000	0.0992	0.9008
35	versicol	versicol	0.0000	1.0000	0.0000
36	virginic	virginic	0.0000	0.0003	0.9997
37	virginic	virginic	0.0000	0.0033	0.9967
38	virginic	virginic	0.0000	0.2056	0.7944
39	setosa	setosa	1.0000	0.0000	0.0000
40	versicol	versicol	0.0000	1.0000	0.0000
41	setosa	setosa	1.0000	0.0000	0.0000
42	versicol	versicol	0.0000	0.9980	0.0020
43	virginic	virginic	0.0000	0.0066	0.9934
44	setosa	setosa	1.0000	0.0000	0.0000
45	versicol	versicol	0.0000	0.9993	0.0007
46	virginic	virginic	0.0000	0.0012	0.9988
47	versicol	versicol	0.0000	0.9764	0.0236
48	setosa	setosa	1.0000	0.0000	0.0000
49	virginic	virginic	0.0000	0.0000	1.0000
50	versicol	versicol	0.0000	1.0000	0.0000
51	versicol	versicol	0.0000	0.9999	0.0001
52	setosa	setosa	1.0000	0.0000	0.0000
53	setosa	setosa	1.0000	0.0000	0.0000
54	versicol	versicol	0.0000	0.7868	0.2132
55	versicol	versicol	0.0000	0.9983	0.0017
56	setosa	setosa	1.0000	0.0000	0.0000
57	versicol	versicol	0.0000	0.9995	0.0005
58	virginic	virginic	0.0000	0.0000	1.0000
59	versicol	versicol	0.0000	0.9923	0.0077
60	versicol	versicol	0.0000	1.0000	0.0000
61	virginic	virginic	0.0000	0.0000	1.0000
62	versicol	versicol	0.0000	1.0000	0.0000
63	versicol	versicol	0.0000	1.0000	0.0000
64	versicol	versicol	0.0000	0.9839	0.0161
65	setosa	setosa	1.0000	0.0000	0.0000
66	versicol	versicol	0.0000	0.9999	0.0001
67	virginic	virginic	0.0000	0.0000	1.0000

68	setosa	setosa	1.0000	0.0000	0.0000
69	virginic	virginic	0.0000	0.0000	1.0000
70	virginic	virginic	0.0000	0.0000	1.0000
71	virginic	virginic	0.0000	0.0000	1.0000
72	virginic	virginic	0.0000	0.0009	0.9991
73	virginic	virginic	0.0000	0.0879	0.9121
74	virginic	virginic	0.0000	0.0000	1.0000
75	setosa	setosa	1.0000	0.0000	0.0000
76	setosa	setosa	1.0000	0.0000	0.0000
77	virginic	virginic	0.0000	0.0012	0.9988
78	setosa	setosa	1.0000	0.0000	0.0000
79	virginic	virginic	0.0000	0.0001	0.9999
80	versicol	versicol	0.0000	0.9999	0.0001
81	virginic	virginic	0.0000	0.0000	1.0000
82	versicol	versicol	0.0000	0.9999	0.0001
83	virginic	virginic	0.0000	0.1590	0.8410
84	virginic	virginic	0.0000	0.0145	0.9855
85	versicol	versicol	0.0000	0.9992	0.0008
86	setosa	setosa	1.0000	0.0000	0.0000
87	versicol	versicol	0.0000	0.9939	0.0061
88	versicol	versicol	0.0000	1.0000	0.0000
89	setosa	setosa	1.0000	0.0000	0.0000
90	versicol	versicol	0.0000	1.0000	0.0000
91	setosa	setosa	1.0000	0.0000	0.0000
92	virginic	virginic	0.0000	0.0000	1.0000
93	versicol	versicol	0.0000	0.9951	0.0049
94	virginic	virginic	0.0000	0.0206	0.9794
95	virginic	virginic	0.0000	0.0071	0.9929
96	virginic	virginic	0.0000	0.0018	0.9982
97	versicol	versicol	0.0000	1.0000	0.0000
98	virginic	virginic	0.0000	0.2122	0.7878
99	virginic	virginic	0.0000	0.0002	0.9998
100	virginic	virginic	0.0000	0.0071	0.9929
101	virginic	virginic	0.0000	0.0012	0.9988
102	versicol	versicol	0.0000	0.9998	0.0002
103	virginic	virginic	0.0000	0.0000	1.0000
104	versicol	versicol	0.0000	1.0000	0.0000
105	setosa	setosa	1.0000	0.0000	0.0000
106	setosa	setosa	1.0000	0.0000	0.0000
107	versicol	versicol	0.0000	0.9997	0.0003
108	setosa	setosa	1.0000	0.0000	0.0000
109	setosa	setosa	1.0000	0.0000	0.0000
110	versicol	versicol	0.0000	1.0000	0.0000
111	setosa	setosa	1.0000	0.0000	0.0000
112	versicol	versicol	0.0000	0.9475	0.0525
113	setosa	setosa	1.0000	0.0000	0.0000
114	virginic	virginic	0.0000	0.0002	0.9998
115	virginic	virginic	0.0000	0.0008	0.9992
116	setosa	setosa	1.0000	0.0000	0.0000
117	setosa	setosa	1.0000	0.0000	0.0000
118	setosa	setosa	1.0000	0.0000	0.0000
119	virginic	virginic	0.0000	0.0000	1.0000
120	setosa	setosa	1.0000	0.0000	0.0000
121	virginic	virginic	0.0000	0.0001	0.9999
122	virginic	virginic	0.0000	0.3033	0.6967
123	versicol	versicol	0.0000	0.9996	0.0004
124	setosa	setosa	1.0000	0.0000	0.0000
125	setosa	setosa	1.0000	0.0000	0.0000
126	versicol	versicol	0.0000	0.9999	0.0001
127	versicol	versicol	0.0000	1.0000	0.0000
128	versicol	versicol	0.0000	0.9999	0.0001
129	versicol	versicol	0.0000	0.9979	0.0021
130	setosa	setosa	1.0000	0.0000	0.0000
131	versicol	versicol	0.0000	0.9994	0.0006
132	setosa	setosa	1.0000	0.0000	0.0000
133	virginic	virginic	0.0000	0.0000	1.0000
134	versicol	versicol	0.0000	0.9925	0.0075
135	setosa	setosa	1.0000	0.0000	0.0000
136	virginic	virginic	0.0000	0.1070	0.8930
137	setosa	setosa	1.0000	0.0000	0.0000
138	setosa	setosa	1.0000	0.0000	0.0000
139	setosa	setosa	1.0000	0.0000	0.0000
140	setosa	setosa	1.0000	0.0000	0.0000
141	virginic	virginic	0.0000	0.0000	1.0000
142	versicol	versicol	0.0000	0.9993	0.0007
143	versicol	versicol	0.0000	1.0000	0.0000
144	versicol	versicol	0.0000	0.6569	0.3431
145	versicol	versicol	0.0000	1.0000	0.0000
146	versicol	versicol	0.0000	1.0000	0.0000
147	virginic	virginic	0.0000	0.0000	1.0000
148	virginic	virginic	0.0000	0.0034	0.9966
149	versicol	versicol	0.0000	1.0000	0.0000
150	setosa	setosa	1.0000	0.0000	0.0000

* Misclassified observation

Discriminant Analysis Classification Summary for Calibration Data: WORK.AD

Between Canonical Structure

	CAN1	CAN2
LONGSEP	0.991468	0.130348
ANCSEP	-0.825658	0.564171
LONGPET	0.999750	0.022358
ANCPET	0.994044	0.108977

Pooled Within Canonical Structure

	CAN1	CAN2
LONGSEP	0.222596	0.310812
ANCSEP	-0.119012	0.863681
LONGPET	0.706065	0.167701
ANCPET	0.633178	0.737242

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2
LONGSEP	-0.686779533	0.019958173
ANCSEP	-0.668825075	0.943441829
LONGPET	3.885795047	-1.645118866
ANCPET	2.142238715	2.164135931

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2
LONGSEP	-.4269548486	0.0124075316
ANCSEP	-.5212416758	0.7352613085
LONGPET	0.9472572487	-.4010378190
ANCPET	0.5751607719	0.5810398645

Raw Canonical Coefficients

	CAN1	CAN2
LONGSEP	-0.829377642	0.024102149
ANCSEP	-1.534473068	2.164521235
LONGPET	2.201211656	-0.931921210
ANCPET	2.810460309	2.839187853

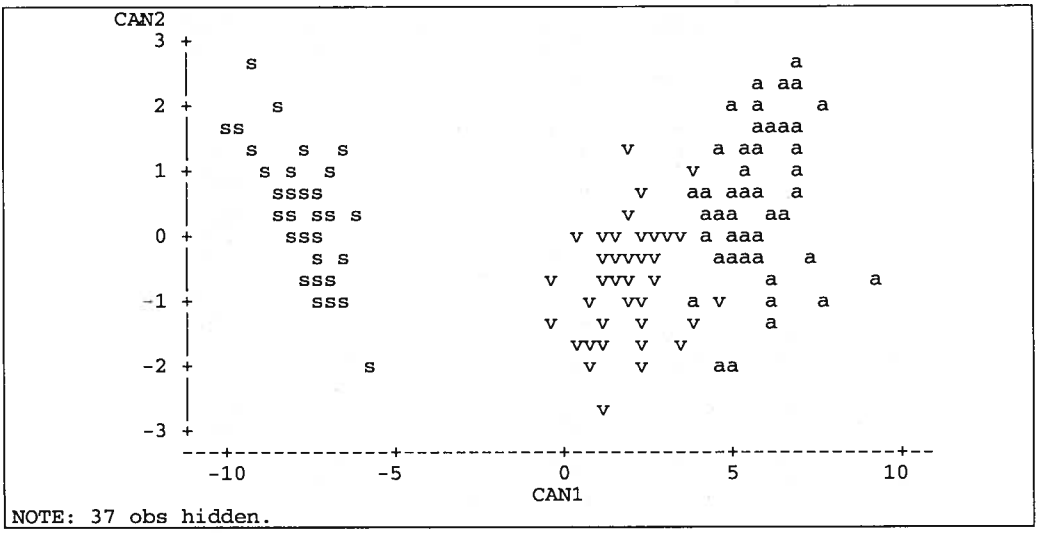
Class Means on Canonical Variables

VARIEDAD	CAN1	CAN2
setosa	-7.607599927	0.215133017
versicol	1.825049490	-0.727899622
virginic	5.782550437	0.512766605

OBS	VARIEDAD	_TYPE_	_NAME_	LONGSEP	ANCSEP	LONGPET	ANCPET
1		N		150.000	150.000	150.000	150.000
2	setosa	N		50.000	50.000	50.000	50.000
3	versicol	N		50.000	50.000	50.000	50.000
4	virginic	N		50.000	50.000	50.000	50.000
5		MEAN		5.843	3.057	3.758	1.199
6	setosa	MEAN		5.006	3.428	1.462	0.246
7	versicol	MEAN		5.936	2.770	4.260	1.326
8	virginic	MEAN		6.588	2.974	5.552	2.026
9	setosa	CSSCP	LONGSEP	6.088	4.862	0.801	0.506
10	setosa	CSSCP	ANCSEP	4.862	7.041	0.573	0.456
11	setosa	CSSCP	LONGPET	0.801	0.573	1.478	0.297
12	setosa	CSSCP	ANCPET	0.506	0.456	0.297	0.544
13	versicol	CSSCP	LONGSEP	13.055	4.174	8.962	2.733
14	versicol	CSSCP	ANCSEP	4.174	4.825	4.050	2.019
15	versicol	CSSCP	LONGPET	8.962	4.050	10.820	3.582
16	versicol	CSSCP	ANCPET	2.733	2.019	3.582	1.916
17	virginic	CSSCP	LONGSEP	19.813	4.594	14.861	2.406
18	virginic	CSSCP	ANCSEP	4.594	5.096	3.498	2.334
19	virginic	CSSCP	LONGPET	14.861	3.498	14.925	2.392
20	virginic	CSSCP	ANCPET	2.406	2.334	2.392	3.696
21		PSSCP	LONGSEP	38.956	13.630	24.625	5.645
22		PSSCP	ANCSEP	13.630	16.962	8.121	4.808
23		PSSCP	LONGPET	24.625	8.121	27.223	6.272
24		PSSCP	ANCPET	5.645	4.808	6.272	6.157
25		BSSCP	LONGSEP	63.212	-19.953	165.248	71.279
26		BSSCP	ANCSEP	-19.953	11.3449	-57.240	-22.933
27		BSSCP	LONGPET	165.248	-57.2396	437.103	186.774
28		BSSCP	ANCPET	71.279	-22.9327	186.774	80.413
29		CSSCP	LONGSEP	102.168	-6.3227	189.873	76.924
30		CSSCP	ANCSEP	-6.323	28.3069	-49.119	-18.124
31		CSSCP	LONGPET	189.873	-49.1188	464.325	193.046
32		CSSCP	ANCPET	76.924	-18.1243	193.046	86.570
33		RSQUARED		0.619	0.4008	0.941	0.929
34	setosa	COV	LONGSEP	0.124	0.0992	0.016	0.010
35	setosa	COV	ANCSEP	0.099	0.1437	0.012	0.009
36	setosa	COV	LONGPET	0.016	0.0117	0.030	0.006
37	setosa	COV	ANCPET	0.010	0.0093	0.006	0.011
38	versicol	COV	LONGSEP	0.266	0.0852	0.183	0.056
39	versicol	COV	ANCSEP	0.085	0.0985	0.083	0.041
40	versicol	COV	LONGPET	0.183	0.0827	0.221	0.073
41	versicol	COV	ANCPET	0.056	0.0412	0.073	0.039
42	virginic	COV	LONGSEP	0.404	0.0938	0.303	0.049
43	virginic	COV	ANCSEP	0.094	0.1040	0.071	0.048
44	virginic	COV	LONGPET	0.303	0.0714	0.305	0.049
45	virginic	COV	ANCPET	0.049	0.0476	0.049	0.075

46		PCOV	LONGSEP	0.265	0.0927	0.168	0.038
47		PCOV	ANCSEP	0.093	0.1154	0.055	0.033
48		PCOV	LONGPET	0.168	0.0552	0.185	0.043
49		PCOV	ANCPET	0.038	0.0327	0.043	0.042
50		BCOV	LONGSEP	0.632	-0.1995	1.652	0.713
51		BCOV	ANCSEP	-0.19953	0.11345	-0.57240	-0.22933
52		BCOV	LONGPET	1.65248	-0.57240	4.37103	1.86774
53		BCOV	ANCPET	0.71279	-0.22933	1.86774	0.80413
54		COV	LONGSEP	0.68569	-0.04243	1.27432	0.51627
55		COV	ANCSEP	-0.04243	0.18998	-0.32966	-0.12164
56		COV	LONGPET	1.27432	-0.32966	3.11628	1.29561
57		COV	ANCPET	0.51627	-0.12164	1.29561	0.58101
58	setosa	STD		0.35249	0.37906	0.17366	0.10539
59	versicol	STD		0.51617	0.31380	0.46991	0.19775
60	virginic	STD		0.63588	0.32250	0.55189	0.27465
61		PSTD		0.51479	0.33969	0.43033	0.20465
62		BSTD		0.79506	0.33682	2.09070	0.89673
63		STD		0.82807	0.43587	1.76530	0.76224
64	setosa	CORR	LONGSEP	1.00000	0.74255	0.26718	0.27810
65	setosa	CORR	ANCSEP	0.74255	1.00000	0.17770	0.23275
66	setosa	CORR	LONGPET	0.26718	0.17770	1.00000	0.33163
67	setosa	CORR	ANCPET	0.27810	0.23275	0.33163	1.00000
68	versicol	CORR	LONGSEP	1.00000	0.52591	0.75405	0.54646
69	versicol	CORR	ANCSEP	0.52591	1.00000	0.56052	0.66400
70	versicol	CORR	LONGPET	0.75405	0.56052	1.00000	0.78667
71	versicol	CORR	ANCPET	0.54646	0.66400	0.78667	1.00000
72	virginic	CORR	LONGSEP	1.00000	0.45723	0.86422	0.28111
73	virginic	CORR	ANCSEP	0.45723	1.00000	0.40104	0.53773
74	virginic	CORR	LONGPET	0.86422	0.40104	1.00000	0.32211
75	virginic	CORR	ANCPET	0.28111	0.53773	0.32211	1.00000
76		PCORR	LONGSEP	1.00000	0.53024	0.75616	0.36451
77		PCORR	ANCSEP	0.53024	1.00000	0.37792	0.47053
78		PCORR	LONGPET	0.75616	0.37792	1.00000	0.48446
79		PCORR	ANCPET	0.36451	0.47053	0.48446	1.00000
80		BCORR	LONGSEP	1.00000	-0.74507	0.99413	0.99977
81		BCORR	ANCSEP	-0.74507	1.00000	-0.81284	-0.75926
82		BCORR	LONGPET	0.99413	-0.81284	1.00000	0.99623
83		BCORR	ANCPET	0.99977	-0.75926	0.99623	1.00000
84		CORR	LONGSEP	1.00000	-0.11757	0.87175	0.81794
85		CORR	ANCSEP	-0.11757	1.00000	-0.42844	-0.36613
86		CORR	LONGPET	0.87175	-0.42844	1.00000	0.96287
87		CORR	ANCPET	0.81794	-0.36613	0.96287	1.00000
88	setosa	STDMEAN		-1.01119	0.85041	-1.30063	-1.25070
89	versicol	STDMEAN		0.11191	-0.65922	0.28437	0.16618
90	virginic	STDMEAN		0.89928	-0.19119	1.01626	1.08453
91	setosa	PSTDMEAN		-1.62656	1.09120	-5.33538	-4.65836
92	versicol	PSTDMEAN		0.18001	-0.84587	1.16653	0.61894
93	virginic	PSTDMEAN		1.44655	-0.24532	4.16885	4.03942
94		CANCORR	CAN1	0.98482	0.98482	0.98482	0.98482
95		CANCORR	CAN2	0.47120	0.47120	0.47120	0.47120
96		STRUCTUR	CAN1	0.79189	-0.53076	0.98495	0.97281
97		STRUCTUR	CAN2	0.21759	0.75799	0.04604	0.22290
98		BSTRUCT	CAN1	0.99147	-0.82566	0.99975	0.99404
99		BSTRUCT	CAN2	0.13035	0.56417	0.02236	0.10898
100		PSTRUCT	CAN1	0.22260	-0.11901	0.70607	0.63318
101		PSTRUCT	CAN2	0.31081	0.86368	0.16770	0.73724
102		SCORE	CAN1	-0.68678	-0.66883	3.88580	2.14224
103		SCORE	CAN2	0.01996	0.94344	-1.64512	2.16414
104		PSCORE	CAN1	-0.42695	-0.52124	0.94726	0.57516
105		PSCORE	CAN2	0.01241	0.73526	-0.40104	0.58104
106		RAWSCORE	CAN1	-0.82938	-1.53447	2.20121	2.81046
107		RAWSCORE	CAN2	0.02410	2.16452	-0.93192	2.83919
108	setosa	CANMEAN	CAN1	-7.60760	-7.60760	-7.60760	-7.60760
109	setosa	CANMEAN	CAN2	0.21513	0.21513	0.21513	0.21513
110	versicol	CANMEAN	CAN1	1.82505	1.82505	1.82505	1.82505
111	versicol	CANMEAN	CAN2	-0.72790	-0.72790	-0.72790	-0.72790
112	virginic	CANMEAN	CAN1	5.78255	5.78255	5.78255	5.78255
113	virginic	CANMEAN	CAN2	0.51277	0.51277	0.51277	0.51277

Plot of CAN2*CAN1. Symbol is value of SIMB.



El archivo de resultados presenta:

. Hay 150 observaciones, 4 variables y 3 clases.

. Número de individuos de cada clase, que es 50 para cada una de ellas.

. Distancias de *Mahalanobis* entre las tres variedades. La distancia mayor es entre la variedad setosa y virginica que es de 179.3847, y la menor es entre versicolor y virginica que es de 17.2011.

. Análisis Multivariante de la Varianza en el que se observa que es significativa la *lambda de Wilk* por lo que existen diferencias significativa entre las tres variedades para las cuatro medidas realizadas.

. Las dos correlaciones canónicas, esto es, máxima correlación múltiple posible de la primera función lineal con los grupos y la segunda máxima correlación múltiple de la segunda función lineal con los grupos.

. Valores propios, el primero explica el 99% de la varianza y el segundo explica el 1% de la varianza

. Pruebas de hipótesis usando la razón de verosimilitud que indican que ambas correlaciones canónicas son significativas, por lo que es adecuado utilizar dos ejes canónicos. La primera prueba coincide con la prueba de la *lambda de Wilk*.

. Una serie de coeficientes de interés diverso. Los más importantes son los encabezados por **Total-Sample Standardized Canonical Coefficients** pues estos son los coeficientes de sendas funciones lineales que aplicados a los datos tipificados proveerán las variables canónicas

. La salida del procedimiento **Print** en el que se imprime multitud de estadísticos del análisis canónico, entre otros tiene interés los últimos seis estadísticos que son las medias de las dos variables canónica de las tres especies. Estos valores pueden representarse en la salida del procedimiento **Plot** para señalar la posición de los tres grupos. Puesto que el tamaño de muestra es el mismo para los tres grupos el radio de los tres grupo es el mismo, este es

$$F_{(4,144;\ 0.05)} = 2.40$$

$$R^2 = 2.4\frac{588}{144} = 9.8685$$

$$LC = \frac{3.1414}{\sqrt{50}} = 0.4443$$

que representados nos da que las tres poblaciones están claramente separadas.

.Salida del procedimiento **Plot** en el que se han representado las dos variables canónicas, se observa que hacia la izquierda del primer eje y en el centro del segundo eje se encuentra el grupo de la variedad *setosa*; en el centro derecha del primer eje esta el grupo de la variedad *versicolor* y hacia la derecha del primer eje esta el grupo de la variedad *virginica*.

Selección de variables en el Análisis Discriminante.-

Recuérdese la analogía que existe entre el análisis discriminante y el de regresión. Como ocurre en la regresión múltiple (ver epígrafe *Importancia relativa de diferentes variables X* del Capítulo 13) en algunas situaciones se dispone de un número considerable de variables que, en la práctica, están correlacionadas, lo que hace más difícil la respuesta a la pregunta de la importancia relativa de cada variable y no se sabe cuáles de ellas contribuyen mejor a discriminar entre los grupos.

Muchas de las variables elegidas pueden contribuir muy poco o nada a la precisión del análisis. Por ejemplo, se puede comenzar con 11 variables y seleccionar tres de ellas que proveen una mejor discriminación.

El problema en este caso es saber cuantas variable se necesitan y cuales son.

Un enfoque lógico para resolver este problemas sería el de elaborar el análisis con todos los posibles subgrupos de las variables, es decir, realizar el análisis simple con cada una de la variables, después todas las combinaciones de dos variables, después con todas las combinaciones de tres variables, *etc.* El conjunto que de una mayor discriminación sería el escogido para hacer la predicción. Si hubiera dos subconjuntos con el mismo poder de discriminación, se elegiría el de menos cantidad de variables. Como es fácil ver, este método falla por la cantidad de cálculos que son necesarios. Si se tienen p variable habría que realizar 2^p-1 análisis, es decir, si se tienen 11 variables habría que realizar 2047 análisis. Aún con un ordenador esto es muy complicado.

Hay varias maneras de resolver esta cuestión, se van ha estudiar aquí el denominado *método ascendente* (*forward*) de selección de variables; el *método descendente* (*backward*) de eliminación de variables; y el *método paso a paso* (*stepwise*) de selección de variables. Para utilizar estos métodos se asume la distribución normal multivariante con la misma matriz de covarianzas.

Se pueden utilizar dos criterios para seleccionar o eliminar una variable: 1) el

nivel de significación de la prueba *F* del análisis de covarianza en el que las variables ya elegidas son las covariables y la variable que se esta evaluando es la variable dependiente; 2) el cuadrado de la correlación parcial de la variable que se esta evaluando con las variable de *clases* teniendo en cuenta las variables ya seleccionadas.

En el método ascendente, se comienza calculando todos los análisis canónicos simples para cada variable, la variable que de una mayor discriminación, medida por la razón de verosimilitud de la *lambda de Wilks*, será la variable seleccionada. Después se calculan todos los análisis canónicos bivariantes en la que aparezca la primera variable seleccionada anteriormente; la pareja que dé una mayor razón de verosimilitud es la seleccionada. Después se calculan todos los análisis con tres variables en las que aparezca la pareja seleccionada anteriormente seleccionada, y así sucesivamente hasta que no se encuentren variables que cumplan el criterio de selección.

En el método descendente primero se realiza el análisis canónico con todas las variables, después en cada paso se elige la variable que menor contribuya al poder discriminante según el criterio de la *lambda de Wilks* y se contrasta el poder discriminante de esta variable, si no cumple el criterio de razón de verosimilitud es eliminada. El proceso se detiene cuando todas las variables cumplan el criterio.

El método de paso a paso es un perfeccionamiento del método ascendente consistente en que en cada paso se considera la inclusión o exclusión de las variables que se habían introducido en pasos anteriores. En el método ascendente, una variable que fuera la mejor en un paso anterior queda definitivamente incluida, sin embargo puede ocurrir que esta variable sea superflua en una fase posterior debido a la relación existente entre dicha variable y variables que se han introducido posteriormente en el modelo, esto lo evalúa el método paso a paso.

Estos métodos no tienen porque elegir las mismas variables, ni ninguno elige las variables que elegiría el método lógico impracticable. Estas diferencias no son muy preocupantes pues la interrelación entre las variables hace que diferentes subgrupos puedan proporcionar discriminaciones semejantes.

El procedimiento **STEPDISC** del *SAS* realiza, por defecto, una selección de variables utilizando el método paso a paso (**stepwise**). Si se desea alguno de los otros dos métodos hay que epecificarlo como opción de método en el procedimiento de la siguiente manera: para el método descendente (**backward**), **PROC STEPDISC METHOD=BW**; para el método ascendente (**forward**), **PROC STEPDISC METHOD=FW**.

Ejemplo.-

Sigamos con el ejemplo de la clasificación de las flores del género *Iris* del trabajo clásico de *Fisher* . Hagamos primero un análisis dicriminante por el método paso a paso (**stepwise**)

Archivo de programa *SAS* (C19-3.SAS).-

```
title 'Análisis discriminante por el método paso a paso';
Options ls=80 ps=60;
data seldis;
infile 'c19-1.dat';
input longsep ancsep longpet ancpet variedad $ clase $ @@;
proc stepdisc ;
 class variedad;
 var longsep ancsep longpet ancpet;
run;
```

Archivo de resultados (C19-3.LST).-

```
                        Stepwise Discriminant Analysis

            150 Observations          4 Variable(s) in the Analysis
              3 Class Levels          0 Variable(s) will be included

             The Method for Selecting Variables will be: STEPWISE
             Significance Level to Enter =  0.1500
             Significance Level to Stay  =  0.1500
                           Class Level Information

             VARIEDAD     Frequency        Weight     Proportion
             setosa              50       50.0000       0.333333
             versicol            50       50.0000       0.333333
             virginic            50       50.0000       0.333333

Stepwise Selection:  Step 1
                     Statistics for Entry, DF = 2, 147

       Variable          R**2           F        Prob > F     Tolerance
       LONGSEP         0.6187       119.265       0.0001        1.0000
       ANCSEP          0.4008        49.160       0.0001        1.0000
       LONGPET         0.9414      1180.161       0.0001        1.0000
       ANCPET          0.9289       960.007       0.0001        1.0000

                      Variable LONGPET will be entered

                The following variable(s) have been entered:
                                  LONGPET

                              Multivariate Statistics
  Wilks' Lambda  = 0.05862828     F( 2, 147) = 1180.161       Prob > F = 0.0001
  Pillai's Trace =   0.941372     F( 2, 147) = 1180.161       Prob > F = 0.0001

             Average Squared Canonical Correlation = 0.47068586

------------------------------------------------------------------------------
Stepwise Selection:  Step 2
                     Statistics for Removal,  DF = 2, 147
                Variable          R**2           F         Prob > F
                LONGPET         0.9414      1180.161        0.0001
                           No variables can be removed

                     Statistics for Entry, DF = 2, 146

                    Partial
       Variable          R**2           F        Prob > F     Tolerance
       LONGSEP         0.3198       34.323       0.0001        0.2400
       ANCSEP          0.3709       43.035       0.0001        0.8164
       ANCPET          0.2533       24.766       0.0001        0.0729

                      Variable ANCSEP will be entered
                The following variable(s) have been entered:
                              ANCSEP   LONGPET
                             Multivariate Statistics
  Wilks' Lambda  = 0.03688411     F( 4, 292) =  307.105       Prob > F = 0.0001
  Pillai's Trace =   1.119908     F( 4, 294) =   93.528       Prob > F = 0.0001

             Average Squared Canonical Correlation = 0.55995394
------------------------------------------------------------------------------
Stepwise Selection:  Step 3
                     Statistics for Removal,  DF = 2, 146
                              Partial
                Variable          R**2           F         Prob > F
                ANCSEP          0.3709       43.035        0.0001
                LONGPET         0.9384     1112.954        0.0001
                           No variables can be removed
                     Statistics for Entry, DF = 2, 145
```

```
                    Partial
       Variable        R**2          F        Prob > F     Tolerance
       LONGSEP       0.1447       12.268        0.0001        0.1323
       ANCPET        0.3229       34.569        0.0001        0.0662
                   Variable ANCPET will be entered
              The following variable(s) have been entered:
                        ANCSEP   LONGPET  ANCPET
                       Multivariate Statistics
  Wilks' Lambda  = 0.02497554    F( 6, 290) =  257.503      Prob > F = 0.0001
  Pillai's Trace =   1.189914    F( 6, 292) =   71.485      Prob > F = 0.0001
            Average Squared Canonical Correlation = 0.59495691
-------------------------------------------------------------------------------
Stepwise Selection:  Step 4

                   Statistics for Removal,  DF = 2, 145

                         Partial
              Variable      R**2          F        Prob > F
              ANCSEP      0.4295       54.577        0.0001
              LONGPET     0.3482       38.724        0.0001
              ANCPET      0.3229       34.569        0.0001

                     No variables can be removed
                   Statistics for Entry, DF = 2, 144
                  Partial

       Variable        R**2          F        Prob > F     Tolerance
       LONGSEP       0.0615       4.721        0.0103        0.0320

                    Variable LONGSEP will be entered
                    All variables have been entered
                       Multivariate Statistics
  Wilks' Lambda  = 0.02343863    F( 8, 288) =  199.145      Prob > F = 0.0001
  Pillai's Trace =   1.191899    F( 8, 290) =   53.466      Prob > F = 0.0001

            Average Squared Canonical Correlation = 0.59594941
-------------------------------------------------------------------------------
Stepwise Selection:  Step 5

                   Statistics for Removal,  DF = 2, 144
                         Partial

              Variable      R**2          F        Prob > F
              LONGSEP     0.0615       4.721        0.0103
              ANCSEP      0.2335      21.936        0.0001
              LONGPET     0.3308      35.590        0.0001
              ANCPET      0.2570      24.904        0.0001

                     No variables can be removed

No further steps are possible

-------------------------------------------------------------------------------
Stepwise Selection:  Summary

                Variable          Number   Partial          F          Prob >
   Step  Entered     Removed        In       R**2       Statistic        F
  ------------------------------------------------------------------------
      1  LONGPET                      1     0.9414       1180.161     0.0001
      2  ANCSEP                       2     0.3709         43.035     0.0001
      3  ANCPET                       3     0.3229         34.569     0.0001
      4  LONGSEP                      4     0.0615          4.721     0.0103

Stepwise Selection:  Summary

                                                          Average
                                                          Squared
           Variable          Number       Wilks'  Prob <   Canonical  Prob >
Step  Entered     Removed       In         Lambda  Lambda  Correlation   ASCC
-------------------------------------------------------------------------------
   1  LONGPET                     1  0.05862828  0.0001   0.47068586  0.0001
   2  ANCSEP                      2  0.03688411  0.0001   0.55995394  0.0001
   3  ANCPET                      3  0.02497554  0.0001   0.59495691  0.0001
   4  LONGSEP                     4  0.02343863  0.0001   0.59594941  0.0001
```

Por defecto se puede tomar como criterio de inclusión la *tolerancia* que es una medida del grado de asociación lineal entre las variables independientes. Para la *i-ésima* variable la tolerancia es $1-R_i^2$ siendo R^2 el coeficiente de determinación múltiple en el que la variable *i* es considerada una variable dependiente y las demás variables son consideradas como independientes. Variables con valores bajos de la tolerancia (inferiores a 0.001) no deben entrar en el análisis.

También se puede utilizar como criterio la prueba ***F*** del análisis de covarianza en la que la variable considerada es la variable dependiente y las variables ya elegidas actúan como covariables.

En el paso 1 (*step* 1), la tolerancia es 1.0 para todas las variables porque todavía no hay ninguna variable en el modelo. Se selecciona la variable *longitud del pétalo* (***LONGPET***) porque tiene una *F*=1180.161 que es la mayor y altamente significativa. En el paso 2, la variable longpet que está incluida en el modelo es probada de nuevo antes de seleccionar una nueva variable, se mantiene esta variable y se selecciona la variable *anchura del sépalo* (***ANCSEP***) que tiene la *F* mayor, 43.035. También es la que tiene la mayor tolerancia (0.8164). Este proceso se repite en el paso 3 y el paso 4.

Análisis Cluster

Este nombre se utiliza para definir una serie de técnicas o algoritmos que tienen como objeto la búsqueda de grupos similares de individuos o de variables que se van agrupando en conglomerados previamente *desconocidos*. Dada una muestra de individuos a los que se les ha medido una serie de variables, el análisis cluster sirve para clasificarlos en grupos lo más homogéneos posibles en base a las variables observadas. La diferencia con el *análisis discriminante* es que en este los grupos se conocen a priori, mientras que en el análisis cluster las características de los grupos derivan o vienen sugeridas por los datos y usualmente no son conocidas a priori. La salida de los análisis cluster se utilizan para realizar agrupaciones desconocidas previamente tipo dendogramas.

El término ingles *cluster* significa racimo, piña, agrupación, conglomerado, montón, *etc.* por lo que cada vez que nos referimos a un grupo, agrupación, *etc* nos estamos refiriendo a un cluster. Dada la utilización generalizada del término ingles, también se seguirá utilizando en este manual.

Este tipo de análisis es muy utilizado en taxonomía numérica, también conocida como *clasificación automática*. Como se sabe, la taxonomía consiste en clasificar los seres vivos en grupos arbitrarios en base a sus características, yendo de lo general a lo particular, esto es, *Reino, Phylum, Clase, Orden, Familia, Genero* y *especie*. Pero los análisis cluster también pueden usarse, por ejemplo, en sanidad para asignar pacientes a una categoría de diagnóstico especifica en base a los síntomas y signos presentes. También se ha usado en antropología para clasificar herramientas de piedra, cerámicas o restos fósiles de las civilizaciones que las han producido. Los consumidores pueden agruparse según la elección de la compra en investigaciones de marketing. Todos estos ejemplos muestran la utilidad y versatilidad del análisis cluster.

El análisis cluster es eminentemente empírico. Diferentes métodos da diferentes agrupaciones tanto en número como en contenido. Con la agravante de que al no existir una idea previa puede resultar difícil juzgar cual de todos los posibles resultados tiene más sentido en el contexto del estudio que se está realizando.

Ejemplo ilustrativo.-

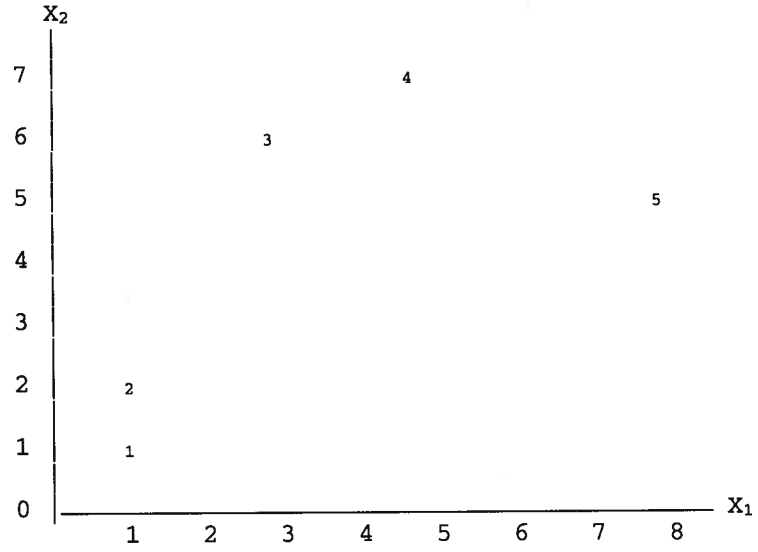
Para ilustrar las ideas que se van ha exponer pongamos un ejemplo simple consistente en cinco individuos a los que se la han medido dos variables.

Individuo	X_1	X_2
1	1	1
2	1	2
3	3	6
4	5	7
5	8	5

Con un ejemplo tan simple el análisis incluso puede realizarse sin ayuda del ordenador.

Conceptos básicos previos a la presentación de los análisis cluster: Diagrama de dispersión.-

En el caso de dos variables, un diagrama de dispersión puede ser muy útil para visualizar las características principales de los grupos. En el caso del ejemplo hipotético el diagrama de dispersión es



En este diagrama se observa que los puntos más próximos son el uno y el dos. Esta observación puede inducirnos a considerar que estos dos puntos son un cluster. Otro cluster puede ser el que contengan a los individuos tres y cuatro, constituyendo en individuo cinco un tercer cluster. Sin embargo puede haber quien opine que los puntos tres, cuatro y cinco son el segundo cluster. Esto ilustra la indeterminación del análisis cluster, puesto que el número de cluster son desconocidos. Nótese también que el concepto de proximidad va implícito en la determinación de las agrupaciones. En el siguiente epígrafe se ampliará este concepto con las definiciones de distancias.

Si el número de variables es pequeño, es posible examinar diagramas de dispersión para cada par de variables y buscar posibles agrupamientos. Pero esta técnica puede hacerse impracticable para más de cuatro variables y especialmente si el número de individuos es grande.

Medidas de distancias.-

En todos los métodos cluster se necesita definir medidas de la cercanía o *similitud* de dos observaciones. El concepto contrario al de similitud es el de la *distancia*. Pero antes de definir medidas de distancias hay que advertir que muchas de estas medidas son muy sensibles a los datos extraños, por lo que es conveniente antes de utilizarlas el realizar una comprobación de que no hay ningún dato erróneo.

La distancia entre los individuos *i-ésimo* y *j-ésimo* es una cantidad que debe cumplir

$$d_{ii} = 0$$
$$d_{ij} \geq 0$$
$$d_{ij} = d_{ji}$$
$$d_{ij} \leq d_{ik} + d_{kj}$$

El valor de una distancia es mayor cuanto mas diverjan los individuos entre los que se miden la distancia. Los valores tipificados de las variables (valores Z) pueden ser considerados como valores de distancias, pues cuanto mayor sea el valor absoluto de la variable tipificada mayor es la lejanía que separa al individuo de la media. Dentro de un momento veremos las distancias de *Mahalanobis*, estas distancias coincide con el valor de Z si se tienen datos univariantes, esto es, la distancia entre el grupo *i-ésimo* y el grupo *j-ésimo* en los que se ha tomado una muestra aleatoria de la variable X es

$$D^2 = \frac{(\overline{X}_i - \overline{X}_j)^2}{S^2}$$

La similitud por contra será mayor cuanto más semejantes sean dos individuos. Los coeficientes de correlación (*Pearson*, *Spearman*, *Kendall*, *etc.*) son medidas de similitud

La distancia más comúnmente usada es la *distancia euclídea*. Si a cada individuos se le ha medido dos variables, se pueden representar los individuos en unos ejes de coordenadas y las distancia euclídea no es sino la aplicación del teorema de *Pitágoras*, esta es

$$d = \sqrt{(X_{11} - X_{12})^2 + (X_{21} - X_{22})^2}$$

Para el ejemplo que se está desarrollando, la distancia euclídea entre el individuo cuatro y cinco es

$$d_{(4,5)} = \sqrt{(5-7)^2 + (8-5)^2} = 3.61$$

Para *p* variables, la distancia euclídea de dos individuos es la raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de cada variable para los dos individuos, por ejemplo, la distancia entre el individuo cuatro y cinco con *p* variables es

$$d = \sqrt{\sum_{j=1}^{p} (X_{4j} - X_{5j})^2}$$

Puesto que la operación de extraer la raíz cuadrada no cambia el orden de las distancias muchas veces se verán estas distancias elevadas al cuadrado. Otra veces se ven como la suma de otras potencias, diferentes del cuadrado, de los valores absolutos de las diferencias. Por ejemplo, si los valores absolutos de las diferencias se elevan a la unidad, la distancia es la suma de las diferencias absolutas, este tipo de distancia se denomina distancia en *manzana* pues para dos dimensiones equivale a lo que habría que andar entre dos puntos en una ciudad cuadriculada en manzanas.

Existen una multitud de distancias pero la más comúnmente utilizada es la que ya se ha empleado en el análisis discriminante, esta es la distancia de *Mahalanobis*. Se ha visto hace un momento que con una sola variable la distancia de *Mahalanobis* entre dos grupos no es sino el cuadrado de la diferencia de las variable tipificada, por lo que la distancia de *Mahalanobis* no es sino la generalización a *p* variables tipificadas, teniendo en cuenta la correlación entre las variables. El cuadrado de la distancia euclídea basada en las variables tipificadas es la suma de los cuadrados de las diferencias, cada una dividida por la correspondiente varianza. Cuando las variables están correlacionadas, se puede definir una distancia teniendo en cuenta esta correlación. La distancia de *Mahalanobis* es justamente lo siguiente: para dos variables, X_1 y X_2, cuyas varianzas muestrales son, respectivamente, S^2_1 y S^2_2 y están correlacionadas con *r*, el cuadrado de la distancia euclidea basada en las variables originales es

$$d^2 = (X_{11} - X_{12})^2 + (X_{21} - X_{22})^2$$

La misma cantidad basada en las variables tipificadas es

$$d_Z^2 = D^2 = \frac{(X_{11} - X_{12})^2}{S_1^2} + \frac{(X_{21} - X_{22})^2}{S_2^2}$$

Si r=0, entonces esta última cantidad es también la distancia de *Mahalanobis*. Si r=/0, las distancias de *Mahalanobis* es

$$D^2 = \frac{1}{1 - r^2}\left[\frac{(X_{11} - X_{12})^2}{S_1^2} + \frac{(X_{21} - X_{22})^2}{S_2^2} - \frac{2r(X_{11} - X_{12})(X_{21} - X_{22})}{S_1 S_2}\right]$$

Para más de dos variables es facil definir la distancia de *Mahalanobis* en base a vectores y matrices. La distancia de *Mahalanobis* entre el grupo *i-ésimo* y el *j-ésimo* es

$$D^2_{i,j} = (\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)' \mathbf{V}^{-1} (\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)$$

siendo **V** la matriz de varianza-covarianza y siendo $\overline{X}_i$ el vector cuyas componentes son las medias de las diferentes variables en el grupo *i-ésimo*.

En la mayoría de las situaciones diferentes medidas de distancias darán, lógicamente, diferentes matrices de distancias y por tanto conducirán a diferentes agrupamientos. Si las unidades en las que se miden las variables son diferentes es recomendable estandarizarlas antes de calcular las distancias. Este procedimiento es particularmente útil cuando el rango de una variable es mucho mayor que el de otra. Además, como ya se ha visto, las distancias de *Mahalanobis* son muy adecuadas cuando las variables están correlacionadas.

Técnicas analíticas de agrupamientos.-

Los métodos relativos al análisis cluster pueden ser de dos tipos: *clasificación jerárquica* y *clasificación no jerárquica*. Una clasificación es no jerárquica cuando se forman grupos homogéneos sin establecer relación entre los grupos; mientras que una clasificación es jerárquica cuando los grupos se van fusionando progresivamente mientras decrece la homogeneidad entre los grupos, cada vez más amplios, que se van formando. La clasificación de los seres vivos de *Linneo* es un ejemplo claro de clasificación jerárquica. El paquete estadístico *SAS* utiliza sendos procedimientos para cada uno de estos métodos; para los métodos jerárquicos utiliza el procedimiento **CLUSTER** y para los métodos no jerárquicos utiliza el procedimiento **FASTCLU**.

Clasificación jerárquica.-

Los métodos jerárquicos pueden, a su vez, diferenciarse en *métodos aglomerativos* o ascendentes y *métodos divisivos* o descendentes. En los *métodos aglomerativos* se comienza con *N* grupos, esto es, cada individuo constituye su propio grupo, para en sucesivos pasos combinar los dos grupos más próximos, de manera que en cada paso se reduce el número de grupos en uno. Al final todas las observaciones están reunidas en un solo cluster. Los *métodos divisivos* siguen un proceso inverso a los aglomerativos. Comienza con un solo cluster que engloba a todos los individuos para en sucesivos pasos ir desgajando grupos en función de sus disimilitudes, acabando el proceso con tanto grupos como individuos.

Los más utilizados, con mucho, son los métodos aglomerativos por lo que son los únicos que estudiaremos. El procedimiento **CLUSTER** del *SAS* provee 11 métodos jerárquicos aglomertivos, alguno de ello son;

. El método de las *distancias mínimas* (**PROC CLUSTER METHOD = SIGLE**) también conocido como el método *del vecino más próximo* (nearest neighbor). Es un método muy simple consistente en medir la distancia entre dos grupos por la distancia entre sus puntos más próximos y con arreglo a esto ir agrupando los individuos o grupos que tienen menor distancia.

. El método de las *distancias máximas* (**PROC CLUSTER METHOD = COMPLETE**), también denominado del vecino más lejano consiste en medir la distancia entre dos grupos por la distancia entre sus puntos más alejados y con arreglo a esto ir agrupando los individuos o grupos que tienen menor distancia.

. El método del *promedio entre grupos* (**PROC CLUSTER METHOD = AVERAGE**), consiste en medir la distancia entre dos grupos por la media de las distancias de todos los pares de individuos que se pueden formar con los individuos de uno y otro grupo, y en base a esto ir agrupando los individuos o grupos que tienen

menor distancia. Como se ve, este método utiliza la información de todas las distancias entre los dos grupos y no solo la distancia mas corta o la más larga como ocurría con los métodos anteriores.

. El método del *centroide* (**PROC CLUSTER METHOD = CENTROID**), consiste en medir la distancia entre dos grupos por la distancia entre las medias de los respectivos grupos y en base a esto ir agrupando los individuos o grupos que tienen menor distancia. El centroide de un grupo es el punto medio de todos los individuos que constituyen el grupo. Si un grupo está formado por un solo individuo entonces los valores de ese individuo es el centroide. Como se ve, este método también utiliza la información de todas las distancias entre los dos grupos.

. El método de *la mínima varianza de Ward* (**PROC CLUSTER METHOD = WARD**), consiste en medir la distancia entre dos grupos por la suma de cuadrados entre los dos grupos (como si fuera una ANOVA de una vía) y en base a esto ir agrupando los individuos o grupos que tienen menor distancia.

De todos los métodos jerárquicos aglomerativos lo más utilizados son el del *centroide*, el de la *media* y el de la *varianza mínima de Ward*. En los tres se utiliza el cuadrado de las distancias euclídeas a no ser que se especifique la opción **NOSQUARE**.

Veamos cual es la idea conductora de estos métodos aglomerativos de clasificación jerárquica desarrollando a mano el *método de las distancias mínimas*, primero, y el *método del centroide*, después, de los datos del ejemplo ilustrativo

Método de las distancias mínimas.-

Las distancias euclídeas de los individuos son

	1	2	3	4	5
1	0	1	5.385	7.211	8.062
2		0	4.472	6.403	7.616
3			0	2.236	5.099
4				0	3.605
5					0

Método del mínimo: como se ve la distancia más pequeña tiene el valor 1 y es la que separa al individuo uno del dos, por lo que se hace un grupo (cluster 1) con estos dos individuos. La distancia más corta entre este cluster y los individuos tres, cuatro y cinco son, respectivamente, 4.472, 6.403 y 7.616, por lo que esas son las que se toman para la nueva matriz de distancias

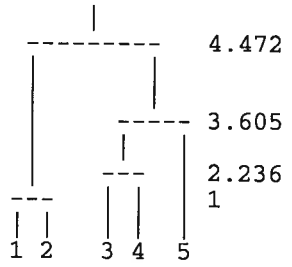
	1,2	3	4	5
1,2	0	4.472	6.403	7.616
3			2.236	5.099
4		0	0	3.605
5				0

la distancia más pequeña tiene el valor 2.236 y es la que separa al individuo tres del cuatro, por lo que se hace un grupo (cluster 2) con estos dos individuos. La distancia más corta entre este cluster y el cluster 1 y el individuo cinco son, respectivamente, 4.472 y 3.605, por lo que esas son las que se toman para la nueva matriz de distancias. Y así también para la última matriz

	1,2	3,4	5
1,2	0	4.472	7.616
3,4		0	3.605
5			0

	1,2	3,4,5
1,2	0	4.472
3,4,5		0

Con estos resultados se puede representar el siguiente dendograma



El paquete estadístico *SAS* realiza esto mismo con el siguiente programa

Archivo de programa *SAS* (C19-4.SAS).-

```
title 'Cluster';
options ls=75 ps=60;
data cluster;
infile 'c19-4.dat';
input indi x1 x2;
proc cluster method=single nonorm;
id indi;
run;
proc tree inc=0.25;
id indi;
run;
```

El procedimiento **CLUSTER** hace el algoritmo y el procedimiento **TREE** realiza el dendograma. La opción **NONORM** es para que nos presente las distancias no normalizadas que son las que se han utilizado para realizar el ejemplo a mano.

Archivo de datos (C19-4.DAT).-

```
1 1 1
2 1 2
3 3 6
4 5 7
5 8 5
```

Archivo de resultados (C19-4.LST).-

```
                    Single Linkage Cluster Analysis
                  Eigenvalues of the Covariance Matrix
          Eigenvalue      Difference      Proportion      Cumulative
  1         12.9570         10.4139        0.835933         0.83593
  2          2.5430               .        0.164067         1.00000
    Root-Mean-Square Total-Sample Standard Deviation = 2.783882
  Number                                      Frequency
    of                                          of New     Minimum
 Clusters    ------Clusters Joined------       Cluster    Distance    Tie
      4                1              2             2    1.000000
      3                3              4             2    2.236068
      2     CL3                       5             3    3.605551
      1     CL4             CL2                     5    4.472136

                                    INDI
                                1     2     3     4     5
                          4.5 +XXXXXXXXXXXXXXXXXXXXXXXXX
                              |XXXXXXX     XXXXXXXXXXXXX
                        M 4.25 +XXXXXXX     XXXXXXXXXXXXX
                        i     |XXXXXXX     XXXXXXXXXXXXX
                        n   4 +XXXXXXX     XXXXXXXXXXXXX
                        i     |XXXXXXX     XXXXXXXXXXXXX
                        m 3.75 +XXXXXXX     XXXXXXXXXXXXX
                        u     |XXXXXXX     XXXXXXXXXXXXX
                        m  3.5 +XXXXXXX     XXXXXXX     .
                              |XXXXXXX     XXXXXXX     .
                        D 3.25 +XXXXXXX     XXXXXXX     .
                        i     |XXXXXXX     XXXXXXX     .
                        s   3 +XXXXXXX     XXXXXXX     .
                        t     |XXXXXXX     XXXXXXX     .
                        a 2.75 +XXXXXXX     XXXXXXX     .
                        n     |XXXXXXX     XXXXXXX     .
                        c  2.5 +XXXXXXX     XXXXXXX     .
                        e     |XXXXXXX     XXXXXXX     .
                          2.25 +XXXXXXX     XXXXXXX     .
                        B     |XXXXXXX     .     .     .
                        e   2 +XXXXXXX     .     .     .
                        t     |XXXXXXX     .     .     .
                        w 1.75 +XXXXXXX     .     .     .
                        e     |XXXXXXX     .     .     .
                        e  1.5 +XXXXXXX     .     .     .
                        n     |XXXXXXX     .     .     .
                          1.25 +XXXXXXX     .     .     .
                        C     |XXXXXXX     .     .     .
                        l   1 +XXXXXXX     .     .     .
                        u     |.     .     .     .     .
                        s 0.75 +.     .     .     .     .
                        t     |.     .     .     .     .
                        e  0.5 +.     .     .     .     .
                        r     |.     .     .     .     .
                        s 0.25 +.     .     .     .     .
                              |.     .     .     .     .
                            0 +.     .     .     .     .
```

Como se ve, al que se denominó *cluster* 1 (individuos 1 y 2) en la resolución manual del ejemplo, el SAS le ha denominado *cluster* 4, y al que se le denominó *cluster* 2 (individuos 3 y 4) en la resolución manual del ejemplo, el SAS le ha denominado *cluster* 3. El *cluster* 2 lo constituyen el cluster 3 (individuo 3 y 4) con el individuo 5, y el *cluster* 1 lo constituye el cluster 4 (individuos 1 y 2) y el cluster 2 (indviduos 3, 4 y 5).

Método del centroide.-

Como la distancia más pequeña tiene el valor 1 y es la que separa al individuo uno del dos, se hace un grupo (cluster 1) con estos dos individuos. Este cluster tiene el valor X_1=1 y X_2=1.5, por lo que la distancias entre este cluster y los idividuos tres, cuatro y cinco son, respectivamente, 4.924, 6.80 y 7.826, por lo que esas son las que se toman para la nueva matriz de distancias

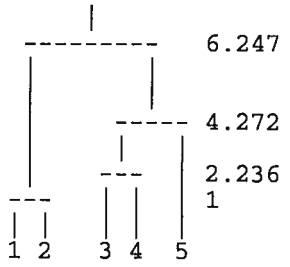
	1,2	3	4	5
1,2	0	4.924	6.80	7.826
3		0	2.236	5.099
4			0	3.605
5				0

la distancia más pequeña tiene el valor 2.236 y es la que separa al individuo tres del cuatro, por lo que se hace un grupo (cluster 2) con estos dos individuos. Este nuevo cluster tiene el valor X_1=4 y X_2=6.5, por lo que la distancia entre este cluster y el cluster 1 y el individuo cinco son, respectivamente, 5.831 y 4.272, por lo que esas son las que se toman para la nueva matriz de distancias. Y así tambien para la última matriz

	1,2	3,4	5
1,2	0	5.831	7.826
3,4		0	4.272
5			0

	1,2	3,4,5
1,2	0	6.247
3,4,5		0

Con estos resultados se puede representar el siguiente dendograma



El paquete estadístico *SAS* realiza esto mismo con el siguiente programa

Archivo de programa *SAS* (C19-5.SAS).-

```
title 'Cluster';
options ls=75 ps=60;
data cluster;
infile 'c19-4.dat';
```

```
input indi x1 x2;
proc cluster method=cen nonorm;
id indi;
run;
proc tree;
id indi;
run;
```

Archivo de resultados (C19-5.LST).-

```
                Centroid Hierarchical Cluster Analysis

                 Eigenvalues of the Covariance Matrix
          Eigenvalue      Difference      Proportion      Cumulative
1          12.9570         10.4139        0.835933         0.83593
2           2.5430             .          0.164067         1.00000

     Root-Mean-Square Total-Sample Standard Deviation = 2.783882
  Number                                        Frequency
    of                                            of New    Centroid
 Clusters    ------Clusters Joined------         Cluster    Distance    Tie
       4              1                 2            2     1.00000
       3              3                 4            2     2.23607
       2    CL3                         5            3     4.27200
       1    CL4             CL2                      5     6.24722

                Centroid Hierarchical Cluster Analysis
                                INDI
                      1     2     3     4     5
                  D   7 +
                  i     |
                  s     |
                  t     |XXXXXXXXXXXXXXXXXXXXXXXXX
                  a     |XXXXXXX     XXXXXXXXXXXXX
                  n   6 +XXXXXXX     XXXXXXXXXXXXX
                  c     |XXXXXXX     XXXXXXXXXXXXX
                  e     |XXXXXXX     XXXXXXXXXXXXX
                        |XXXXXXX     XXXXXXXXXXXXX
                  B     |XXXXXXX     XXXXXXXXXXXXX
                  e   5 +XXXXXXX     XXXXXXXXXXXXX
                  t     |XXXXXXX     XXXXXXXXXXXXX
                  w     |XXXXXXX     XXXXXXXXXXXXX
                  e     |XXXXXXX     XXXXXXXXXXXXX
                  e     |XXXXXXX     XXXXXXX     .
                  n   4 +XXXXXXX     XXXXXXX     .
                        |XXXXXXX     XXXXXXX     .
                  C     |XXXXXXX     XXXXXXX     .
                  l     |XXXXXXX     XXXXXXX     .
                  u     |XXXXXXX     XXXXXXX     .
                  s   3 +XXXXXXX     XXXXXXX     .
                  t     |XXXXXXX     XXXXXXX     .
                  e     |XXXXXXX     XXXXXXX     .
                  r     |XXXXXXX     XXXXXXX     .
                        |XXXXXXX     .     .     .
                  C   2 +XXXXXXX     .     .     .
                  e     |XXXXXXX     .     .     .
                  n     |XXXXXXX     .     .     .
                  t     |XXXXXXX     .     .     .
                  r     |XXXXXXX     .     .     .
                  o   1 +XXXXXXX     .     .     .
```

El procedimiento **CLUSTER** hace el algoritmo y el procedimiento **TREE** realiza el dendograma. La opción **NONORM** es para que nos presente las distancias no normalizadas que son las que se han utilizado para realizar el ejemplo a mano.

Como se ve en el archivo de resultados, al que se denominó *cluster* 1

(individuos 1 y 2) en la resolución manual del ejemplo, el SAS le ha denominado *cluster* 4, y al que se le denominó *cluster* 2 (individuos 3 y 4) en la resolución manual del ejemplo, el SAS le ha denominado *cluster* 3. El *cluster* 2 lo constituyen el cluster 3 (individuo 3 y 4) con el individuo 5, y el *cluster* 1 lo constituye el cluster 4 (individuos 1 y 2) y el cluster 2 (indviduos 3, 4 y 5).

Ventajas e inconvenientes de los procedimientos jerárquicos.-

Los procedimientos jerárquicos tienen la ventaja de ser más rápidos que, por ejemplo, examinar todas las posibles combinaciones de observaciones, por lo que son muy útiles para utilizar en aplicaciones taxonómicas. Pero estos métodos pueden producir errores en ciertas situaciones como consecuencia de que, por ejemplo, una combinación indeseada que se produzca al principio del proceso persiste hasta el final del análisis lo que puede producir resultados artificiales. Se debe de realizar varias veces el análisis para ir eliminando o corrigiendo observaciones sospechosamente erróneas.

Otro inconveniente es que para tamaños de muestra grande los dendogramas pueden hacerse inmanejables y de difícil o imposible lectura.

Un problema importante es el de seleccionar un número de cluster pues no existe un procedimiento establecido para este fin. Puede servir de guía las distancias entre cluster en sucesivos pasos. Se puede parar cuando las distancias excedan un valor especifico o cuando las sucesivas diferencias de las distancias entre pasos den un brusco salto. También situaciones subyacentes pueden sugerir un número de cluster. Si se conoce previamente ese número son muy apropiados los métodos no jerárquicos.

Ejemplo.-

Volvamos a utilizar parte de los datos de *Fisher* del género *Iris*. Supóngase que se tiene una muestra de 30 individuos que no pueden ser asignados a grupo alguno por no existir criterio previo para dicha asignación. Los datos son los del archivo C19-6.DAT

Archivo de programa *SAS* (C19-6.SAS).-

```
title 'Clasificaci¢n sin criterio previo';
options ls=75 ps=60;
data cluster;
infile 'c19-6.dat';
input longsep anchsep longpet anchpet variedad $ simb $ @@;
proc cluster method=cen;
id simb ;
run;
proc tree inc=.1 tickpos=3 space=1;
id simb ;
run;
```

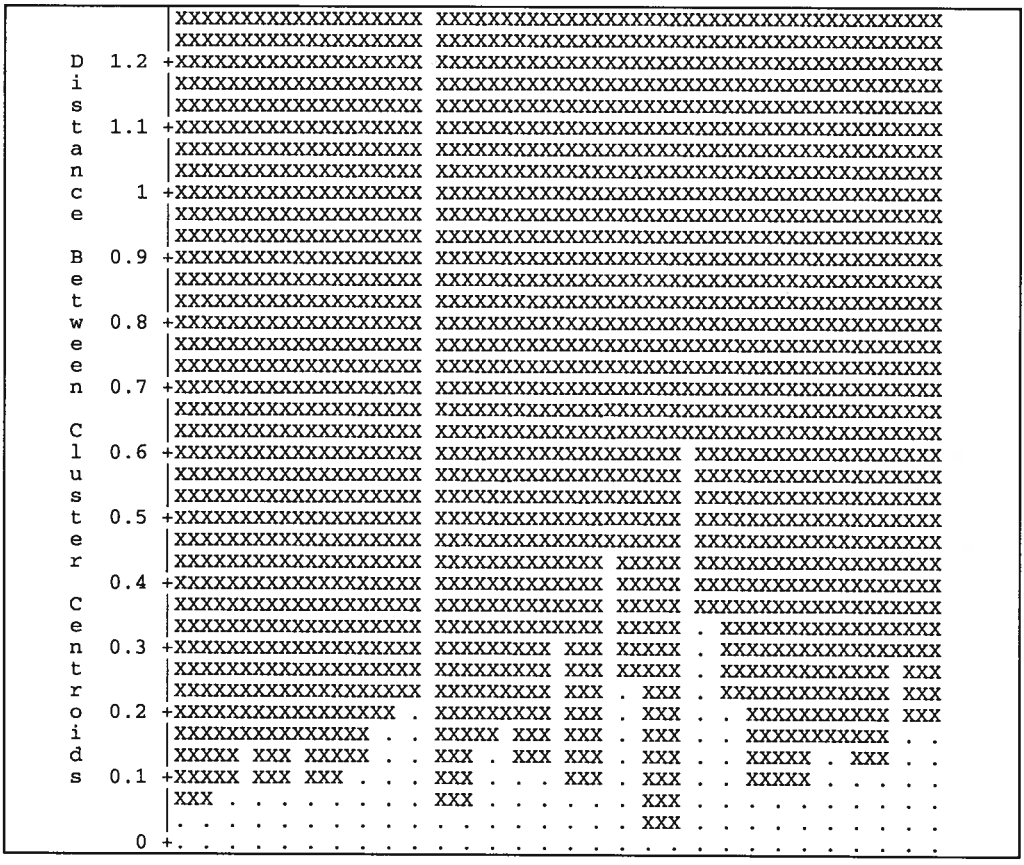
Archivo de datos C19-6.DAT.-

```
5.0 3.3 1.4 0.2 setosa      1s   5.8 2.7 4.1 1.0 versicolor  6v
4.6 3.4 1.4 0.3 setosa      2s   6.0 2.9 4.5 1.5 versicolor  7v
```

4.6	3.6	1.0	0.2	setosa	3s	5.7	2.6	3.5	1.0	versicolor	8v
5.1	3.3	1.7	0.5	setosa	4s	4.9	2.4	3.3	1.0	versicolor	9v
5.5	3.5	1.3	0.2	setosa	5s	6.6	2.9	4.6	1.3	versicolor	10v
4.8	3.1	1.6	0.2	setosa	6s	7.3	2.9	6.3	1.8	virginica	1a
5.2	3.4	1.4	0.2	setosa	7s	6.7	2.5	5.8	1.8	virginica	2a
4.9	3.6	1.4	0.1	setosa	8s	6.5	3.0	5.8	2.2	virginica	3a
4.4	3.2	1.3	0.2	setosa	9s	6.9	3.1	5.4	2.1	virginica	4a
5.0	3.5	1.6	0.6	setosa	10s	6.7	3.1	5.6	2.4	virginica	5a
5.8	2.6	4.0	1.2	versicolor	1v	6.3	2.8	5.1	1.5	virginica	6a
5.5	2.6	4.4	1.2	versicolor	2v	6.5	3.0	5.2	2.0	virginica	7a
5.0	2.3	3.3	1.0	versicolor	3v	6.5	3.0	5.5	1.8	virginica	8a
6.7	3.1	4.4	1.4	versicolor	4v	5.8	2.7	5.1	1.9	virginica	9a
5.6	3.0	4.5	1.5	versicolor	5v	6.8	3.2	5.9	2.3	virginica	10ª

Archivo de salida C19-6.LST.-

```
                    Centroid Hierarchical Cluster Analysis

                     Eigenvalues of the Covariance Matrix
              Eigenvalue       Difference       Proportion       Cumulative
   1             4.31514          4.14954         0.947282          0.94728
   2             0.16561          0.10697         0.036355          0.98364
   3             0.05864          0.04274         0.012873          0.99651
   4             0.01590           .              0.003490          1.00000

        Root-Mean-Square Total-Sample Standard Deviation = 1.067156
        Root-Mean-Square Distance Between Observations   = 3.018373

        Number                               Frequency   Normalized
          of                                  of New      Centroid
       Clusters    --Clusters Joined--        Cluster     Distance     Tie
            29     9v          3v                   2     0.046854
            28     1s          7s                   2     0.074082
            27     6v          1v                   2     0.081153
            26     4s          10s                  2     0.087655
            25     2s          9s                   2     0.104768      T
            24     10v         4v                   2     0.104768
            23     CL28        8s                   3     0.111123
            22     5a          10a                  2     0.114767
            21     3a          CL22                 3     0.109881
            20     7a          8a                   2     0.119453
            19     7v          5v                   2     0.136600
            18     CL25        6s                   3     0.146300
            17     CL23        CL26                 5     0.152124
            16     4a          CL20                 3     0.152724
            15     CL27        2v                   3     0.157151
            14     CL17        CL18                 8     0.162441
            13     CL21        CL16                 6     0.174263
            12     CL15        CL19                 5     0.197629
            11     CL14        3s                   9     0.201907
            10     6a          9a                   2     0.214710
             9     CL11        5s                  10     0.226084
             8     2a          CL13                 7     0.231716
             7     8v          CL29                 3     0.270171
             6     CL8         CL10                 9     0.300550
             5     CL12        CL24                 7     0.319601
             4     1a          CL6                 10     0.378319
             3     CL5         CL7                 10     0.452719
             2     CL3         CL4                 20     0.630959
             1     CL9         CL2                 30     1.295372


                    Centroid Hierarchical Cluster Analysis
                                     SIMB
                  1                     1                 1
            1 7 8 4 0 2 9 6 3 5 6 1 2 7 5 0 4 8 9 3 1 2 3 5 0 4 7 8 6 9
            s s s s s s s s s s v v v v v v v v v v a a a a a a a a a a
        1.3 +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```



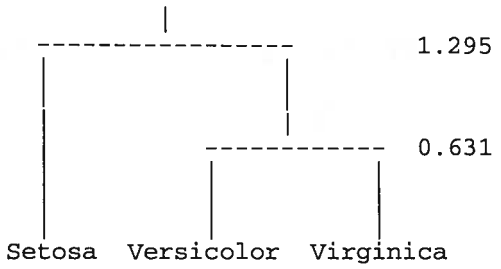
En la salida del procedimiento **CLUSTER** se observa:

. Hay solo una coincidencia de valores (**TIE**) entre los individuos 2 y 9 de las setosas (con objeto de contrastar se hará la discusión haciendo referencia a la *variedad*, que en estos tipos de análisis nunca se conoce a priori).

. El cluster 1, del que cuelgan todos los individuos, se subdivide en dos cluster, el 2 formado por los individuos *virginica* y *versicolor*, y el 9 formado solo por individuos *setosa*. Esto se produce a una distancia normalizada entre los centroides de 1.295. Es decir, que en la raíz de la jerarquía hay dos aglomeraciones claramente diferenciadas, por un lado las setosas y por el otro lado las virginica y versicolor.

. A una distancia de 0.631 (la mitad de la distancia anterior) se separan las dos ramas del cluster 2, constituyendo el cluster 3 que son los individuos versicolor, y el cluster 4 que son los individuos virginica.

Esto se observa gráficamente en el dendograma de la salida del procedimiento **TREE**. Dada la pequeña magnitud de las demás distancias, este dendograma se podría simplificar de la siguiente manera:



Clasificación no jerárquica.-

Estos métodos de clasificación se les conoce también como agrupamientos disgregativos (*disjoint*). Su objetivo es realizar una partición, en dos o tres pasos, de los individuos de *K* grupos. Por lo que el investigador tiene que tener claro cuantos grupos desea o, lo que en cierta manera es equivalente, saber cual es el radio mínimo de los cluster. La asignación de los individuos a los grupos se hace mediante un proceso que optimiza el criterio de selección. Estos métodos permite clasificar simultáneamente a los individuos y a las variables, siendo sus salidas el número de cluster así como las distancia entre los centroides de los cluster.

Método de las *K-medias*.-

Este método es el más utilizado de los métodos no jerárquicos. Especificados el número *K* de cluster el algoritmo procede de la siguiente manera.

1. Divide los datos en *K* cluster iniciales. Los miembros de estos cluster pueden ser especificados por el investigador o por el programa
2. Calcula las medias o centroides de cada uno de los *K* cluster.
3. Para todos y cada uno de los individuos calcula la distancia a cada centroide. Si el individuo esta próximo al centroide del cluster al que está asignado, permanece en el, en caso contrario es reasignado al cluster que ha dado una menor medida.
4. Repite los pasos 2 y 3 hasta que no se reasigna ningún individuo.

Ejemplo.-

Realicemos este algoritmo con los datos del ejemplo hipotético para formar dos grupos. En primer lugar se puede hacer dos grupo con los dos individuos mas alejados, estos son el individuo uno (cluster 1) y el cinco (cluster 2). Se toma el individuo más próximo al cluster 1, este es el individuo dos y se mide sus distancias a ambos clustes, estas son 1 y 7.62, respectivamente, por lo que se asigna al cluster 1. Se toma el individuo más próximo al cluster 2, este es el individuo cuatro y se mide sus distancias a ambos cluster, estas son 6.40 y 3.61, respectivamente, por lo que se asigna al cluster 2. Se toma el individuo tres y se mide sus distancias a ambos clustes, estas son, 4.92 y 3.50, respectivamente, por lo que se asigna al cluster 2. Ya están asignados todos los individuos a los dos cluster cuyos centroides valen (1, 1.5) y (5.33, 6), respectivamente. Ahora se mide la distancia de todos los individuos a cada uno de los centroides y si no se reasignan ninguno, como así ocurre, se finaliza el proceso.

Este método es muy útil cuando se tienen muestra grandes. El procedimiento del *SAS* para realizar este método es el **FASTCLUS**. El programa puede ser el siguiente

Archivo de programa *SAS* (C19-7.SAS).-

```
Title 'Clasificación no jerárquica';
options ls=75 ps=60;
data cluster;
infile 'c19-4.dat';
input indi x1 x2;
proc fastclus maxc=2 otu=cluster;
var x1 x2;
id indi;
run;
proc print data=cluster;
run;
```

La opción **MAXC=2** es para indicarle que haga un máximo de dos grupos.

Archivo de resultados (C19-7.LST).-

```
     FASTCLUS Procedure:  Replace=FULL  Radius=0  Maxclusters=2  Maxiter=1

                              Initial Seeds
                      Cluster            X1            X2
                      ------------------------------------
                         1          1.00000       1.00000
                         2          8.00000       5.00000

                    Criterion Based on Final Seeds = 1.2315

                              Cluster Summary
                                Maximum Distance
                      RMS Std     from Seed      Nearest   Distance Between
Cluster  Frequency  Deviation  to Observation   Cluster  Cluster Centroids
-------------------------------------------------------------------------
   1           2      0.5000      0.5000            2             6.2472
   2           3      1.9149      2.8480            1             6.2472

                          Statistics for Variables
Variable       Total STD      Within STD       R-Squared       RSQ/(1-RSQ)
-------------------------------------------------------------------------
X1              2.966479        2.054805        0.640152          1.778947
X2              2.588436        0.912871        0.906716          9.720000
OVER-ALL        2.783882        1.589899        0.755376          3.087912

                                Pseudo F Statistic =     9.26
                  Approximate Expected Over-All R-Squared =    .
                               Cubic Clustering Criterion =    .
      WARNING: The two above values are invalid for correlated variables.

                              Cluster Means
                      Cluster            X1            X2
                      ------------------------------------
                         1          1.00000       1.50000
                         2          5.33333       6.00000

                          Cluster Standard Deviations
                      Cluster            X1            X2
                      ------------------------------------
                         1          0.00000       0.70711
                         2          2.51661       1.00000


                OBS     INDI      X1      X2     CLUSTER     DISTANCE
                 1        1       1       1         1         0.50000
                 2        2       1       2         1         0.50000
                 3        3       3       6         2         2.33333
                 4        4       5       7         2         1.05409
                 5        5       8       5         2         2.84800
```

Ejemplo.-

Utilicemos de nuevo datos completos de *Fisher*, suponiéndose, como antes, que es una muestra de 150 individuos que no pueden ser asignados a grupo alguno por no existir criterio previo para dicha asignación. Los datos son los mismos de ejemplos anteriores (**C19-1.DAT**)

Archivo de programa *SAS* (C19-8.SAS).-

```
Title 'Clasificación no jerárquica';
options ls=80 ps=60;
data kmedias;
infile 'c19-1.dat';
input longsep anchsep longpet anchpet variedad $ simb $ @@;
proc fastclus data=kmedias maxc=2 out=cluster;
var longsep anchsep longpet anchpet;
id variedad;
run;
proc freq;
tables cluster*variedad;
run;
proc fastclus data=kmedias maxc=3 out=cluster;
var longsep anchsep longpet anchpet;
id variedad;
run;
proc freq;
tables cluster*variedad;
run;
proc fastclus data=kmedias maxc=4 out=cluster;
var longsep anchsep longpet anchpet;
id variedad;
run;
proc freq;
tables cluster*variedad;
run;
```

Archivo de resultados (C19-8.LST).-

```
     FASTCLUS Procedure:  Replace=FULL  Radius=0  Maxclusters=2  Maxiter=1

                              Initial Seeds
        Cluster       LONGSEP       ANCHSEP       LONGPET       ANCHPET
        ------------------------------------------------------------------
           1          7.70000       2.60000       6.90000       2.30000
           2          4.30000       3.00000       1.10000       0.10000

                  Criterion Based on Final Seeds = 0.5378

                              Cluster Summary
                              Maximum Distance
                    RMS Std       from Seed      Nearest    Distance Between
Cluster  Frequency  Deviation    to Observation  Cluster  Cluster Centroids
-------------------------------------------------------------------------
   1          93      0.5459      2.2414            2              3.8405
   2          57      0.4746      1.9347            1              3.8405

                        Statistics for Variables
Variable     Total STD      Within STD      R-Squared      RSQ/(1-RSQ)
-------------------------------------------------------------------
LONGSEP       0.828066       0.535061       0.585283         1.411285
```

ANCHSEP	0.435866	0.390560	0.202476	0.253880
LONGPET	1.765298	0.714752	0.837164	5.141160
ANCHPET	0.762238	0.363578	0.774009	3.424964
OVER-ALL	1.069224	0.520069	0.765004	3.255398

Pseudo F Statistic = 481.80
Approximate Expected Over-All R-Squared = 0.51539
Cubic Clustering Criterion = 13.854
WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	6.33763	2.90430	5.01828	1.72258
2	5.03684	3.30702	1.70175	0.34561

Cluster Standard Deviations

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	0.618890	0.314820	0.740145	0.402473
2	0.356887	0.490211	0.670951	0.288512

TABLE OF CLUSTER BY VARIEDAD

CLUSTER(Cluster) Frequency Percent Row Pct Col Pct	setosa	versicol	virginic	Total
1	0 0.00 0.00 0.00	43 28.67 46.24 86.00	50 33.33 53.76 100.00	93 62.00
2	50 33.33 87.72 100.00	7 4.67 12.28 14.00	0 0.00 0.00 0.00	57 38.00
Total	50 33.33	50 33.33	50 33.33	150 100.00

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=3 Maxiter=1

Initial Seeds

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	4.90000	2.50000	4.50000	1.70000
2	5.50000	4.20000	1.40000	0.20000
3	7.70000	3.80000	6.70000	2.20000

Criterion Based on Final Seeds = 0.37097

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	67	0.4180	1.8532	3	1.8341
2	50	0.2780	1.2480	1	3.4252
3	33	0.3883	1.2923	1	1.8341

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
LONGSEP	0.828066	0.448242	0.710915	2.459187

ANCHSEP	0.435866	0.324819	0.452092	0.825123
LONGPET	1.765298	0.429764	0.941527	16.101961
ANCHPET	0.762238	0.238707	0.903243	9.335201
OVER-ALL	1.069224	0.370171	0.881751	7.456709

Pseudo F Statistic = 548.07
Approximate Expected Over-All R-Squared = 0.62728
Cubic Clustering Criterion = 24.559
WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	5.94776	2.76119	4.45224	1.45373
2	5.00600	3.42800	1.46200	0.24600
3	6.90000	3.09697	5.82727	2.12727

Cluster Standard Deviations

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	0.483158	0.295397	0.536080	0.301174
2	0.352490	0.379064	0.173664	0.105386
3	0.501248	0.290995	0.457761	0.240147

TABLE OF CLUSTER BY VARIEDAD

CLUSTER(Cluster) VARIEDAD

Frequency / Percent / Row Pct / Col Pct	setosa	versicol	virginic	Total
1	0	50	17	67
	0.00	33.33	11.33	44.67
	0.00	74.63	25.37	
	0.00	100.00	34.00	
2	50	0	0	50
	33.33	0.00	0.00	33.33
	100.00	0.00	0.00	
	100.00	0.00	0.00	
3	0	0	33	33
	0.00	0.00	22.00	22.00
	0.00	0.00	100.00	
	0.00	0.00	66.00	
Total	50	50	50	150
	33.33	33.33	33.33	100.00

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=4 Maxiter=1

Initial Seeds

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	4.90000	2.40000	3.30000	1.00000
2	5.50000	4.20000	1.40000	0.20000
3	7.00000	3.20000	4.70000	1.40000
4	7.70000	2.60000	6.90000	2.30000

Criterion Based on Final Seeds = 0.31521

Cluster Summary

Maximum Distance

Cluster	Frequency	RMS Std Deviation	from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids

1	28	0.3005	0.9914	3	1.4412
2	50	0.2780	1.2480	1	2.8484
3	55	0.3430	0.9971	1	1.4412
4	17	0.3653	1.1592	3	1.5370

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
LONGSEP	0.828066	0.352174	0.822764	4.642196
ANCHSEP	0.435866	0.319822	0.472436	0.895504
LONGPET	1.765298	0.341198	0.963395	26.318470
ANCHPET	0.762238	0.245647	0.898233	8.826343
OVER-ALL	1.069224	0.317439	0.913632	10.578411

Pseudo F Statistic = 514.82
Approximate Expected Over-All R-Squared = 0.69068
Cubic Clustering Criterion = 29.783
WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	5.53214	2.63571	3.96071	1.22857
2	5.00600	3.42800	1.46200	0.24600
3	6.33091	2.90182	5.00182	1.76182
4	7.24118	3.16471	6.15294	2.13529

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=4 Maxiter=1

Cluster Standard Deviations

Cluster	LONGSEP	ANCHSEP	LONGPET	ANCHPET
1	0.318624	0.269725	0.390004	0.186304
2	0.352490	0.379064	0.173664	0.105386
3	0.334372	0.275876	0.409377	0.339112
4	0.450082	0.337159	0.384249	0.264436

TABLE OF CLUSTER BY VARIEDAD

CLUSTER(Cluster) Frequency Percent Row Pct Col Pct	setosa	versicol	virginic	Total
1	0 0.00 0.00 0.00	27 18.00 96.43 54.00	1 0.67 3.57 2.00	28 18.67
2	50 33.33 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	50 33.33
3	0 0.00 0.00 0.00	23 15.33 41.82 46.00	32 21.33 58.18 64.00	55 36.67
4	0 0.00 0.00 0.00	0 0.00 0.00 0.00	17 11.33 100.00 34.00	17 11.33
Total	50 33.33	50 33.33	50 33.33	150 100.00

Como se ve, se realizar tres análisis no jerárquicos para agrupar los individuos en

MAXC=2, **MAXC=3** y **MAXC=4** cluster, respectivamente.

Archivo se salida C19-8.LST.-

En la salida de los tres procedimientos **FASTCLUS** da, entre otros valores de interés, el valor de los centroides, las distancias entre ellos, las desviaciones típicas de los cluster y el número de individuos de cada cluster. Si se quisiera saber cuales son los individuos de cada cluster se pone detrás de cada procedimiento **fastclus** un procedimiento **PRINT DATA=CLUSTER**. En este caso se ha puesto un procedimiento **FREQ** con **TABLES CLUSTER*VARIEDAD** con objeto de ver si mezcla individuos de diferentes variedad en el mismo cluster.

Si se observa la salida del primer procedimiento *freq* que recoge los resultados del análisis que agrupa a los individuos en dos cluster, se observa que en el primer cluster están todos los individuos setosa más 7 individuos versicolor, y en el segundo cluster están todos los individuos virginica y la mayoría (43) de los individuos versicolor.

Si se observa la salida del segundo procedimiento *freq* que recoje los resultados del análisis que agrupa a los individuos en tres cluster, se observa que en el primer cluster están solamente los 50 individuos setosa, en el segundo cluster están todos los individuos versicolor y el 30% (15) de los individuos virginica, y en el tercer cluster están el 70% (35) restante de los individuos versicolor.

Si se observa la salida del tercer procedimiento *freq* que recoge los resultados del análisis que agrupa a los individuos en cuatro cluster, se observa que en el primer cluster están solamente los 50 individuos setosa; en el segundo cluster solamente hay individuos virginica (17, el 34%); en el tercer cluster esta formado prácticamente solo por la mitad de los individuos versicolor (26 versicolor y 1 virginica); y en el tercer cluster está formado por la otra mitad de los individuos versicolor (24) y el 64% (32) restante de los individuos virginica.

La discusión de estos resultados es eminentemente empírica, esto es, está muy asociada a la finalidad planteada en el estudio concreto, por lo que no se puede hacer una discusión generalizada.

Métodos no paramétricos.-

Como se dijo en la introducción del Análisis Discriminante, todos estos análisis se sustentan en el supuesto de que la distribución dentro de los grupos es normal, por lo que se usan métodos paramétricos. Si no se puede realizar dicho supuesto se pueden usar métodos no paramétricos en la función discriminante (**PROC DISCRIM**), pero no en el análisis canónico discriminante ni en el análisis cluster.

Los métodos no paramétricos de la función discriminante incluyen los métodos de *Kernel* y el de *k-vecinos-cercanos*. En el método de los k-vecinos-cercanos para calcular las distancias de Mahalanobis se usa la matriz de covarianza general, mientras que en método de kernel se puede usar tanto las matrices de covarianzas dentro de grupo como la matriz de covarianza general

Ejemplo.-

Analicemos de nuevo los datos completos de *Fisher* que trata de la clasificación de tres especies del género *Iris*, como si estos datos fueran no normales (que si lo son)

Archivo de programa *SAS* (C19-9.SAS).-

```
title 'Métodos no paramétricos';
Options ls =80 ps=60;
data ad;
infile 'c19-1.dat';
input longsep anchsep longpet anchpet variedad $ simb $ @@;
proc discrim outd=outd method=npar kernel=epa pool=yes r=1.8
             listerr crosslisterr posterr;

class variedad;
var longsep anchsep longpet anchpet;
run;
proc print data=outd;
run;
```

Como se ve, las diferencias con el ejemplo C19-1 son:

Se ha expresado que se use un método no paramétrico (**METHOD=NPAR**), por defecto, se realiza el método paramétrico.

Que use un método de Kernel con un diámetro igual para cada clase (**POOL=YES**) con valor de 1.8 (**R=1.8**).

La opción **LISTERR** da el listado de las observaciones que han sido mal clasificadas mediante el método empírico. Y la opción **CROSSLISTERR** da el listado de las observaciones que han sido mal clasificadas mediante una validación cruzada.

Archivo de resultados (C19-9.LST).-

```
                          Discriminant Analysis
                  150 Observations          149 DF Total
                    4 Variables             147 DF Within Classes
                    3 Classes                 2 DF Between Classes

                               Class Level Information
              Output                                                    Prior
VARIEDAD      SAS Name   Frequency      Weight      Proportion      Probability
setosa        SETOSA            50     50.0000        0.333333        0.333333
versicol      VERSICOL          50     50.0000        0.333333        0.333333
virginic      VIRGINIC          50     50.0000        0.333333        0.333333

              Posterior Probability of Membership in VARIEDAD:
Obs      From          Classified
         VARIEDAD      into VARIEDAD      setosa      versicol      virginic

 13      virginic      versico1 *          0.0000        0.5931        0.4069
 25      versicol      virginic *          0.0000        0.4492        0.5508
 34      versicol      virginic *          0.0000        0.2807        0.7193

                          * Misclassified observation

 Discriminant Analysis      Classification Summary for Calibration Data:
WORK.AD
```

Resubstitution Summary using Epanechnikov Kernel Density
Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each VARIEDAD:

$$F(X|j) = n_j^{-1} \sum_i \left(1.0 - D^2_{ji}(X,Y_i) / R^2 \right)$$

$$Pr(j|X) = PRIOR_j \, F(X|j) \Big/ \sum_k PRIOR_k \, F(X|k)$$

Number of Observations and Percent Classified into VARIEDAD:

From VARIEDAD	setosa	versicol	virginic	Total
setosa	50	0	0	50
	100.00	0.00	0.00	100.00
versicol	0	48	2	50
	0.00	96.00	4.00	100.00
virginic	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
Percent	33.33	32.67	34.00	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for VARIEDAD:

	setosa	versicol	virginic	Total
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

Discriminant Analysis Classification Results for Calibration Data: WORK.AD

Resubstitution Results using Epanechnikov Kernel Density
Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each VARIEDAD:

$$F(X|j) = n_j^{-1} \sum_i \left(1.0 - D^2_{ji}(X,Y_i) / R^2 \right)$$

$$Pr(j|X) = PRIOR_j \, F(X|j) \Big/ \sum_k PRIOR_k \, F(X|k)$$

Number of Observations and Average Posterior Probabilities
Classified into VARIEDAD:

From VARIEDAD	setosa	versicol	virginic
setosa	50	0	0
	1.0000	.	.
versicol	0	48	2
	.	0.9832	0.6351
virginic	0	1	49
	.	0.5931	0.9723
Total	50	49	51
	1.0000	0.9753	0.9591
Priors	0.3333	0.3333	0.3333

Posterior Probability Error Rate Estimates for VARIEDAD:

Estimate	setosa	versicol	virginic	Total
Stratified	0.0000	0.0442	0.0218	0.0220
Unstratified	0.0000	0.0442	0.0218	0.0220
Priors	0.3333	0.3333	0.3333	

Posterior Probability of Membership in VARIEDAD:

Obs	From VARIEDAD	Classified into VARIEDAD		setosa	versicol	virginic
13	virginic	versicol	*	0.0000	0.8668	0.1332
20	virginic	OTHER	P	.	.	.
25	versicol	virginic	*	0.0000	0.2276	0.7724
34	versicol	virginic	*	0.0000	0.0000	1.0000

Discriminant Analysis Classification Results for Calibration Data: WORK.AD

Cross-validation Results using Epanechnikov Kernel Density

Posterior Probability of Membership in VARIEDAD:

Obs	From VARIEDAD	Classified into VARIEDAD		setosa	versicol	virginic
54	versicol	virginic	*	0.0000	0.2116	0.7884
71	virginic	OTHER	P	.	.	.
73	virginic	OTHER	P	.	.	.
111	setosa	OTHER	P	.	.	.
119	virginic	OTHER	P	.	.	.
122	virginic	versicol	*	0.0000	1.0000	0.0000
141	virginic	OTHER	P	.	.	.

Discriminant Analysis Classification Summary for Calibration Data: WORK.AD

Cross-validation Summary using Epanechnikov Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each VARIEDAD:

$$F(X|j) = n_j^{-1} \, SUM_i \, (1.0 - D^2(X,Y_{ji}) / R^2)$$

$$Pr(j|X) = PRIOR_j \, F(X|j) \, / \, SUM_k \, PRIOR_k \, F(X|k)$$

Number of Observations and Percent Classified into VARIEDAD:

From VARIEDAD	setosa	versicol	virginic	OTHER	Total
setosa	49	0	0	1	50
	98.00	0.00	0.00	2.00	100.00
versicol	0	47	3	0	50
	0.00	94.00	6.00	0.00	100.00
virginic	0	2	43	5	50
	0.00	4.00	86.00	10.00	100.00
Total	49	49	46	6	150
Percent	32.67	32.67	30.67	4.00	100.00
Priors	0.3333	0.3333	0.3333		

Error Count Estimates for VARIEDAD:

	setosa	versicol	virginic	Total
Rate	0.0200	0.0600	0.1400	0.0733
Priors	0.3333	0.3333	0.3333	

Discriminant Analysis Clasification Results for Calibration Data: WORK.AD

Cross-validation Results using Epanechnikov Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each VARIEDAD:

$$F(X|j) = n_j^{-1} \sum_i \left(1.0 - D_{ji}^2(X,Y) \right) / R^2$$

$$Pr(j|X) = PRIOR_j F(X|j) / \sum_k PRIOR_k F(X|k)$$

Number of Observations and Average Posterior Probabilities
Classified into VARIEDAD:

From VARIEDAD	setosa	versicol	virginic
setosa	49	0	0
	1.0000	.	.
versicol	0	47	3
	.	0.9894	0.8536
virginic	0	2	43
	.	0.9334	0.9692
Total	49	49	46
	1.0000	0.9871	0.9617
Priors	0.3333	0.3333	0.3333

Posterior Probability Error Rate Estimates for VARIEDAD:

Estimate	setosa	versicol	virginic	Total
Stratified	0.0200	0.0326	0.1153	0.0560
Unstratified	0.0200	0.0326	0.1153	0.0560
Priors	0.3333	0.3333	0.3333	

OBS	LONGSEP	ANCHSEP	LONGPET	ANCHPET	VARIEDAD	SIMB	SETOSA	VERSICOL	VIRGINIC
1	5.0	3.3	1.4	0.2	setosa	s	3.30936	0.00000	0.00000
2	4.8	3.0	1.4	0.3	setosa	s	1.60360	0.00000	0.00000
3	4.7	3.2	1.3	0.2	setosa	s	3.28953	0.00000	0.00000
4	6.4	2.8	5.6	2.2	virginic	a	0.00000	0.00000	0.77344
5	5.1	3.8	1.6	0.2	setosa	s	1.28139	0.00000	0.00000
6	4.6	3.1	1.5	0.2	setosa	s	2.32986	0.00000	0.00000
7	6.5	2.8	4.6	1.5	versicol	v	0.00000	1.06444	0.00000
8	6.1	3.0	4.9	1.8	virginic	a	0.00000	0.15286	0.78044
9	6.9	3.2	5.7	2.3	virginic	a	0.00000	0.00000	1.07558
10	6.7	3.1	5.6	2.4	virginic	a	0.00000	0.00000	0.74182
11	4.8	3.4	1.9	0.2	setosa	s	0.45556	0.00000	0.00000
12	6.2	2.9	4.3	1.3	versicol	v	0.00000	1.90294	0.00000
13	6.3	2.8	5.1	1.5	virginic	a	0.00000	0.31434	0.21569
14	5.0	3.0	1.6	0.2	setosa	s	1.86169	0.00000	0.00000
15	7.4	2.8	6.1	1.9	virginic	a	0.00000	0.00000	0.39137
16	4.6	3.4	1.4	0.3	setosa	s	2.00582	0.00000	0.00000
17	5.0	3.2	1.2	0.2	setosa	s	1.66888	0.00000	0.00000
18	5.9	3.0	4.2	1.5	versicol	v	0.00000	1.13816	0.00000
19	6.9	3.1	5.1	2.3	virginic	a	0.00000	0.00000	0.29438
20	6.1	2.6	5.6	1.4	virginic	a	0.00000	0.00000	0.16837
21	5.1	3.4	1.5	0.2	setosa	s	3.32625	0.00000	0.00000
22	6.2	2.2	4.5	1.5	versicol	v	0.00000	0.24571	0.00000
23	6.4	2.8	5.6	2.1	virginic	a	0.00000	0.00000	0.94910
24	5.0	3.5	1.3	0.3	setosa	s	2.46471	0.00000	0.00000
25	5.9	3.2	4.8	1.8	versicol	v	0.00000	0.26071	0.31972
26	4.3	3.0	1.1	0.1	setosa	s	1.81346	0.00000	0.00000
27	5.6	2.8	4.9	2.0	virginic	a	0.00000	0.00000	0.50077
28	4.6	3.6	1.0	0.2	setosa	s	0.93324	0.00000	0.00000
29	5.8	4.0	1.2	0.2	setosa	s	0.17122	0.00000	0.00000
30	6.0	2.2	4.0	1.0	versicol	v	0.00000	0.25254	0.00000
31	6.1	3.0	4.6	1.4	versicol	v	0.00000	1.51892	0.02451
32	5.1	3.8	1.9	0.4	setosa	s	0.59122	0.00000	0.00000
33	7.3	2.9	6.3	1.8	virginic	a	0.00000	0.00000	0.37628
34	6.0	2.7	5.1	1.6	versicol	v	0.00000	0.16837	0.43152
35	6.7	3.1	4.4	1.4	versicol	v	0.00000	1.15810	0.00000
36	6.7	2.5	5.8	1.8	virginic	a	0.00000	0.00000	0.23912
37	6.5	3.0	5.2	2.0	virginic	a	0.00000	0.00000	1.11533
38	6.2	2.8	4.8	1.8	virginic	a	0.00000	0.06703	0.81524
39	4.9	3.1	1.5	0.1	setosa	s	2.45892	0.00000	0.00000
40	5.6	2.5	3.9	1.1	versicol	v	0.00000	1.47088	0.00000
41	4.9	3.0	1.4	0.2	setosa	s	2.07015	0.00000	0.00000

42	6.7	3.1	4.7	1.5	versicol	v	0.00000	1.46467	0.00000
43	6.5	3.0	5.5	1.8	virginic	a	0.00000	0.09304	0.69716
44	5.1	3.5	1.4	0.2	setosa	s	2.95321	0.00000	0.00000
45	6.3	2.3	4.4	1.3	versicol	v	0.00000	0.35753	0.00000
46	5.8	2.7	5.1	1.9	virginic	a	0.00000	0.00000	0.74200
47	5.6	3.0	4.5	1.5	versicol	v	0.00000	0.86879	0.00000
48	5.4	3.7	1.5	0.2	setosa	s	2.00499	0.00000	0.00000
49	6.8	3.2	5.9	2.3	virginic	a	0.00000	0.00000	0.96966
50	5.8	2.7	4.1	1.0	versicol	v	0.00000	0.55213	0.00000
51	5.6	3.0	4.1	1.3	versicol	v	0.00000	1.10177	0.00000
52	5.1	3.3	1.7	0.5	setosa	s	0.82387	0.00000	0.00000
53	5.0	3.4	1.6	0.4	setosa	s	2.46303	0.00000	0.00000
54	6.3	2.5	4.9	1.5	versicol	v	0.00000	0.20800	0.15073
55	5.7	2.8	4.5	1.3	versicol	v	0.00000	1.21671	0.01730
56	4.6	3.2	1.4	0.2	setosa	s	2.70909	0.00000	0.00000
57	6.1	2.8	4.7	1.2	versicol	v	0.00000	0.62851	0.01541
58	6.2	3.4	5.4	2.3	virginic	a	0.00000	0.00000	0.37687
59	6.0	2.9	4.5	1.5	versicol	v	0.00000	1.38503	0.04472
60	6.4	2.9	4.3	1.3	versicol	v	0.00000	1.54034	0.00000
61	7.7	3.8	6.7	2.2	virginic	a	0.00000	0.00000	0.18227
62	5.7	2.6	3.5	1.0	versicol	v	0.00000	0.48381	0.00000
63	5.1	2.5	3.0	1.1	versicol	v	0.00000	0.24408	0.00000
64	6.3	3.3	4.7	1.6	versicol	v	0.00000	0.81159	0.00000
65	5.7	4.4	1.5	0.4	setosa	s	0.29147	0.00000	0.00000
66	5.7	2.8	4.1	1.3	versicol	v	0.00000	1.89637	0.00000
67	6.7	3.3	5.7	2.5	virginic	a	0.00000	0.00000	0.66445
68	5.0	3.6	1.4	0.2	setosa	s	2.63728	0.00000	0.00000
69	6.5	3.0	5.8	2.2	virginic	a	0.00000	0.00000	0.66031
70	7.6	3.0	6.6	2.1	virginic	a	0.00000	0.00000	0.38314
71	7.7	3.0	6.1	2.3	virginic	a	0.00000	0.00000	0.16837
72	6.9	3.1	5.4	2.1	virginic	a	0.00000	0.00000	0.75360
73	4.9	2.5	4.5	1.7	virginic	a	0.00000	0.00000	0.16837
74	6.3	3.4	5.6	2.4	virginic	a	0.00000	0.00000	0.47821
75	5.4	3.9	1.3	0.4	setosa	s	0.52559	0.00000	0.00000
76	5.5	3.5	1.3	0.2	setosa	s	0.72111	0.00000	0.00000
77	5.8	2.7	5.1	1.9	virginic	a	0.00000	0.00000	0.74200
78	5.1	3.5	1.4	0.3	setosa	s	2.78898	0.00000	0.00000
79	6.7	3.0	5.2	2.3	virginic	a	0.00000	0.00000	0.40683
80	5.7	2.9	4.2	1.3	versicol	v	0.00000	1.72654	0.00000
81	7.2	3.6	6.1	2.5	virginic	a	0.00000	0.00000	0.39273
82	7.0	3.2	4.7	1.4	versicol	v	0.00000	0.96155	0.00000
83	7.2	3.0	5.8	1.6	virginic	a	0.00000	0.00000	0.27166
84	6.5	3.2	5.1	2.0	virginic	a	0.00000	0.00104	0.75789
85	6.4	3.2	4.5	1.5	versicol	v	0.00000	1.28072	0.00000
86	5.4	3.4	1.5	0.4	setosa	s	0.86072	0.00000	0.00000
87	6.1	2.9	4.7	1.4	versicol	v	0.00000	1.16106	0.09171
88	6.1	2.8	4.0	1.3	versicol	v	0.00000	1.22411	0.00000
89	5.2	4.1	1.5	0.1	setosa	s	0.32588	0.00000	0.00000
90	5.6	2.9	3.6	1.3	versicol	v	0.00000	0.40828	0.00000
91	4.8	3.1	1.6	0.2	setosa	s	2.47391	0.00000	0.00000
92	7.1	3.0	5.9	2.1	virginic	a	0.00000	0.00000	0.56880
93	6.9	3.1	4.9	1.5	versicol	v	0.00000	1.15453	0.00000
94	5.9	3.0	5.1	1.8	virginic	a	0.00000	0.15187	0.68230
95	6.4	3.1	5.5	1.8	virginic	a	0.00000	0.07470	0.61539
96	6.4	2.7	5.3	1.9	virginic	a	0.00000	0.00000	0.90590
97	5.5	2.4	3.8	1.1	versicol	v	0.00000	1.39888	0.00000
98	6.0	3.0	4.8	1.8	virginic	a	0.00000	0.14455	0.66499
99	6.8	3.0	5.5	2.1	virginic	a	0.00000	0.00000	1.03360
100	6.3	2.5	5.0	1.9	virginic	a	0.00000	0.00000	0.47837
101	6.3	2.9	5.6	1.8	virginic	a	0.00000	0.07374	0.48714
102	5.5	2.5	4.0	1.3	versicol	v	0.00000	1.22465	0.00000
103	6.4	3.2	5.3	2.3	virginic	a	0.00000	0.00000	0.64933
104	4.9	2.4	3.3	1.0	versicol	v	0.00000	0.52128	0.00000
105	4.8	3.4	1.6	0.2	setosa	s	2.18686	0.00000	0.00000
106	5.2	3.4	1.4	0.2	setosa	s	2.54537	0.00000	0.00000
107	5.6	2.7	4.2	1.3	versicol	v	0.00000	1.75733	0.00000
108	4.8	3.0	1.4	0.1	setosa	s	2.34900	0.00000	0.00000
109	4.9	3.6	1.4	0.1	setosa	s	1.45495	0.00000	0.00000
110	5.7	3.0	4.2	1.2	versicol	v	0.00000	0.87699	0.00000

111	4.5	2.3	1.3	0.3	setosa	s	0.16837	0.00000	0.00000
112	5.4	3.0	4.5	1.5	versicol	v	0.00000	0.37238	0.00000
113	5.5	4.2	1.4	0.2	setosa	s	0.34852	0.00000	0.00000
114	5.7	2.5	5.0	2.0	virginic	a	0.00000	0.00000	0.56046
115	7.9	3.8	6.4	2.0	virginic	a	0.00000	0.00000	0.18227
116	4.9	3.1	1.5	0.2	setosa	s	2.91928	0.00000	0.00000
117	5.7	3.8	1.7	0.3	setosa	s	1.37419	0.00000	0.00000
118	4.4	3.2	1.3	0.2	setosa	s	1.83725	0.00000	0.00000
119	7.7	2.6	6.9	2.3	virginic	a	0.00000	0.00000	0.16837
120	5.1	3.8	1.5	0.3	setosa	s	1.93537	0.00000	0.00000
121	6.7	3.3	5.7	2.1	virginic	a	0.00000	0.00000	0.48916
122	6.0	2.2	5.0	1.5	virginic	a	0.00000	0.07398	0.16837
123	5.5	2.3	4.0	1.3	versicol	v	0.00000	0.58864	0.00000
124	5.0	3.5	1.6	0.6	setosa	s	0.45768	0.00000	0.00000
125	5.4	3.9	1.7	0.4	setosa	s	1.52083	0.00000	0.00000
126	6.6	3.0	4.4	1.4	versicol	v	0.00000	1.41981	0.00000
127	5.8	2.6	4.0	1.2	versicol	v	0.00000	1.60408	0.00000
128	6.6	2.9	4.6	1.3	versicol	v	0.00000	1.25884	0.00000
129	6.8	2.8	4.8	1.4	versicol	v	0.00000	0.90424	0.00000
130	4.4	3.0	1.3	0.2	setosa	s	2.06088	0.00000	0.00000
131	5.2	2.7	3.9	1.4	versicol	v	0.00000	0.64543	0.00000
132	5.4	3.4	1.7	0.2	setosa	s	2.05420	0.00000	0.00000
133	7.7	2.8	6.7	2.0	virginic	a	0.00000	0.00000	0.36303
134	6.0	3.4	4.5	1.6	versicol	v	0.00000	0.48936	0.00000
135	5.1	3.7	1.5	0.4	setosa	s	1.99944	0.00000	0.00000
136	6.3	2.7	4.9	1.8	virginic	a	0.00000	0.08049	0.77854
137	5.0	3.4	1.5	0.2	setosa	s	3.48537	0.00000	0.00000
138	5.2	3.5	1.5	0.2	setosa	s	2.90439	0.00000	0.00000
139	4.7	3.2	1.6	0.2	setosa	s	2.12959	0.00000	0.00000
140	4.4	2.9	1.4	0.2	setosa	s	1.58656	0.00000	0.00000
141	5.8	2.8	5.1	2.4	virginic	a	0.00000	0.00000	0.16837
142	5.5	2.6	4.4	1.2	versicol	v	0.00000	0.66678	0.00000
143	5.0	2.0	3.5	1.0	versicol	v	0.00000	0.52162	0.00000
144	6.7	3.0	5.0	1.7	versicol	v	0.00000	0.44978	0.13202
145	5.0	2.3	3.3	1.0	versicol	v	0.00000	0.76354	0.00000
146	5.5	2.4	3.7	1.0	versicol	v	0.00000	1.12145	0.00000
147	6.3	3.3	6.0	2.5	virginic	a	0.00000	0.00000	0.22614
148	7.2	3.2	6.0	1.8	virginic	a	0.00000	0.00000	0.38714
149	5.8	2.7	3.9	1.2	versicol	v	0.00000	1.56462	0.00000
150	5.3	3.7	1.5	0.2	setosa	s	2.17642	0.00000	0.00000

En el archivo de salida nos informa:

. Hay 150 observaciones, 4 variables y 3 clases.

. Número de individuos de cada clase, 50 para cada una de ellas.

. 9 observaciones de virginica han sido clasificadas como versicolor, 17 observaciones de versicolor han sido clasificadas como virginicas y 23 observaciones de versicolor han sido clasificadas como virginicas

. Hay 6 individuos clasificados en el grupo **OTHER** pues son individuos que caen fuera de los tres elipsoides

. Las demás salidas tienen un sentido semejante al del ejemplo C19-1.

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

CAPÍTULO 20

Datos de Respuesta Categórica

CAPÍTULO 20

Datos de Respuesta Categórica

Introducción.-

Todas las técnicas estadísticas estudiadas en los capítulos anteriores (si no se ha indicado otra situación) han consistido en la estimas de los parámetros de la distribución Normal y pruebas de hipótesis relativas a los parámetros de dicha distribución. Los datos que se han manejado hasta ahora eran datos cuantitativos continuos (variable de proporción) y, por lo tanto, se adoptaba el modelo probabilístico Normal. En los ejemplos en los que los datos no eran cuantitativos continuos, para usar la misma metodología, se realizaba una aproximación a la normal como consecuencia del Teorema Central del Límite o se aproxima a la normal mediante una *transformación* de los datos.

Toda la estadística inferencial usada hasta el momento, puesto que se basaba en el supuesto de normalidad, se ha basado en los modelos de *ANOVA* y *Regresión*.

Pero en muchos casos este supuesto es difícilmente justificable cuando no es claramente erróneo, y en otras ocasiones, aunque sea justificable, las desviaciones de las condiciones paramétricas del modelo normal son lo suficientemente grandes como para invalidar la utilización de las técnicas de la estadística vista hasta ahora.

Para operar con datos de variables no normales o datos en los que no es fácil determinar la distribución original, hay dos opciones

1) Utilizar los *estadísticos de distribución libre* o *independientes de la distribución*, es decir, procedimientos que no dependan de una distribución específica, sino que sean válidos bajo condiciones distribucionales muy amplias. En los que no se especifica la naturaleza de la distribución, por lo que no se utilizan parámetros. Los estadísticos no paramétricos comparan distribuciones y no parámetros. Estos estadísticos pueden ser sensibles a cambios de localización, de dispersión o a ambos cambios.

2) Utilizar los métodos propios del modelo *loglinear* y modelo *logístico* que son modelos lineales del *Modelo General Lineal* y, por lo tanto, muy análogos a los modelos de regresión pero para respuestas categóricas.

En el presente capítulo se van a exponer metodologías uni y multivariante para describir y realizar inferencias de variables de respuesta categórica. Las pruebas serán una veces pruebas no paramétricas y otras veces pruebas paramétricas pero en las que, aunque los modelos que se van a utilizar son muy semejantes a los modelos de regresión (de variables de respuesta continua), se asumen como distribución de la variable de respuesta la distribución *binomial*, la *multinomial* o la de *Poisson*, en lugar de la *normal*.

El modelo logístico se va a utilizar con las variables de respuesta binomial o multinomial, mientras que el modelo loglineal se va ha utilizar con las variables Poisson. Aunque existen muchas equivalencias entre estos dos modelos.

Para presentar estas pruebas se utiliza la misma ordenación o programa de los capítulos anteriores, utilizando muchas veces, inclusive, la misma denominación con el calificativo de *no paramétrica* o *para datos categóricos*.

Como ya se explicó en el Capítulo 2, una variable categórica es la que representa un número o frecuencia de unidades que caen dentro de una categorías o clase de entre un conjunto de clases. Por ejemplo, la ideología política puede medirse como de *izquierda, centro* y *derecha*; el status de los fumadores puede medirse como *nunca fuma, fuma esporádicamente* y *fuma asiduamente*; *etc.*

Estos tipos de datos son muy corrientes en Ciencias Sociales y Ciencias Biomédicas. También se presentan con cierta frecuencia en Etología, Ecología, Sanidad Publica, Ciencias de la Educación, Marketing. Y tiene campos de aplicación muy interesantes en los Controles de Calidad Industrial en los que hay que evaluar características a menudo muy subjetivas como puede ser la característica al tacto de cierto producto, o el buen sabor de un cierto alimento, *etc.*

Hay muchos tipos de variables categóricas y se puede clasificar de varias maneras, ya vistas en el Capítulo 2.

Clasificación.-

Los datos categóricos lo son de variables que se pueden clasificar. Normalmente es posible clasificar los miembros de una población de muchos modos. Por ejemplo, las personas se pueden clasificar por su sexo, por su estado civil, por su mayoría o no de edad, *etc.* Estas son clasificaciones dicotómicas, pero también hay clasificaciones múltiples como pueden ser, en personas, las encuestas de opinión política, *etc.* Ya sea una clasificación dicotómica (binomial) o múltiple (multinomial), ésta tiene que ser *exhaustiva* y las categorías en las que se separan los miembros de la población tienen que ser *mutuamente excluyentes*.

Clasificación exhaustiva: Una clasificación es exhaustiva cuando hay suficientes

categorías o clases como para acomodar a todos los miembros de la población.

Clasificación mutuamente excluyente: Las categorías o clases son mutuamente excluyentes cuando cada miembro de la población puede ser colocado sólo en una de las clases.

Ocurrirán ciertos casos en que para conseguir que la clasificación sea exhaustiva y mutuamente excluyente se tenga que incluir una clase de *Otros*, o de *No sabe/no contesta*, o de mayor que, o menor que, o ambos, *etc.*

Uso de las pruebas de datos categóricos.-

Los análisis que se van a estudiar en este capítulo se aplican en los siguientes casos.

1) Cuando solo sea posible el clasificar los datos por clases o categorías (variables nominales) por falta de una escala de medida adecuada o posible. En este caso lo ideal es hacer una de estas prueba. Otras veces, el tomar los datos como categorías, es una manera rápida de recolectar datos, siendo, por tanto, una de estas pruebas la adecuada para las necesidades del investigador.

2) Cuando es posible ordenar los datos de mayor a menor, se dispone de datos categóricos. Por ejemplo, pueden asignarse ordenaciones por textura o sabor a productos alimenticios; o bien clasificar plantas, animales o personas por el grado de infección de virus, o parcelas por una plaga de insectos; en un ensayo de variedades que implique muchas localizaciones, las varianzas pueden ser heterogéneas no cumpliendo los supuestos usuales para un análisis de varianza válido, y las ordenaciones pueden ser la mejor medida para el análisis.

Análisis de datos categóricos cuando se tiene una muestra.-

Primeramente se van a presentar las pruebas estadísticas que se van a utilizar para contrastar hipótesis que requieren solamente una muestra (y una variable). Estas pruebas son las paralelas a las pruebas de la distribución normal del Capítulo 3 y principio del Capítulo 4 y, como aquellas, se va a probar si la muestra que se ha tomado pertenece a una población específica.

Las pruebas con una muestra son usualmente de las del tipo de bondad de ajuste.

Prueba de bondad de ajuste.-

Frecuentemente, lo que se desea no es saber algo acerca de los parámetros de una supuesta distribución, sino acerca de la forma de la distribución. Es decir, se desea

probar la hipótesis de que los datos muestrales provienen de una distribución específica. Para realizar dichas pruebas se necesita variables categóricas.

Las pruebas de bondad de ajuste fueron desarrolladas por *Pearson* a principio de siglo teniendo, desde el primer momento, un impacto revolucionario en el tratamiento estadístico de los datos categóricos y siendo uno de los primeros métodos estadísticos inferenciales. Esta prueba evalúa la probabilidad de que cierta distribución sea igual a una distribución hipotética.

Realización de la prueba de bondad de ajuste. Distribución binomial.-

Entre sus muchas aplicaciones, estas pruebas se utilizan asiduamente en *Genética* para contrastar si la herencia de los caracteres se ajustan a las conocidas como leyes de *Mendel*, que no son sino pruebas de ajuste a una distribución binomial (un carácter) o multinomial (dos o más caracteres). Para comprender la idea básica de estas pruebas se pondrá un ejemplo de la distribución binomial, primero, y multinomial después, referidos ambos a la herencia Mendeliana de diferentes caracteres.

Ejemplo.-

Supóngase que se ha realizado un experimento de cruce entre dos heterozigotos de la F_1 y se obtiene una descendencia en la F_2 de 60 del fenotipo que se supone *dominante* y 40 individuos con el fenotipo *recesivo*, es decir, se observa una proporción 6:4 (3:2) cuando la proporción esperada por las leyes de *Mendel* es 3:1. Si se expresa en forma de frecuencias: $f_1 = pn = 0.75*100 = 75$ y $f_2 = qn = 0.25*100 = 25$, donde n es el tamaño de muestra.

La cuestión es si la desviación observada con respecto a lo esperado es debida al azar. La respuesta a esta pregunta ya se vio en el Capítulo 3 y Capítulo 4 en los que, haciendo uso del Teorema Central del Límite, se hacía una aproximación a la distribución normal; y dado que se tiene una distribución *Bin*(100,0.75) no hay más que comprobar si la p observada (p=0.60) es igual que la p paramétrica (p=0.75).

Otro método, basado en el mismo principio, es utilizar los límites de confianza para las proporciones binomiales haciendo, también, uso del teorema Central del Límite, tal como se ha visto en el Capítulo 4.

Un tercer procedimiento es hacer uso de la teoría de muestreo, con lo que no se tendrá que hacer aproximaciones a la normal y son métodos válidos tanto para

grandes como para pequeñas muestras. Esto es la prueba de bondad de ajuste que se puede desarrollar de la siguiente forma

Fenotipo	*Frecuencias observadas* o	*Frecuencias esperadas* e	*Desviaciones* $(o-e)$	$(o-e)^2$	$\frac{(o-e)^2}{e}$
Dominante	60	75	-15	225	3
Recesivo	40	25	15	225	9
total	100	100	0.0		$\chi^2=12$

En primer lugar se ha medido las desviaciones de las frecuencias observadas con respecto a las esperadas. Lógicamente, la suma de estas desviaciones es igual a cero por la misma razón que lo es la suma de las desviaciones respecto a la media (Capítulo 2); por tanto, en un ejemplo de dos clases, como es este, las desviaciones siempre son de la misma magnitud y diferente signo. Al igual que se hizo en el Capítulo 2, para hacer que la medida de las desviaciones y su suma sean siempre positivas se elevan al cuadrado tal como se expresa en la quinta columna de la tabla anterior. Pero esta cantidad debe expresarse como una proporción de la frecuencia esperada pues si la frecuencia esperada fuera, por ejemplo, de 14.0, una desviación de 15 sería extremadamente grande pues constituye casi el 100% de la frecuencia esperada, pero tal desviación representa sólo el 10% si la frecuencia esperada fuera de 150.0. De esta manera se obtiene la última columna cuya suma es 12.00. Pues bien, este estadístico, que se simboliza como χ^2, se denomina de esta manera porque se ajusta a la distribución del mismo nombre. Recuérdese que la función de densidad de la distribución χ^2 es

$$\chi^2 = \Sigma_i \frac{(X_i - \mu)^2}{\sigma^2}$$

es decir, el numerador es la suma de cuadrados y el denominador es la varianza paramétrica; mientras que la operación que se ha realizado en la tabla anterior es

$$\chi^2 = \Sigma_i \frac{(o_i - e_i)^2}{e_i} =$$

Se puede demostrar que esta expresión es igual a

$$= \frac{(o_1 - p\,n)^2}{pqn} + \frac{(o_2 - qn)^2}{pqn}$$

es decir, un sumatorio en el que el numerador es el cuadrado de la desviación respecto a lo esperado con un grado de libertad (puesto que se tiene dos sumandos) y el denominador es la varianza de la distribución binomial, o sea, una varianza paramétrica, por lo que, aunque ésta no sea una distribución continua y la distribución χ^2 sí lo es, parece razonable que la expresión calculada en la tabla anterior se distribuya como una χ^2 con un grado de libertad.

Si la muestra (el experimento realizado) hubiera salido exactamente la proporción 3:1, entonces el valor de χ^2 hubiera sido cero. A medida que aumenta la desviación respecto de lo esperado, mayor será el valor de χ^2.

Dado que las desviaciones están elevadas al cuadrado, el valor de χ^2 siempre será positivo.

Ahora se necesita saber si el valor de χ^2 del ejemplo es estadísticamente igual a cero o es diferente de cero y para ello se necesita saber cuantos grados de libertad se tienen para poder compararla con el valor de la distribución χ^2 (Tabla 4).

Dado que el total (100) es constante, para elaborar las frecuencias esperadas se puede adoptar cualquier hipótesis siempre y cuando se mantenga el total. Por tanto, dado que se tiene dos clases, si se supone un valor cualquiera en una clase, el valor de la otra clase viene obligado por la suma total. Esto hace que se tenga un grado de libertad.

Por tanto, si el valor del ejemplo es $\chi^2 = 12.00$ y este valor es mayor que el valor de χ^2 de la tabla, tanto al α=0.05 como al α=0.01 ($\chi^2_{(1;\ 0.01)}$=6.635), se rechaza la hipótesis nula de 3:1 y se concluye que la proporción del tipo dominante es inferior que 0.75, por lo que no se confirma que la herencia de estos caracteres se mendeliana.

Realización de la prueba de bondad de ajuste. Distribución multinomial.-

Estas pruebas de bondad de ajuste pueden aplicarse a una distribución con más de dos clases. Véase otro ejemplo.

Ejemplo.-

La pezuña normal (hendida) en el cerdo se supone determinada por el alelo recesivo del gen que controla este carácter, mientras que la pezuña sin hendir (*pie de mula*) esta determinado por el alelo dominante. Por otro lado, el color blanco de la piel en el cerdo se supone determinada por el alelo dominante del gen que controla el color de la piel, mientras que el color negro esta determinado por el alelo recesivo. En la siguiente tabla se muestra el resultado de cruces de dihíbridos en los

que la proporción esperada en la descendencia es de 9:3:3:1, es decir, $p_1 = 9/16$, $p_2 = 3/16$, $p_3 = 3/16$ y $p_4 = 1/16$.

Fenotipo	*o*	*e*	(*o*-*e*)	$(o-e)^2$	$\frac{(o-e)^2}{e}$
blanco-hendido	10	11.25	-1.25	1.5625	0.139
blanco-mula	5	3.75	1.25	1.5625	0.417
negro-hendido	3	3.75	-0.75	0.5625	0.150
negro-mula	2	1.25	0.75	0.5625	0.450
total	20	20.00	0.00		χ^2=1.15

Si suponemos que este χ^2 se distribuye como la χ^2 teórica, se necesita saber

cuántos grados de libertad hay en este ejemplo.

Siguiendo el mismo razonamiento del primer ejemplo, si se tiene cuatro clases, tres de ellas pueden variar libremente pero la cuarta ha de compensar la diferencia entre la suma total y la suma de las otras tres. Por tanto, en un caso de cuatro clases se tendrá tres grados de libertad y, en general, cuando se tiene *t* clases, se tendrá *t*-1 grados de libertad.

Dado que $\chi^2_{(3;\ 0.05)} = 7.815$ y que este valor es mayor que el observado en el problema, no se rechaza la hipótesis nula por lo que se puede concluir que el cruce realizado sigue la proporción 9:3:3:1.

Grados de libertad en las pruebas de bondad de ajuste.-

En los dos ejemplos vistos hasta el momento, la hipótesis que se quería comprobar era fruto de los conocimientos previos del investigador. Los valores de p=0.75 y q=0.25, eran consecuencia de la hipótesis 3:1 y no se obtuvieron a partir de los datos. Lo mismo ocurría con el segundo ejemplo, la hipótesis 9:3:3:1 se basaba también en la teoría genética, por lo que las frecuencias esperadas se basan en una hipótesis *extrínseca*, es decir, una hipótesis externa a los datos. Pero existen muchos casos en los que los parámetros para la hipótesis nula se obtienen a partir de los datos de la muestra, por lo tanto las frecuencias esperadas representan una hipótesis *intrínseca*; en tal caso, para obtener el número correcto de grados de libertad para la prueba χ^2 de bondad de ajuste se deberá restar, del número de clases en que se han repartido los datos, no sólo un grado de libertad por el motivo anteriormente estudiado, sino también un grado de libertad por cada parámetro de la población estimado a partir de la muestra.

Ejemplo.-

Se ha contabilizado el número de conservas contaminadas, de cierta marca, encontradas en un muestreo de 98 establecimientos y se quiere saber si se ajusta a una distribución de Poisson.

Nº Conservas	o	e	(o-e)	$(o-e)^2$	$\frac{(o-e)^2}{e}$
0	3	4.78	-1.78	3.17	0.66
1	17	14.44	2.56	6.55	0.45
2	26	21.81	4.19	17.56	0.80
3	16	21.96	-5.96	35.52	1.62
4	18	16.58	1.42	2.02	0.12
5	9	10.02	-1.02	1.04	0.10
6	3	5.04	-2.04	4.16	0.83
7	5	2.18	2.82	7.95	3.65
8	0	0.82	-0.82	0.67	0.82
9	1	0.27	0.73	0.53	1.97
10	0	0.08	-0.08	0.006	0.08
11 ó más	0	0.03	-0.03	0.000	0.03
	98	98.01	-0.01		χ^2=8.26

El único paso nuevo es el de los grados de libertad. En este ejemplo *gl*=*t*-1=12-1=11; pero como para obtener las frecuencias esperadas se ha tenido que estimar λ (parámetro de la distribución de Poisson) con los datos de la muestra (λ=3.0204), se tiene que restar otra unidad a los grados de libertad. Siendo entonces

$$gl = t - 1 - 1 = 10$$

$$\chi^2_{(10;\,0.05)} = 1.8307$$

por tanto, el χ^2 es no significativo y se acepta, por el momento, la hipótesis nula, concluyendo que el ajuste a la distribución de Poisson es satisfactorio.

Otras pruebas de bondad de ajuste serán las relativas a la distribución Binomial y a la distribución Normal. Para la distribución Binomial, los grados de libertad serán el número de clases menos uno y menos uno si se estima el parámetro *p* a partir de la muestra, si no fuera este el caso, porque se conociera *p* sin utilizar los datos de la muestra, los grados de libertad serían el número de clases menos uno, tal como los ejemplos anteriores, en los que la herencia de unos caracteres se ajustaban a una binomial con *p* determinada por una teoría genética. En el caso de la distribución Normal, ambos parámetros μ y σ^2 generalmente serán estimados a partir de la muestra, por lo que los grados de libertad serán, en total, el número de clases menos tres.

Fórmula general para el cálculo del estadístico χ^2.-

Antes de proceder al análisis de diseños específicos veamos como se puede simplificar el cálculo de χ^2. La expresión dada anteriormente

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

se puede aplicar a cualquier prueba de bondad de ajuste aún cuando se tiene que calcular desviaciones, elevarlas al cuadrado y dividir por las frecuencias esperadas. Esta fórmula, análoga a la fórmula teórica para una suma de cuadrados, no es práctica para el cálculo si se utiliza una calculadora de bolsillo, por lo que se puede desarrollar con objeto de simplificarla

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \sum_i \left(\frac{o_i^2 - 2o_i e_i + e_i^2}{e_i} \right) = \sum_i \frac{o_i^2}{e_i} - 2\sum_i o_i + \sum_i e_i =$$

$$= \sum_i \frac{o_i^2}{e_1^2} - 2N + N = \sum_i \frac{o_i^2}{e_i} - N$$

Prueba de razón de verosimilitudes.-

Existen otras técnicas que pueden utilizarse para comprobar la concordancia de las frecuencias observadas con las esperadas sobre la base de una hipótesis. El método que se va a ver a continuación es mucho más reciente que el χ^2 y tiene propiedades que lo hacen preferible a éste.

Si se dividen las frecuencia observada por la esperada se tiene que, en la hipótesis nula, el valor observado es igual al esperado, al ser los dos valores iguales su cociente (L) sería igual a uno. A medida que la diferencia entre o y e sea mayor, mayor será el cociente. Esto indica que el cociente de estas dos probabilidades o verosimilitudes puede utilizarse como medida del grado de concordancia entre las frecuencias esperadas y observadas. En este cociente se basa la llamada *prueba de razón de verosimilitudes.*

La distribución teórica de este cociente es complicada; sin embargo se ha demostrado que la distribución de

$$G = 2\sum_i o_i \ln\left(\frac{o_i}{e_i}\right)$$

se aproxima a una distribución χ^2, y los grados de libertad son los mismos que los estudiados anteriormente.

Ejemplo.-

Siguiendo con primer ejemplo se tiene

Fenotipo	*o*	*e*	$\left(\frac{o}{e}\right)$	$o \ln\left(\frac{o}{e}\right)$
Dominante	60	75	0.8	-13.3886
Recesivo	40	25	1.6	18.8001
total	100	100		5.4115

$$G = 2 \times 5.4115 = 10.823 **$$

$$\chi^2_{(1;\,0.01)} = 6.635$$

si se compara este valor observado con el χ^2 de la tabla para un grado de libertad, se observa que es significativo como ya lo fue la prueba anterior. En general, *G* será numéricamente muy semejante a χ^2.

Esta prueba *G* tiene varias ventajas sobre la prueba χ^2; una de ellas es la simplificación del cálculo, pero la ventaja más importante es que la *G* se ajusta más fielmente a la distribución teórica χ^2 por lo que las conclusiones que se obtengan con esta prueba serán más exactas que con la prueba tradicional si bien muy pocas veces habrá diferencias manifiestas entre ambas.

Corrección de continuidad.-

En las pruebas de bondad de ajuste que únicamente contienen dos clases, los valores de χ^2 y de *G* contienen un sesgo que puede modificarse aplicando una corrección de continuidad, haciendo que los valores de χ^2 o de *G* se aproximen más a la distribución χ^2. Esta corrección consiste, al igual que se vio anteriormente, en sumar o restar 0.5 a las frecuencias observadas, de manera que se minimice el valor de χ^2 y de *G*.

En el caso del χ^2, esto puede hacerse mediante la resta de 0.5 de los valores absolutos de la desviación (*o*-*e*), quedando la fórmula de la siguiente manera

$$\chi^2_{adj} = \sum_i \frac{(|o_i - e_i| - 0.05)^2}{e_i}$$

Para la corrección de continuidad de la prueba *G*, simplemente se ajustan las frecuencias observadas transformándolas mediante la suma de ± 0.5 a fin de reducir la diferencia entre ellas y las correspondientes frecuencias esperadas, quedando la fórmula de la siguiente manera

$$G_{adj} = 2\sum_i o_i \ln\left(\frac{o_i \pm 0.5}{e_i}\right)$$

Hay autores que recomiendan emplear esta fórmula en todos los casos que haya dos clases, mientras que otros opinan que únicamente es necesaria para muestras de tamaño inferior a 200 datos. En realidad, esta corrección proporciona únicamente una pequeña diferencia de χ^2 y de G, incluso cuando el tamaño de la muestra es menor de 200. Se puede establecer la siguiente regla para comprobar los casos con dos clases: para muestras en las que $n>200$ se utilizan las fórmulas normales de χ^2 o G; para tamaños entre 25 y 200 se usa la corrección de continuidad; y para tamaños menores o iguales a 25 se buscan las probabilidades exactas tal como veremos más adelante.

Ejemplo.-

Siguiendo con el primer ejemplo de la experiencia genética, se tiene

Fenotipo	*o*	*e*	$\frac{(\lvert o-e\rvert-0.5)^2}{e}$
Dominante	60	75	2.8033
Recesivo	40	25	8.4100
total	100	100	$\chi^2_{adj} = 11.2133$

$$\chi^2_{adj} = 11.2133 \text{ * * *}$$

$$\chi^2_{(1;\,0.001)} = 10.828$$

Fenotipo	*o*	*e*	$\left(\frac{o \pm 0.5}{e}\right)$	$o\ln\left(\frac{o \pm 0.5}{e}\right)$
Dominante	60	75	0.8067	-12.8907
Recesivo	40	25	1.5800	18.2970
total	100	100		5.4063

$$G = 2 \times 5.4063 = 10.8126 \text{ * *}$$

$$\chi^2_{(1;\,0.01)} = 6.635$$

Como se ve, tanto para χ^2 como para la G, el valor cambia ligeramente manteniéndose las misma significación. Esto quiere decir que si los valores observados están alejados de los valores críticos, tanto por arriba como por abajo, la conclusión de la prueba no cambiará si se usa la corrección por continuidad.

Clasificaciones de una vía o clasificación única.-

Ahora se aplicará la prueba de bondad de ajuste a frecuencias ordenadas en una clasificación única, donde las frecuencias esperadas se basan en una hipótesis intrínseca a los datos. Estas pruebas de bondad de ajuste pueden aplicarse a cualquier distribución (Normal, Binomial, Poisson, *etc.*). Ya se han visto pruebas para binomial, multinomial y Poisson. Ahora se va ha ver que si se tiene una variable cuantitativa continua (de proporción) se puede agrupar en clases y realizar la prueba de bondad de ajuste a la normal.

Ejemplo.-

Se tiene un carácter productivo medido en una ganadería de 200 individuos y se quiere saber si la variable continua, con la que se mide dicho carácter, se ajusta a una distribución normal.

Clases			*Marcas*	*O*	*e*	$\frac{(o-e)^2}{e}$	$o\ln\left(\frac{o}{e}\right)$
<		14.92	13.575	1	0.69	0.1393	0.3711
14.92	-	17.61	16.265	2	2.02	0.0002	-0.0199
17.61	-	20.30	18.955	5	5.86	0.1262	-0.7936
20.30	-	22.99	21.645	14	13.40	0.0269	0.6132
22.99	-	25.68	24.335	22	24.18	0.1965	-2.0786
25.68	-	28.38	27.025	39	34.46	0.5981	4.8267
28.38	-	31.07	29.727	40	38.78	0.0384	1.2390
31.07	-	33.76	32.415	28	34.46	1.2110	-7.8886
33.76	-	36.45	35.105	25	24.18	0.0278	0.8337
36.45	-	39.14	37.795	13	13.40	0.0119	-0.3940
39.14	-	41.83	40.485	10	5.86	2.9248	5.3443
41.83		44.52	43.175	0	2.02	2.0200	0.0000
44.52		>	45.865	1	0.69	0.3100	0.3711
			Total	200	200.00	7.4605	4.5005

Los valores esperados se han obtenido de la tabla *Z* (Tabla 1) usando la estima de los parámetros $\overline{X} = 29.72$ y $S^2 = 30.04$

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 7.460 ns$$

$$G = 2\sum_i O_i \ln\left(\frac{o}{e}\right) = 2 \times 4.5005 = 9.001 ns$$

$$gl = 13(\text{número clases}) - 1 - 1(\text{estima } \mu) - 1(\text{estima } \sigma) = 10$$

$$\chi^2_{(10;\,0.05)} = 18.307$$

El ajuste a la normal es bueno.

Otras pruebas no paramétricas para una sola variable de clasificación.-

Estas otras pruebas son:

Prueba de *Kolmogorov-Smirnov* para una muestra.-
Prueba de *Shapiro Wilk* para varias muestras.-
Independencia de los errores.-
Pruebas de rachas.-
Prueba de rachas para variables binomiales.-
Pruebas de rachas para datos cuantitativos.-
Coeficiente de correlación serial.-

Todas estas pruebas se han estudiado en el Capítulo 6.

Datos categóricos para dos muestras independientes. Tablas de contingencia.-

Al igual que se vio en el Capítulo 5 con las pruebas de datos normales, al estudiar las diferencias entre dos grupos se pueden usar muestras apareadas o muestras independientes. Dos muestras son independientes cuando son tomadas al azar de dos poblaciones o se le asigna al azar dos tratamientos a individuos de origen arbitrario (ver *Introducción* del Capítulo 5). En cualquier caso no es necesario que las dos muestras sean del mismo tamaño.

En los epígrafes anteriores de este capítulo, a cada unidad se le media una sola variable. Pero muy frecuentemente se medirán dos variables por unidad experimental o individuo, y cuando se tengan medidas de dos variables se representaran en tablas de contingencia de dos vías (doble entrada). Por ejemplo, se toman 100 individuos aleatoriamente y se le mide (observa) la tonalidad del color del pelo (*claro* o *oscuro*) y el color de los ojos, representándose los datos de la siguiente manera

		Pelo		
		Claro	*Oscuro*	*total*
Ojos	*Claro*	19	5	24
	Oscuro	9	67	76
	total	28	72	100

En estos casos, los datos se registran en tablas de doble entrada, una para cada variable clasificatoria, denominadas tabla de contingencia o tablas de asociación, con las que se va a estudiar si dos acontecimientos son independientes o están asociados.

Una relación de este tipo (bivariante) viene definida por la distribución conjunta de las dos variables. La distribución conjunta determina la distribución marginal y la distribución condicional.

Sean X e Y las dos variables respuesta categóricas, teniendo X, i niveles e Y, j niveles. Cuando se clasifican individuos con respecto a ambas variables, hay ij posibles combinaciones de clasificación. Si se elige aleatoriamente un individuo de una población, la respuesta (X, Y) tiene una distribución de probabilidad. Esta distribución se puede expresar en una tabla de doble entrada que tenga *i filas* para las categoría de X y *j columnas* para las categorías de Y. Las *casillas* de la tabla representan los ij posibles eventos, de probabilidades π_{ij}, esto es, π_{ij} expresa la probabilidad de que (X, Y) caiga en la casilla determinada por el cruce de la *i-ésima* fila, *j-ésima* columna.

La distribución de probabilidad π_{ij} es la *distribución conjunta* de X e Y. Las *distribuciones marginales* son los totales de las filas y las columnas y se obtienen sumando las probabilidades conjuntas. Las probabilidades marginales se simbolizan, $\pi_{i.}$ para la variable de las filas y $\pi_{.j}$ para la variable de las columnas. Las distribuciones marginales dan información de las variables individualmente, sin permitir detectar posible asociación entre las dos variables.

Se cumple que

$$\sum_i \pi_{i.} = \sum_j \pi_{.j} = \sum_i \sum_j \pi_{ij} = 1$$

Diseños muestrales en tablas de contingencia de dos o más vías.-

Antes de continuar con las probabilidades conjuntas y probabilidades marginales hagamos notar que existen diferentes modelos de tablas de contingencia relacionados con diferentes diseños experimentales o diseños muestrales. El tipo de muestreo determinará si las filas, las columnas o ambas son fijas o aleatorias. Se pueden presentar tres situaciones.

	Número de		
Tipo de Diseño	*Variables dependientes*	*Variables Independientes*	*Marginales fijos*
Muestreo aleatorio simple (Modelo I)	2	0	0
Muestreo aleatorio estratificado (Modelo II)	1	1	1
Experimento aleatorio (Modelo III)	0	2	2

Muestreo aleatorio simple o Modelo I.-

En la tabla de contingencia anterior los individuos se clasifican de acuerdo con dos variables. Otro ejemplo puede ser: se toman 100 individuos, los cuales pueden clasificarse como *fumadores* o *no fumadores*, y al mismo tiempo se puede clasificar como individuos *con enfermedad coronaria* o *sin enfermedad coronaria*; hay dos variables de clasificación cuyos totales marginales variaran de una experiencia a otra (serán aleatorios), reflejando los parámetros de la población, sin que el investigador pueda incidir sobre ellos; mientras que el total global si es fijo (determinado por el

experimentador). Este es el denominado *Modelo* I.

La distribución de frecuencias en dichas tablas es la distribución multinomial. Hay dos variables dependientes y la prueba de no asociación significará independencia de las dos variables aleatorias.

La prueba *G* de razón de verosimilitud fue desarrollada para inferir en modelos de tipo I.

Un tipo especial de *Modelo* I es el de datos emparejados (ver Capítulo 4), por ejemplo, estudiar la intención de voto en marido y mujer, o estudiar cierta característica en los dos ojos de cada individuo, *etc.* Una manera muy usual de datos emparejados es el autoemparejamiento que consiste en someter a los mismos individuos a dos tratamientos sucesivos, de manera que un individuo es su propio control, siendo el tamaño de muestra lo único fijado por el investigador, pero la hipótesis que hay que probar es diferente por lo que se estudiará aparte.

Muestreo aleatorio simple estratificado o Modelo II.-

En otros casos, los individuos pueden asignarse a grupos y luego clasificarse dentro de cada grupo con respecto a alguna variable. Esto es, un individuo puede asignarse a un grupo de tratamiento o a un grupo de control y posteriormente clasificarse según la respuesta a un estímulo o a un tratamiento, por ejemplo, se puede tomar 50 individuos y aplicarle un supuesto analgésico y otro 50 individuos y aplicarle un placebo (control), y medir si se alivia o no el dolor. Por tanto uno de los totales marginales viene determinado por el experimentador cuando diseña la experiencia, mientras que el otro marginal es aleatorio. Este es el denominado *Modelo* II.

En estos modelos se tiene que una variable es dependiente o de respuesta (la aleatoria) y la otra es una variable independiente o explicativa otra (la fija), y la prueba de no asociación significará que la probabilidad de aliviar el dolor en la población del supuesto analgésico es la misma que la probabilidad de aliviar el dolor en la población del placebo.

En tablas de contingencia del Modelo II no tiene sentido la noción de distribución conjunta de ambas variables, pues para cada nivel fijo de la variable *X*, *Y* tiene una distribución de probabilidad, por lo que lo adecuado es estudiar la distribución de *Y* para cada nivel fijo de *X*, esto es, se tendrá una distribución de probabilidad condicional de *Y* para cada nivel *i* de *X*, de manera que

$$\sum_j \pi_{j|i} = 1$$

Experimento aleatorio o Modelo III.-

Un tercer modelo para el diseño de una tabla de contingencia es aquél en el que ambos totales marginales son fijos. Este caso lo ilustra *Fisher* con el siguiente ejemplo: Tomando el té, cierta señora afirma que es capaz de distinguir por el sabor que es lo primero que se ha vertido en la taza, el té o la leche; para probar dicha afirmación se preparan cuatro tazas en las que se vierte primero el té y otras cuatro en las que se vierte primero la leche; la señora sabe que hay cuatro tazas de cada tipo y tiene que adivinar cuáles son, los resultados son

		Dedujo que se vertió primero		
		Leche	*Té*	*total*
Se vertió primero	*Leche*	3	1	4
	Té	1	3	4
	total	4	4	8

En este modelo las dos variables son independientes o explicatorias y los dos totales marginales son fijos. Por tanto, la distribución de probabilidad de las celdillas se ajustan a una distribución *hipergeométrica.*

Análisis de las tablas de contingencia de dos o más vías.-

Las tablas de contingencia pueden analizarse por:

(a) Métodos Inferenciales Tradicionales de **Pruebas de Asociación o Independencia;**
(b) Modelos Lineales y **Modelos Loglineales**, o
(c) Análisis Factorial de Correspondencias

Pruebas de Asociación o independencia.-

Las pruebas de independencia propuestas para las tablas de contingencia son de tres clases:

(a) **Prueba χ^2 de *Pearson*,**
(b) **Prueba *G* de Razón de Verosimilitud** y
(c) **Prueba *exacta de Fisher*.**

La prueba χ^2 es la más utilizada y no esta desarrollada específicamente para ninguno de los modelos, mientras que la prueba *G* se desarrolló específicamente para el Modelo I y la prueba exacta de *Fisher* se desarrollo para el Modelo III.

Modelo Lineal y Loglineal.-

Las tablas de contingencia, si son del Modelo I o Modelo II se pueden analizar por medio del modelo lineal y loglineal, respectivamente.

Se utilizará el modelo loglineal en/para:

1) En el modelo aleatorio simple (Modelo I), es decir, en las tablas de contingencia en las que todas las variables son dependientes y no hay ningún marginal fijo.
2) Interpretaciones de no asociación significa independencia estadística.
3) Puede asimilarse a un análisis de independencia entre variables.
4) Generalmente se usa el análisis de máxima verosimilitud.

Se utilizará el modelo lineal en/para:

1) En el modelo aleatorio simple estratificado (Modelo II), es decir, en las tablas de contingencia en las que algunas variables son variables independientes que definen poblaciones. Por ello, las probabilidades se comparan entre estas poblaciones.
2) Interpretaciones de no asociación significa igualdad de probabilidades.
3) Puede asimilarse a un análisis de igualdad de probabilidades.
4) Generalmente se usa el análisis de mínimos cuadrados ponderados.

Análisis Factorial de Correspondencias.-

El Análisis Factorial de Correspondencias o simplemente Análisis de Correspondencias es el método más reciente de análisis multivariante y de los más utilizados. Es una técnica para el estudio de las relaciones de dependencia entre variables categóricas de tipo nominal representadas en tablas de contingencia. Se basa en **distancias** χ^2 por lo que tiene una fuerte relación con dicha prueba. Sin embargo, el análisis de correspondencias, además de analizar la relación existente entre las variables, permite analizar como está estructurada esta relación, lo que lo hace útil incluso en tablas de contingencia de doble entrada cuando el número de clases de cada clasificación es elevado.

Pruebas de independencia o asociación para tablas de contingencia.-

Sea la siguiente tabla de contingencia de tamaño 2x2

	Columna 1	*Columna* 2	*total*
Fila 1	n_{11}	n_{12}	$n_{1.}$
Fila 2	n_{21}	n_{22}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	N

A partir de esta tabla de dos variables y basándose en la hipótesis nula de independencia, se puede calcular las frecuencias esperadas y compararlas con las frecuencias observadas con las formulas vistas anteriormente. La frecuencia esperada de, por ejemplo, n_{22} sería

$$\hat{n}_{22} = \frac{n_{2.}\ n_{.2}}{N}$$

es decir, las frecuencias esperadas, para cada celdilla de la tabla, se calculan multiplicando los totales de la fila y la columna en las que se encuentra la celdilla, dividido por el tamaño total de muestra.

Los totales de los esperados serán los mismos que los observados ya que las frecuencias esperadas han sido calculadas sobre la base de los totales correspondientes a la tabla de frecuencias observadas.

Como se ha dicho anteriormente la prueba de asociación de estas tablas se puede hacer mediante el χ^2 de *Pearson*, mediante la razón de verosimilitud (*G*) o mediante la prueba exacta de *Fisher*. La prueba χ^2 y la prueba *G* se hacen tal como se ha visto anteriormente para una sola variable.

Los criterios para elegir una de entre las tres pruebas, tal como se dijo en el epígrafe anterior, se basan en el modelo del diseño muestral. Aunque también se utiliza como criterio para utilizar la prueba exacta de *Fisher*, el pequeño tamaño de la muestra, aunque, afortunadamente, proporcionan resultados bastante similares en todos los casos.

Las formulas para los dos primeros estadísticos es

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$$G = 2\ \sum_i o_i \ln\left(\frac{o}{e}\right)$$

Pero existen fórmulas más simples cuanto se tienen tablas de contingencia. Para tablas 2x2, la prueba χ^2 es

$$\chi^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2 N}{n_{1.}\ n_{2.}\ n_{.1}\ n_{.2}}$$

Y para la prueba *G*, razón de verosimilitud

$$G = 2[\sum (O_{\text{en cada casilla}} \ln O_{\text{en cada casilla}}) -$$
$$- \sum (O_{\text{totales de filas y columnas}} \ln = \text{totales de filas y columnas}) +$$
$$+ N \ln N]$$

Aunque la fórmula de la prueba de χ^2 parece ser más sencilla que la de *G*, en realidad el cálculo de esta última suele resultar más fácil, especialmente teniendo en cuenta que los *ln* se obtienen directamente de la calculadora de bolsillo. Si se realizan las operaciones con calculadora de bolsillo, muchas veces ocurrirá que solo se pueda hacer la *G*, esto es debido a que el χ^2 requiere el producto sucesivo de los cuatro

totales marginales, que no solo es una operación engorrosa, sin otro algoritmo de solución, sino que además puede causar, frecuentemente, que el producto resultante exceda la capacidad de la calculadora.

Para tablas de contingencia mayores de 2x2, como puede ser esta

	Columna 1	*Columna* 2	...	*Columna c*	*total*
Fila 1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
Fila 2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
Fila f	n_{f1}	n_{f2}	...	n_{fc}	$n_{f.}$
total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	N

la fórmula de cálculo reducido es

$$\chi^2 = \frac{N}{n_{1.}}\left(\frac{n_{11}^2}{n_{.1}} + \frac{n_{12}^2}{n_{.2}} + \ldots + \frac{n_{1c}^2}{n_{.c}}\right) + \frac{N}{n_{2.}}\left(\frac{n_{21}^2}{n_{.1}} + \frac{n_{22}^2}{n_{.2}} + \ldots + \frac{n_{2c}^2}{n_{.c}}\right) +$$
$$+ \ldots\ldots\ldots\ldots\ldots + \frac{N}{n_{f.}}\left(\frac{n_{f1}^2}{n_{.1}} + \frac{n_{f2}^2}{n_{.2}} + \ldots + \frac{n_{fc}^2}{n_{.c}}\right) - N$$

Y para la prueba *G*, razón de verosimilitud, sería igual que la anterior, esta es

$$G = 2[\Sigma\,(O_{\text{en cada casilla}} \ln O_{\text{en cada casilla}}) -$$
$$- \Sigma\,(O_{\text{totales de filas y columnas}} \ln O_{\text{totales de filas y columnas}}) +$$
$$+ N \ln N]$$

Grados de libertad para las tablas de contingencia.-

Los grados de libertad para una tabla de contingencia se pueden calcular por la misma regla expuesta anteriormente para una sola variable. Existen *k* casillas en la tabla y se debe restar un grado de libertad por cada parámetro independiente que se haya obtenido a partir de los datos. Por tanto, si se han calculado (*F*-1) frecuencias esperadas en cada columna (siendo *F* el número de filas) puesto que el valor de la última casilla viene determinado como el total de la columna menos la suma de los *F*-1 valores calculados; y se ha hallado (*C*-1) frecuencias esperadas en cada fila (siendo *C* el número de columnas), los grados de libertad son

$$gl = K - (F-1) - (C-1) - 1 = k - F - C + 1$$

y dado que $k=C^*F$

$$gl = (F \times C) - F - C + 1 = (F-1) \times (C-1)$$

Esto es, los grados de libertad en una tabla de contingencia es igual al número de filas menos uno por el número de columnas menos uno.

Corrección por continuidad.-

Hay autores que recomiendan el uso de la corrección de *Yates* por continuidad para tablas con un grado de libertad y tamaño de muestra es pequeño (n<20). Esta corrección, como la vista para una sola variable, ajusta las frecuencias observadas añadiendo o restando 0.5, de manera que se reduzcan las desviaciones con respecto a las frecuencias esperadas. En una prueba de independencia, esto significa que se añade 0.5 a las frecuencias n_{11} y n_{22} y se resta 0.5 de las frecuencias n_{12} y n_{21} cuando la cantidad $n_{11}\,n_{22}$ - $n_{12}\,n_{21}$ es negativa. Si esta cantidad fuese positiva se haría la suma y la resta contraria. Este ajuste puede realizarse automáticamente mediante el uso de la siguiente expresión:

$$\chi^2_{adj} = \frac{\left(\left| n_{11}\,n_{22} - n_{12}\,n_{21} \right| - \frac{N}{2} \right)^2 N}{n_{1.}\;\; n_{2.}\;\; n_{.1}\;\; n_{.2}}$$

En el caso de la G no existe una ecuación simplificada para la corrección de *Yates*, sino que simplemente se deben ajustar las frecuencias n_{11}, n_{12}, n_{21} y n_{22}, como se ha descrito más arriba. Si $n_{11}\,n_{22}$ - $n_{12}\,n_{21}$ es negativo, se le suma 0.5 a las casillas n_{11} y n_{22} y se le resta 0.5 a las casillas n_{12} y n_{21} y si es positivo se hace la suma y la resta contraria.

Si bien hay autores que han demostrado que la aplicación de la corrección de *Yates* da casi siempre como resultado una prueba demasiado conservadora (el error de Tipo I es mucho menor que el deseado). Parece, por tanto, que el uso de esta corrección es innecesaria, incluso con muestras de pequeño tamaño, y en todo caso, si se hace, es conveniente hacer también la prueba sin corrección y comparar ambos resultados.

Otra opción es realizar en estos casos la prueba exacta de *Fisher*

Otras medidas de asociación.-

Existen otras medidas de asociación entre dos variables categóricas, tres de ellas son: el coeficiente *Phi* (φ), el coeficiente de contingencia P y la V de *Cramer*.

El coeficiente *Phi* (φ) se deriva del valor de χ^2. Su campo de variación es

$$-1 < \varphi < +1$$

por lo que puede ser considerado como un coeficiente de correlación de puntos entre las filas y las columnas.

Este coeficiente se calcula, en una tabla 2x2,

$$\varphi = \frac{n_{11}\ n_{22} - n_{12}\ n_{21}}{\sqrt{n_{1.}\ n_{2.}\ n_{.1}\ n_{.2}}}$$

Y para tablas mayores se calcula,

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

El *coeficiente de contingencia* (*P*) mide la intensidad de asociación existente entre las filas y las columnas. Su campo de variación es

$$0 < P < +1$$

Este coeficiente se calcula

$$P = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

La *V* de *Cramer* es una tercera medida de asociación derivada del valor de χ^2. Su campo de variación es

$$-1 < V < +1$$

Para las tablas 2x2 se tiene que $V=\varphi$, en los demás casos se tiene

$$V = \sqrt{\frac{\frac{\chi^2}{N}}{\min\ (F-1),(C-1)}}$$

Ejemplo.-

Una muestra de 100 ratones se dividió en dos grupos: 50 ratones fueron vacunados y pasados unos días se les inoculó una dosis del virus, mientras que el segundo grupo de 50 ratones recibieron únicamente el virus, sin vacuna previa. Después de que hubiese pasado un tiempo suficiente para que se desarrollara el periodo de incubación y para que la enfermedad siguiera su curso, manifestaron los síntomas de la enfermedad 34 ratones, mientras que 66 ratones no manifestaron síntomas alguno. De los que manifestaron la enfermedad, 11 habían sido vacunados, mientras que 23 habían recibido únicamente virus. La cuestión es si la vacuna protege, de tal manera que hay una mayor proporción de no enfermos en el grupo de vacunados.

Se ve claramente que es un Modelo II o muestreo aleatorio simple estratificado.

Los datos se disponen convenientemente en forma de una tabla de dos factores, que en este caso se trata de una tabla de 2x2

	Respuesta		
	Enfermos	*No enfermos*	*total*
Vacunados	11	39	50
No vacunados	23	27	50
total	34	66	100

Los resultados son mostrados claramente en esta tabla; observándola, por ejemplo, se sabe que 11 ratones de los que fueron vacunados, enfermaron, esto es, enfermaron un 22% de los ratones vacunados; mientras que, de los no vacunados, enfermaron 23, esto es, un 46%. Los totales marginales de la tabla proporcionan el número de ratones que exhiben cualquier propiedad: 50 ratones fueron vacunados; 66 ratones no enfermaron, *etc.* El último marginal representa el número de ratones que intervienen en el experimento, es decir, el total de la muestra.

Los valores esperados en cada casilla son

$$\hat{n}_{11} = \frac{34 \times 50}{100} = 17.0 \quad \hat{n}_{12} = \frac{66 \times 50}{100} = 33.0$$

$$\hat{n}_{21} = \frac{34 \times 50}{100} = 17.0 \quad \hat{n}_{22} = \frac{66 \times 50}{100} = 33.0$$

Quedando la tabla

	Enfermos	*No enfermos*	*total*
Vacunados	17.0	33.0	50.0
No vacunados	17.0	33.0	50.0
total	34.0	66.0	100.0

Los grados de libertad son

$$gl = (2-1)(2-1) = 1$$

por lo que los valores críticos estarán a partir de

$$\chi^2_{(1;\,0.05)} = 3.841$$

La prueba estadística por el método general del χ^2 de *Pearson* es

$$\chi^2 = \frac{(11 - 17.0)^2}{17.0} + \frac{(39 - 33.0)^2}{33.0} + \frac{(23 - 17.0)^2}{17.0} + \frac{(27 - 33.0)^2}{33.0} =$$
$$= 6.4171^*$$

Y por el método general de la razón de verosimilitud, es

$$G = 2\left[11\ln\left(\frac{11}{17}\right) + 39\ln\left(\frac{39}{33}\right) + 23\ln\left(\frac{23}{17}\right) + 27\ln\left(\frac{27}{33}\right)\right] = 6.522^*$$

Recuérdese que el método de la razón de verosimilitud se desarrolló para los muestreos aleatorios simples (Modelo I), aunque también se puede emplear para los muestreos aleatorios simples estratificados (Modelo II), como es el de este ejemplo.

Si se emplean las fórmulas más simples, propias de las tablas 2x2, sería

$$\chi^2 = \frac{[(11\times 27) - (23\times 39)]^2 100}{50\times 50\times 34\times 66} = 6.417^*$$

$$G = 2\,[(11\ln 11 + 39\ln 39 + 23\ln 23 + 27\ln 27) - (50\ln 50 + 50\ln 50 + 34\ln 34 + 66\ln 66)$$
$$+ 100\ln 100] = 6.522^*$$

Se concluye, por tanto, que la probabilidad de enfermar de estos ratones no es la misma en la población de vacunados que en la población de no vacunados. Nótese que el porcentaje de enfermos en la población de vacunados es del 22%, mientras que el porcentaje de enfermos en la población de no vacunados es del 46% (más del doble). Se puede, por tanto, afirmar que el efecto de la vacuna ha sido positivo en la reducción de la manifestación de la enfermedad.

Si se piensa que se debe hacer la corrección por continuidad, las pruebas serían

$$\chi^2_{adj} = \frac{[\,/\,(11\times 27) - (23\times 39)\,/\, - 50]^2 100}{50\times 50\times 34\times 66} = 5.392^*$$

$$G_{adj} = 2\,[(11.5\ln 11.5 + 38.5\ln 38.5 + 22.5\ln 22.5 + 27.5\ln 27.5) -$$
$$- (50\ln 50 + 50\ln 50 + 34\ln 34 + 66\ln 66) +$$
$$+ 100\ln 100] =$$
$$= 5.465^*$$

Como es el objetivo de la corrección, ha disminuido ligeramente el valor de ambos estadísticos pero las conclusiones son las mismas que sin corrección.

Las medidas de asociación para los datos de esta tabla son

$$\varphi = \frac{11\times27 - 23\times39}{\sqrt{50\times50\times34\times66}} = -0.253$$

$$P = \sqrt{\frac{6.417}{100 + 6.417}} = 0.246$$

$$V = \sqrt{\frac{\frac{6.417}{100}}{1}} = 0.253$$

Archivo del programa *SAS* (C20-1.SAS).-

```
title 'Tabla de contingencia, análisis tradicional';
options ls=75 ps=60;
data chi2;
infile 'c20-1.dat';
input vacunado $ respues $ n @@;
proc freq order=data;
weight n;
table vacunado * respues / chisq;
run;
```

Archivo de datos (C20-1.DAT).-

```
si enfermo 11   si noenfer 39
no enfermo 23   no noenfer 27
```

Archivo de resultados (C20-1.LST).-

```
                Tabla de contingencia, análisis tradiciona1

                   TABLE OF VACUNADO BY RESPUES

                VACUNADO      RESPUES

                Frequency|
                Percent  |
                Row Pct  |
                Col Pct  |enfermo |noenfer |   Total
                ---------+--------+--------+
                si       |     11 |     39 |     50
                         |  11.00 |  39.00 |  50.00
                         |  22.00 |  78.00 |
                         |  32.35 |  59.09 |
                ---------+--------+--------+
                no       |     23 |     27 |     50
                         |  23.00 |  27.00 |  50.00
                         |  46.00 |  54.00 |
                         |  67.65 |  40.91 |
                ---------+--------+--------+
                Tota1          34       66      100
                            34.00    66.00   100.00
```

```
STATISTICS FOR TABLE OF VACUNADO BY RESPUES
Statistic                          DF      Value         Prob
-------------------------------------------------------------
Chi-Square                          1      6.417        0.011
Likelihood Ratio Chi-Square         1      6.522        0.011
Continuity Adj. Chi-Square          1      5.392        0.020
Mantel-Haenszel Chi-Square          1      6.353        0.012
Fisher's Exact Test (Left)                              9.78E-03
                    (Right)                                0.997
                    (2-Tail)                               0.020
Phi Coefficient                          -0.253
Contingency Coefficient                   0.246
Cramer's V                               -0.253
Sample Size = 100
```

Veamos un ejemplo de una tabla mayor de 2x2.

Ejemplo.-

En un estudio de toxicidad se tomó una muestra de 1000 ratones hembras preñadas de menos de diez días y se dividió en cinco grupos de grupos: al primer grupo de 200 ratones no se le administro nada, es decir, fueron utilizados como control; a los cuatro grupos restantes se les administro una de cuatro concentraciones de la sustancia estudiada. Dos días después se examinaron los fetos, clasificandolos en tres clases: *muertos*, *malformados* y *normales*. Los datos son

	Respuesta			
Mg/Kg día	*Muertos*	*Deformes*	*Normales*	*total*
Control	14	2	184	200
100	16	1	183	200
200	21	8	171	200
300	39	58	103	200
400	102	88	10	200
total	192	157	651	1000

Se ve claramente que es un Modelo II o muestreo aleatorio simple estratificado. Los datos se disponen convenientemente en forma de una tabla de dos factores, que en este caso se trata de una tabla de 5x3.

Los resultados son mostrados claramente en esta tabla; observándola se sabe, por ejemplo, que hubo 14 fetos muertos espontáneamente y que hubo 58 malformaciones para la concentración 300, *etc.* Se ve que, conforme aumenta la concentración del tóxico, la probabilidad de muertos y deformes aumenta e, inversamente, la probabilidad de normales disminuye. Los marginales de la respuesta proporcionan el número de ratones que exhiben cualquier propiedad: en total hubo 192 fetos muertos, 157 deformes y 651 fetos normales. El último marginal representa el número de ratones que intervienen en el experimento, es decir, el total de la muestra.

Los valores esperados en cada casilla de la respuesta de *muertos*, *malformados* y *normales* son, respectivamente

$$\hat{n}_{f1} = \frac{192 \times 200}{1000} = 38.4$$

$$\hat{n}_{f2} = \frac{157 \times 200}{1000} = 31.4$$

$$\hat{n}_{f3} = \frac{651 \times 200}{1000} = 130.2$$

Quedando la tabla

mg/Kg día	*Muertos*	*Malformados*	*Normales*	*total*
Control	38.4	31.4	130.2	200
100	38.4	31.4	130.2	200
200	38.4	31.4	130.2	200
300	38.4	31.4	130.2	200
400	38.4	31.4	130.2	200
total	192.0	157.0	651.0	1000

Los grados de libertad son

$$gl = (5 - 1)(3 - 1) = 8$$

por lo que los valores críticos estarán a partir de

$$\chi^2_{(8;\,0.05)} = 15.507$$

La prueba estadística por el método general del χ^2 de *Pearson* es

$$\chi^2 = \frac{(14-38.4)^2}{38.4} + \frac{(2-31.4)^2}{31.4} + \frac{(184-130.2)^2}{130.2} +$$
$$+ \frac{(16-38.4)^2}{38.4} + \frac{(1-31.4)^2}{31.1} + \frac{(183-130.2)^2}{130.2} +$$
$$+ \frac{(21-38.4)^2}{(38.4} + \frac{(8-31.4)^2}{31.4} + \frac{(171-130.2)^2}{130.2} +$$
$$+ \frac{(39-38.4)^2}{38.4} + \frac{(58-31.4)^2}{31.4} + \frac{(103-130.2)^2}{130.2} +$$
$$+ \frac{(102-38.4)^2}{38.4} + \frac{(88-31.4)^2}{31.4} + \frac{(10-130.2)^2}{130.2} =$$
$$= 513.8358 \; ***$$

Y por el método general de la razón de verosimilitud, es

$$G = 2\left[14\ln\left(\frac{14}{38.4}\right)+2\ln\left(\frac{2}{31.4}\right)+184\ln\left(\frac{184}{130.2}\right)+\right.$$
$$+16\ln\left(\frac{16}{38.4}\right)+1\ln\left(\frac{1}{31.4}\right)+183\ln\left(\frac{183}{130.2}\right)+$$
$$+21\ln\left(\frac{21}{38.4}\right)+8\left(\frac{8}{31.4}\right)+171\ln\left(\frac{171}{130.2}\right)+$$
$$+39\ln\left(\frac{39}{38.4}\right)+58\left(\frac{58}{31.4}\right)+103\ln\left(\frac{103}{130.2}\right)+$$
$$\left.+102\ln\left(\frac{102}{38.4}\right)+88\ln\left(\frac{88}{31.4}\right)+10\ln\left(\frac{10}{130.2}\right)\right]=$$
$$=577.1449***$$

Recuérdese que el método de la razón de verosimilitud se desarrolló para los muestreos aleatorios simples (Modelo I), aunque también se puede emplear para los muestreos aleatorios simples estratificados (Modelo II), como es el de este ejemplo.

Si se emplean las fórmulas más simples, propias de las tablas *fxc*, sería

$$\chi^2 = \frac{1000}{200}\left(\frac{14^2}{192}+\frac{2^2}{157}+\frac{184^2}{651}\right)+$$
$$+\frac{1000}{200}\left(\frac{16^2}{192}+\frac{1^2}{157}+\frac{183^2}{651}\right)+$$
$$+\frac{1000}{200}\left(\frac{21^2}{192}+\frac{8^2}{157}+\frac{171^2}{651}\right)+$$
$$+\frac{1000}{200}\left(\frac{39^2}{192}+\frac{58^2}{157}+\frac{103^2}{651}\right)+$$
$$+\frac{1000}{200}\left(\frac{102^2}{192}+\frac{88^2}{157}+\frac{10^2}{651}\right)-1000=$$
$$=513.8385***$$

$$
\begin{aligned}
G = 2\,[&(14\ln 14+2\ln 2+184\ln 184 + \\
&+16\ln 16+1\ln 1+183\ln 183 + \\
&+21\ln 21+8\ln 8+171\ln 171 + \\
&+39\ln 39+58\ln 58+103\ln 103 + \\
&+102\ln 102+88\ln 88+10\ln 10) - \\
&-(200\ln 200)5 + \\
&+192\ln 192+157\ln 157+651\ln 651) + \\
&+1000\ln 1000] = \\
&= 577.1449 ***
\end{aligned}
$$

Se concluye, por tanto, que la probabilidad de muerte o malformación de estos ratones no es la misma en la población de no tóxico que en las poblaciones de altas concentraciones del tóxico. Nótese que el porcentaje de muertos, malformados y normales en la población de no tóxico es, respectivamente, 7%, 1% y 92%, mientras que en la población de 400 *mg/Kg día* es 51%, 44% y 5%. Se puede, por tanto, afirmar que el efecto del tóxico ha sido positivo en el aumento de la mortalidad y malformación.

Las medidas de asociación para los datos de esta tabla son

$$\varphi = \sqrt{\frac{513.8385}{1000}} = 0.7168$$

$$P = \sqrt{\frac{513.8385}{1000 + 513.8385}} = 0.5826$$

$$V = \sqrt{\frac{\frac{513.8385}{1000}}{2}} = 0.5069$$

Archivo del programa *SAS* (C20-2.SAS).-

```
title 'Prueba de independencia';
options ls=75 ps=60;
data chi2;
infile 'c20-2.dat';
input cantidad $ respues $ n @@;
proc freq order=data;
weight n;
table cantidad * respues / chisq nocol nopercent;
run;
```

Archivo de datos (C20-2.DAT).-

```
  0 muerto  14    0 defor  2    0 normal 184
100 muerto  16  100 defor  1  100 normal 183
200 muerto  21  200 defor  8  200 normal 171
300 muerto  39  300 defor 58  300 normal 103
400 muerto 102  400 defor 88  400 normal  10
```

Archivo de resultados (C20-2.LST).-

Prueba de independencia

TABLE OF CANTIDAD BY RESPUES

CANTIDAD / Frequency / Row Pct	muerto	defor	normal	Total
0	14 7.00	2 1.00	184 92.00	200
100	16 8.00	1 0.50	183 91.50	200
200	21 10.50	8 4.00	171 85.50	200
300	39 19.50	58 29.00	103 51.50	200
400	102 51.00	88 44.00	10 5.00	200
Total	192	157	651	1000

STATISTICS FOR TABLE OF CANTIDAD BY RESPUES

Statistic	DF	Value	Prob
Chi-Square	8	513.836	0.001
Likelihood Ratio Chi-Square	8	577.145	0.001
Mantel-Haenszel Chi-Square	1	310.552	0.001
Phi Coefficient		0.717	
Contingency Coefficient		0.583	
Cramer's V		0.507	

Sample Size = 1000

En el archivo de resultados (**C20-2.LST**) se observa que tanto el χ^2 (**Chi-Square**) como la razón de verosimilitud (**Likelihood Ratio Chi-Square**) son significativos, por lo que se concluye que la probabilidad de muerte o malformación de estos ratones no es la misma en la población de no tóxico que en las poblaciones de altas concentraciones del tóxico. Nótese que el porcentaje de muertos, malformados y normales en la población de no tóxico es, respectivamente, 7%, 1% y 92%, mientras que en la población de 400 *mg/Kg día* es 51%, 44% y 5%. Se puede, por tanto, afirmar que el efecto del tóxico ha sido positivo en el aumento de la mortalidad y malformación. Si bien esto último es una apreciación y, por el momento, no se puede hacer un análisis mas pormenorizado.

Riesgo relativo y su intervalo de confianza.-

Esta claro con los datos del ejemplo c20-1 que la proporción de individuos no vacunados que se vieron afectados por la enfermedad es considerablemente mayor que la proporción de individuos afectados y vacunados. En otras palabras, el riesgo de ser afectados si se ha vacunado es menor que el riesgo de ser afectado si no se ha vacunado.

Pero una diferencia fija en dos proporciones puede tener más importancia si ambas proporciones están próximas al 0 o al 1, que si ambas proporciones se encuentran próximas al 0.5. Por ejemplo, supóngase que en el ejemplo anterior, la proporción de los no vacunados que enfermaron fuera de 0.010 y la proporción de vacunados que enfermaron fuera de 0.001, esta diferencia es mucho más significativa que la misma diferencia numérica entre 0.410 y 0.401.

En dicho caso, la razón de las proporciones es también una útil medida descriptiva de asociación. Esta razón es la conocida como *riesgo relativo*, pues dará la probabilidad (riesgo) de que un individuo vacunado padezca la enfermedad con respecto a un individuo no vacunado, o viceversa. Genéricamente, si se tiene una variable predictiva y una variable de respuesta, el riesgo relativo se define como

$$RR = \frac{\frac{n_{11}}{n_{1.}}}{\frac{n_{21}}{n_{2.}}}$$

siendo

$n_{1.}$ el marginal fijo del primer nivel de la variable explicativa.
$n_{2.}$ el marginal fijo del segundo nivel de la variable explicativa.
n_{11} el valor de la primera respuesta dentro del primer nivel de la variable explicativa (se podía haber elegido la respuesta *b*).
n_{21} el valor de la primera respuesta dentro del segundo nivel de la variable explicativa (se podía haber elegido la respuesta *d*).

Esta razón puede ser cualquier número real no negativo (desde 0 a ∞). Un riesgo de 1/3 indica que los individuos pertenecientes al segundo nivel de la variable explicativa tienen una probabilidad 3 veces superior de manifestar la respuesta uno de la variable respuesta que los individuos que pertenecen al nivel primero de la variable explicativa; un riesgo relativo de 1 corresponde a la independencia entre las dos variables, es decir, no asociación o no riesgo de respuesta diferente para los diferentes niveles de la variable explicativa; y un riesgo relativo de 3 es que los individuos pertenecientes al primer nivel de la variable explicativa tienen una probabilidad 3 veces superior de manifestar la respuesta que los individuos pertenecientes al segundo nivel de la variable explicativa.

El intervalo de confianza del riesgo relativo sería

$$LC_{RR} = RR\ e^{\pm Z_{(\alpha/2)}\sqrt{V}}$$

siendo

$$V = Var\,(\ln\ RR) = \frac{1-P_{11}}{n_{11}} + \frac{1-P_{21}}{n_{21}}$$

Ejemplo.-

Calcúlese, en el ejemplo anterior, el riesgo relativo de que enfermen los vacunados con respecto a que enfermen los no vacunados

	Enfermos	*No enfermos*	*total*
Vacunados	11	39	50
No vacunados	23	27	50
total	34	66	100

$$RR = \frac{\frac{11}{50}}{\frac{23}{50}} = 0.478 = \frac{1}{2.09}$$

Esto indica que los vacunados tienen una probabilidad 0.478 inferior de contraer la enfermedad con respecto a los no vacunados, o lo que es lo mismo, los no vacunados tienen una probabilidad 2.09 veces superior de contraer la enfermedad con respecto a los vacunados.

Los límites de confianza sería

$$V = Var\,(\ln RR) = \frac{1-\frac{11}{50}}{11} + \frac{1-\frac{23}{50}}{23} = 0.0944$$

$$L_i = 0.478\, e^{-1.96\sqrt{0.0944}} = 0.262$$

$$L^s = 0.478\, e^{1.96\sqrt{0.0944}} = 0.873$$

Esto indica que los vacunados tienen, como mínimo, una probabilidad 0.262 inferior de contraer la enfermedad y, como máximo, tienen una probabilidad 0.873 inferior de contraer la enfermedad.

Si se desea el riesgo relativo de que no enfermen los vacunados con respecto a que no enfermen los no vacunados, este riesgo es

$$RR = \frac{\frac{39}{50}}{\frac{27}{50}} = 1.444$$

Lo que indica que la probabilidad de que no enfermen los no vacunados es 1.44 veces superior que la probabilidad de que no enfermen los no vacunados.

Los límites de confianza sería

$$V = Var(\ln RR) = \frac{1-\frac{39}{50}}{39} + \frac{1-\frac{27}{50}}{27} = 0.0227$$

$$L_i = 1.444\, e^{-1.96\sqrt{0.0227}} = 1.075$$

$$L^s = 1.444\, e^{1.96\sqrt{0.0227}} = 1.940$$

Esto indica que los vacunados tienen, como mínimo, una probabilidad 1.075 superior de no contraer la enfermedad y, como máximo, tienen una probabilidad 1.94 superior de no contraer la enfermedad.

Índice de desviación y su intervalo de confianza.-

El llamado *índice de desviación* (odds ratio) puede ser considerado como otra medida de riesgo relativo y viene dado por la expresión

$$ID = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

siendo

n_{11} el valor de la primera respuesta dentro del primer nivel de la variable explicativa.
n_{12} el valor de la segunda respuesta dentro del primer nivel de la variable explicativa.
n_{21} el valor de la primera respuesta dentro del segundo nivel de la variable explicativa.
n_{22} el valor de la segunda respuesta dentro del segundo nivel de la variable explicativa.

También se le denomina índice de *producto cruzado* puesto que es igual a la razón de los productos de las probabilidades de las dos diagonales.

El campo de variación de este índice es el mismo del riesgo relativo, este es, entre cero e infinito. Si el índice de desviación vale 4, significa que la desviación de la primera respuesta es cuatro veces mayor en el primer nivel de la variable explicativa que en el segundo nivel de la variable explicativa (esto no es lo mismo que decir que la probabilidad de la primera respuesta en el primer nivel de la variable explicativa sea cuatro veces mayor que la probabilidad de la misma probabilidad en el segundo nivel de la variable explicativa, pues esto es el $RR = 4$). Si el índice de desviación vale 1/4, significa que la desviación de la primera respuesta es cuatro veces menor en el primer nivel de la variable explicativa que en el segundo nivel de la variable explicativa. Un valor de uno significa no asociación o no desviación.

El índice de desviación no cambia si se cambia la tabla de contingencia de orientación.

El índice de desviación es el riesgo relativo cuando la tabla refleja el resultado

de un estudio retrospectivo, esto es, cuando la variable respuesta se ha medido antes que la variable explicativa; estos tipos de estudios son denominados de *caso-control*

El intervalo de confianza del índice de desviación sería

$$LC_{ID} = ID\ e^{\pm Z_{(\alpha/2)}\sqrt{V}}$$

siendo

$$V = Var\,(\ln ID) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Ejemplo.-

Calcúlese, en el ejemplo C20-1, el de la tabla de contingencia de tamaño 2x2, el índice de desviación de los vacunados que enfermaron

	Enfermos	*No enfermos*	*total*
Vacunados	11	39	50
No vacunados	23	27	50
total	34	66	100

$$ID = \frac{11\times 27}{23\times 39} = 0.3311$$

Esto indica que la desviación de los enfermos vacunados es 0.331 veces inferior que la desviación de los enfermos no vacunados

Los límites de confianza sería

$$V = Var\,(\ln\ ID) = \frac{1}{11} + \frac{1}{39} + \frac{1}{23} + \frac{1}{27} = 0.1971$$

$$L_i = 0.331\,e^{-1.96\sqrt{0.1971}} = 0.1387$$

$$L^s = 0.331\,e^{1.96\sqrt{0.1971}} = 0.7904$$

Esto indica que la desviación de los enfermos vacunados es, como mínimo, 0.1387 inferior que los enfermos no vacunados y la desviación de los enfermos vacunados es, como máximo, 0.7904 inferior a la de los enfermos no vacunados.

Archivo del programa *SAS* (C20-3.SAS).-

```
title 'Riesgo relativo e índice de desviación';
options ls=80;
data chi2;
infile 'c20-1.dat';
input vacunado $ respues $ n @@;
proc freq order=data;
weight n;
table vacunado * respues / measures nocol nopercent;
run;
```

Archivo de resultados (C20-3.LST) .-

```
                Riesgo relativo e índice de desviación

                    TABLE OF VACUNADO BY RESPUES

 VACUNADO      RESPUES
 Frequency|
 Row Pct  |enfermo |noenfer |  Total
 ---------+--------+--------+
 si       |     11 |     39 |     50
          |  22.00 |  78.00 |
 ---------+--------+--------+
 no       |     23 |     27 |     50
          |  46.00 |  54.00 |
 ---------+--------+--------+
 Total          34       66      100

       Estimates of the Relative Risk (Row1/Row2)
                                                95%
Type of Study            Value      Confidence Bounds
-------------------------------------------------------------
Case-Control             0.331        0.139       0.790
Cohort (Col1 Risk)       0.478        0.262       0.873
Cohort (Col2 Risk)       1.444        1.075       1.940

Sample Size = 100
```

Intervalos de confianza para las casillas.-

Los datos presentados en una tabla de contingencia de un muestreo aleatorio simple estratificado son las frecuencias de los diferentes niveles de una variable respuesta observadas en dos poblaciones (las dos filas), por tanto, estas frecuencias observadas son las estimas de las medias de las respuestas en las dos poblaciones, por lo que esta estima tiene su error típico.

El error típico de la proporción observada en la casilla n_{11} es

$$ET_{11} = \sqrt{\frac{\frac{n_{11}}{n_{1.}} - \left(\frac{n_{11}}{n_{1.}}\right)^2}{n_{1.}}}$$

y el de la casilla n_{32} es

$$ET_{32} = \sqrt{\frac{\frac{n_{32}}{n_{3.}} - \left(\frac{n_{32}}{n_{3.}}\right)^2}{n_{3.}}}$$

siendo

n_{11} el valor de la primera respuesta dentro del primer nivel de la variable explicativa.
n_{32} el valor de la segunda respuesta dentro del tercer nivel de la variable explicativa.
$n_{1.}$ el valor total de la primera población o variable explicativa.
$n_{3.}$ el valor total de la tercera población o variable explicativa.

Por lo que el intervalo de confianza para estas casillas son, respectivamente

$$LC_{11} = \frac{n_{11}}{n_{1.}} \pm ET_{11} Z_{(\alpha/2)}$$

$$LC_{32} = \frac{n_{32}}{n_{3.}} \pm ET_{32} Z_{(\alpha/2)}$$

Ejemplo.-

Calcúlese los límites de confianza de los ejemplo C20-1 y C20-2:

a) Ejemplo de la tabla 2x2, la probabilidad de enfermar estando vacunado.

	Enfermos	*No enfermos*	*total*
Vacunados	11	39	50
No vacunados	23	27	50
total	34	66	100

$$ET_{11} = \sqrt{\frac{\frac{11}{50} - \left(\frac{11}{50}\right)^2}{50}} = 0.05858$$

$$L_i = \frac{11}{50} - 1.96 \times 0.05858 = 0.1052$$

$$L^s = \frac{11}{50} + 1.96 \times 0.05858 = 0.3348$$

Con lo que se concluye que, con un 95% de confianza, la proporción mínima de individuos vacunados que enferma es de 0.1052 y la proporción máxima es de 0.3348.

b) Ejemplo de la tabla 5x2, la probabilidad de morir habiendo recibido una dosis de 300 *mg/Kg*

mg/Kg día	*Muertos*	*Malformados*	*Normales*	*total*
Control	14	2	184	200
100	16	1	183	200
200	21	8	171	200
300	39	58	103	200
400	102	88	10	200
total	192	157	651	1000

$$ET_{41} = \sqrt{\frac{\frac{39}{200} - \left(\frac{39}{200}\right)^2}{200}} = 0.0280$$

$$L_i = \frac{39}{200} - 1.96 \times 0.0280 = 0.1401$$

$$L^s = \frac{39}{200} + 1.96 \times 0.0280 = 0.2499$$

Con lo que se concluye que, con un 95% de confianza, la proporción mínima de individuos que mueren recibiendo una dosis de 300 *mg/Kg* es de 0.1401 y la proporción máxima es de 0.2499.

Modelos lineales.-

Como ya se indicó en el epígrafe *Análisis de las tablas de contingencia de dos vías*, se puede hacer inferencia de dichas tablas utilizando los modelos lineales o loglineales. Se dijo, así mismo, que si es un muestreo aleatorio simple (Modelo I) se analiza mediante el modelo loglineal, mientras que si es un muestreo aleatorio simple estratificado (Modelo II) se analiza mediante el modelo lineal.

Como los ejemplos que se han utilizado hasta el momento correspondía a un muestreo aleatorio simple estratificado, se comenzará estudiando los modelos lineales para analizar con ellos este mismo ejemplo.

Se utilizan modelos con objeto de reducir (y por lo tanto explicar) el conjunto de datos que se tienen a un pequeño conjunto de números significativos (estimas de parámetros). La utilización de modelos facilitará la respuesta a cuestiones de asociación, mejorando la probabilidad de tomar las decisiones adecuadas.

Supóngase que se quiere estudiar la asociación entre dos variables, *A* y *B*, y que denominamos como **Y** el conjunto de los valores de la variable *A* y como **X** el conjunto de los valores de la variable *B* y como β el conjunto de parámetros. Entonces, se puede contemplar estos datos en el siguiente modelo.

Observaciones de la variable A	=	*Combinación de:* 1. *Parámetros* 2. *Observaciones de la variable B*	+	*Variaciones aleatorias que solapan las de la asociación*

$$Y = f(X, \beta) + \varepsilon$$

Esta parte contiene toda la información pertinente a la asociación entre las dos variables

El utilizar modelos tiene muchas ventajas, algunas de éstas son:

- Es fácil controlar muchas variables; simplemente se incluyen en el modelo.
- Los parámetros representan medidas de asociación que pueden estimarse.
- Se pueden estimar muchas asociaciones simultáneamente.
- Los parámetros estimados pueden usarse para predecir valores de la variable *Y* en función exclusiva de la asociación.
- Hay relativamente pocos parámetros y todas las cuestiones sobre la asociación son aplicables a cuestiones sobre los parámetros.

En el modelo anterior se ha dividido la variabilidad de la respuesta en dos espacios:

- Espacio de la variación debida a la asociación entre ambas variables (espacio del modelo) y
- Espacio de la variación debida a las desviaciones aleatorias (espacio del error).

El espacio del modelo es para la variación que es significativamente diferente de cero y, esta variación, se divide en las debidas a diferentes fuentes de variación, denominadas efectos. Cada efecto corresponde a uno o más parámetros del modelo.

El espacio del error es para la variación estadísticamente igual a cero. Recoge la variación residual, esto es, la variación que queda después de medida la variación debida al modelo.

El modelo más simple (para solo dos variables) sería

$$Y = \alpha + \beta X$$

donde α es la ordenada en el origen y representa la media de la respuesta, y β el incremente o el decremento, con respecto a la media (ordenada en el origen) de la respuesta en la primera población.

Si hubiera mas variables, por ejemplo tres, *A*, *B* y *C*, el modelo sería

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

que puesto en términos de análisis de la varianza, sería

$$A = \mu + B + C + BC$$

Ejemplo.-

Resuélvase por medio del modelo lineal los ejemplos C20-1 y C20-2

Archivo del programa *SAS* (C20-4.SAS).-

```
title 'Modelo lineal para tabla 2x2';
options ls=75 ps=60;
data mod_line;
infile 'c20-1.dat';
input vacunado $ respues $ n @@;
proc catmod order=data;
 weight n;
 response marginals;
 model respues = vacunado / predict;
run;
```

Archivo de resultados (C20-4.LST).-

```
                    Modelo lineal para tabla 2x2

                          CATMOD PROCEDURE
Response: RESPUES                      Response Levels (R)=     2
Weight Variable: N                     Populations     (S)=     2
Data Set: MOD_LINE                     Total Frequency (N)=   100
Frequency Missing: 0                   Observations  (Obs)=     4

                       POPULATION PROFILES
                                    Sample
                    Sample  VACUNADO  Size
                    ---------------------
                        1      si       50
                        2      no       50

                        RESPONSE PROFILES
                        Response  RESPUES
                        ----------------
                             1    enfermo
                             2    noenfer

                                    DESIGN
                          Response  MATRIX
                 Sample   Function    1       2
                 ---------------------------------
                     1     0.22000     1       1
                     2     0.46000     1      -1

                   ANALYSIS-OF-VARIANCE TABLE
          Source                    DF   Chi-Square      Prob
          ---------------------------------------------------
          INTERCEPT                  1        55.05    0.0000
          VACUNADO                   1         6.86    0.0088
          RESIDUAL                   0            .         .
```

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.3400	0.0458	55.05	0.0000
VACUNADO	2	-0.1200	0.0458	6.86	0.0088

PREDICTED VALUES FOR RESPONSE FUNCTIONS

Sample	Function Number	Observed Function	Observed Standard Error	Predicted Function	Predicted Standard Error	Residual
1	1	0.22	0.05858327	0.22	0.05858327	0
2	1	0.46	0.07048404	0.46	0.07048404	0

El archivo de resultados (**C20-4.LST**) nos informa que hay dos poblaciones, la primera de *si* vacunados con tamaño de muestra 50 y la segunda de *no* vacunados y tamaño de muestra 50. La variable respuesta tiene dos niveles, el primero de *enfermo* y el segundo de *noenfer*. La proporción de la respuesta uno en la población uno es de 0'22 y la proporción de la respuesta uno en la población dos es de 0'46.

La ANALYSIS OF VARIANCE TABLE nos muestra que el efecto de la variable VACUNADO es significativo (P<0'01), esto es, que la proporción de enfermos es diferente para los diferentes niveles del factor VACUNADO (si vacunado o no vacunado). La media (INTERCEPT) también es estadísticamente diferente de cero. Como se ha introducido en el modelo todos los efectos que pueden intervenir no queda residuo.

En la tabla ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES, se muestra que el valor medio de la respuesta (INTERCEPT) es de 0'34 (0'22+0'46/2) que es significativamente diferente de cero (P<0'001). Y muestra, asimismo que los vacunados enfermos tienen una desviación negativa (-0'12) con respecto a la media, y esta desviación es estadísticamente diferente de cero (P<0'01), esto indica que hay menos vacunados enfermos que no vacunados enfermos.

Por lo tanto la conclusión es la misma que con el análisis del χ^2 y de la G, esta es que la probabilidad de enfermar de estos ratones no es la misma en la población de vacunados que en la población de no vacunados. La desviación, con respecto a la media, de los vacunados enfermos es significativamente negativa, lo que indica que hay menos enfermos vacunados, por lo que se puede concluir que la vacuna protege de la enfermedad.

Archivo del programa *SAS* (C20-5.SAS).-

```
title 'Modelo lineal de datos categóricos';
options ls=75 ps=60;
data mod_line;
infile 'c20-2.dat';
input cantidad $ respues $ n @@;
proc catmod order=data;
 weight n;
 response marginals;
 model respues = cantidad ;
run;
```

Archivo de resultados (C20-5.LST).-

Modelo lineal de datos categóricos

CATMOD PROCEDURE

Response: RESPUES
Weight Variable: N
Data Set: MOD_LINE
Frequency Missing: 0

Response Levels (R)= 3
Populations (S)= 5
Total Frequency (N)= 1000
Observations (Obs)= 15

POPULATION PROFILES

Sample	CANTIDAD	Sample Size
1	0	200
2	100	200
3	200	200
4	300	200
5	400	200

RESPONSE PROFILES

Response	RESPUES
1	muerto
2	defor
3	normal

Sample	Response Functions 1	2	DESIGN MATRIX 1	2	3	4	5
1	0.07000	0.01000	1	1	0	0	0
2	0.08000	0.00500	1	0	1	0	0
3	0.10500	0.04000	1	0	0	1	0
4	0.19500	0.29000	1	0	0	0	1
5	0.51000	0.44000	1	-1	-1	-1	-1

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	2	1069.99	0.0000
CANTIDAD	8	2048.53	0.0000
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.1920	0.0113	288.20	0.0000
	2	0.1570	0.0101	243.77	0.0000
CANTIDAD	3	-0.1220	0.0180	46.05	0.0000
	4	-0.1470	0.0114	165.19	0.0000
	5	-0.1120	0.0187	35.97	0.0000
	6	-0.1520	0.0108	199.10	0.0000
	7	-0.0870	0.0202	18.47	0.0000
	8	-0.1170	0.0147	63.28	0.0000
	9	0.00300	0.0245	0.02	0.9024
	10	0.1330	0.0268	24.61	0.0000

El archivo de resultados (**C20-5.LST**) nos informa que hay cinco poblaciones y tres respuestas. La proporción de la respuesta uno (*Muertos*) en la población uno es de 0.07 y la proporción de la respuesta dos (*Deformes*) en la población uno es de 0.01, *etc.*

La ANALYSIS OF VARIANCE TABLE nos muestra que el efecto de la variable CANTIDAD es significativo ($P<0.001$), esto es, que la proporción de muertos y

deformes es diferente para los diferentes niveles del factor CANTIDAD (las diferentes concentraciones del tóxico). La media (INTERCEPT) también es estadísticamente diferente de cero. Como se ha introducido en el modelo todos los efectos que pueden intervenir no queda residuo.

En la tabla ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES, se muestra que el valor medio de la respuesta *Muerto* (INTERCEPT) es de 0.192 (0.07+0.08+0.105+0.195+0.51/5) que es significativamente diferente de cero (P<0.001) y el valor medio de la respuesta *Deform* (INTERCEPT) es de 0.157 (0.01+0.005+0.04+0.29+0.44/5) que es significativamente diferente de cero (P<0.001).

Y se muestra, asimismo que en la población uno (concentración de 0 *mg/Kg*), la desviación de los *Muertos* con respecto a la media de muertos es significativamente (P<0.001) negativa (-0.122) y la desviación de *Deformes* con respecto a su media es significativamente (P<0.001) negativa (-0.147).

Las desviaciones de todas las poblaciones (o concentraciones) son significativamente negativas hasta llegar a la concentración del tóxico de 300 *mg/Kg* que se hacen significativamente positivas. La única desviación que no es estadísticamente diferente de la media es la de la respuesta uno (*Muertos*) en la población tres (concentración del tóxico de 300 *mg/Kg*) que tiene una desviación de 0.003 que no es estadísticamente diferente de cero (P>0.005). Esto indica que las concentraciones del tóxico inferiores a 300 *mg/Kg* produce muertos y malformaciones por debajo de la media, mientras que las concentraciones iguales o superiores a 300 *mg/Kg* producen muertos y malformaciones por encima de la media.

La conclusión es la misma que con el análisis del χ^2 y de la G, esta es que la probabilidad de muerte o malformación de estos ratones no es la misma en la población de poca concentración del tóxico que en la población de mayor concentración del tóxico. A menor concentración del tóxico se producen significativamente menos muertes y malformaciones que a mayor concentración del tóxico. Se puede, por tanto, afirmar que el efecto de la concentración ha sido positivo en el aumento de la manifestación de las muertes y malformaciones de los fetos.

Análisis de Correspondencias.-

Concebido originariamente para el estudio de tablas de contingencia, se ha revelado eficaz para el estudio de cualquier matriz de números no negativos. Entre los objetivos de este análisis se puede destacar:

(a) Resumir la información de una tabla de contingencia describiendo sintéticamente pautas de relaciones entre variables y categorías que sería difícil de entresacar directamente de una tabla grande. Para ello se hace una reducción del espacio factorial, tal como se hace con el Análisis de Componentes Principales, condensándose el máximo de información en uno o pocos factores. En la representación gráfica de los resultados, las categorías similares aparecen juntas por lo que es fácil contestar a preguntas como ¿qué categorías de una variable son más similares y podría unirse en unas sola? ¿Cuáles son las categorías más diferentes?

¿Cuál es la asociación entre las filas y las columnas de la tabla de contingencia?

(b) Las variables resumen o factores que calcula el análisis de correspondencias son cuantitativas, por lo que en cierto sentido es un método que cuantifica datos cualitativos. Los factores extraídos de este análisis pueden utilizarse como variable de entrada en otros análisis.

(c) Como ocurre con el análisis de Componentes Principales, los factores suelen ser interpretables, tienen su nombre y representan las dimensiones que más diferencian al colectivo de individuos en cuanto a las características consideradas.

(d) Aunque es un método descriptivo, también se pueden probar los diferentes factores, por lo que también es un método inferencial.

El número de factores a elegir nos lo da la prueba χ^2 de cada valor propio. Los grados de libertad para el *i-ésimo* valor propio es

$$gl_i = F + C - (2i - 1)$$

teniendo en cuenta que se comienza por el segundo valor propio, pues el primer valor propio no se considera al valer la unidad, solo se consideran los que son inferiores a 1. Por lo que se tiene

Valor propio	χ^2	*gl*
λ_2	$\chi^2_2 = N\lambda_2$	$F + C - 3$
λ_3	$\chi^2_3 = N\lambda_3$	$F + C - 5$
...	...	...
λ_c	$\chi^2_c = N\lambda_c$	$F + C - (2c-1)$
Total	χ^2	(F-1)(C-1)

Ejemplo.-

Resuélvase por medio del Análisis de Correspondencias el ejemplo C20-2

Archivo del programa *SAS* (C20-6.SAS).-

```
title 'Análisis de correspondencias';
Options ls=75 ps=30;
Data chi2;
Infile 'c20-2.dat';
Input cantidad $ respues $ n @@;
Proc corresp out=corr;
Tables cantidad, respues;
Weight n;
run;
proc plot;
plot dim2 * dim1 = _name_;
run;
```

Archivo de resultados (C20-6.LST).-

```
                    Análisis de correspondencias

                The Correspondence Analysis Procedure

                Inertia and Chi-Square Decomposition
Singular  Principal Chi-
Values    Inertias  Squares Percents   20   40   60   80  100
                                    ----+----+----+----+----+---
0.70859   0.50209   502.094  97.71% ************************
0.10836   0.01174    11.742   2.29% *
          -------   -------
          0.51384   513.836 (Degrees of Freedom = 8)

                        Row Coordinates
                             Dim1          Dim2
                0          -0.57071      -0.02458
                100        -0.56305      -0.05027
                200        -0.43491      -0.03731
                300         0.31093       0.21083
                400         1.25774      -0.09868

               Summary Statistics for the Row Points
                     Quality          Mass         Inertia
          0          1.00000       0.200000       0.127010
          100        1.00000       0.200000       0.124379
          200        1.00000       0.200000       0.074163
          300        1.00000       0.200000       0.054931
          400        1.00000       0.200000       0.619516

        Partial Contributions to Inertia for the Row Points
                               Dim1          Dim2
                  0         0.129740      0.010291
                  100       0.126281      0.043039
                  200       0.075343      0.023706
                  300       0.038509      0.757112
                  400       0.630126      0.165852

           Indices of the Coordinates that Contribute
               Most to Inertia for the Row Points
                        Dim1          Dim2          Best
          0                1             0             1
          100              1             0             1
          200              0             0             1
          300              0             2             2
          400              1             1             1

               Squared Cosines for the Row Points
                               Dim1          Dim2
                  0         0.998148      0.001852
                  100       0.992092      0.007908
                  200       0.992695      0.007305
                  300       0.685029      0.314971
                  400       0.993882      0.006118

                        Column Coordinates
                             Dim1          Dim2
                defor        1.11041       0.18497
                muerto       0.84002      -0.18142
                normal      -0.51554       0.00890

               Summary Statistics for the Column Points
                     Quality          Mass         Inertia
          defor      1.00000       0.157000       0.387196
          muerto     1.00000       0.192000       0.275968
          normal     1.00000       0.651000       0.336836
```

```
        Partial Contributions to Inertia for the Column Points
                                  Dim1          Dim2
                    defor      0.385554      0.457446
                    muerto     0.269836      0.538164
                    normal     0.344611      0.004389

                Indices of the Coordinates that Contribute
                   Most to Inertia for the Column Points
                              Dim1          Dim2          Best
              defor              2             2             2
              muerto             2             2             2
              normal             1             0             1

                    Squared Cosines for the Column Points
                                  Dim1          Dim2
                    defor      0.973002      0.026998
                    muerto     0.955436      0.044564
                    normal     0.999702      0.000298

                 Plot of DIM2*DIM1.  Symbol is value of _NAME_.
 0.4 +
     |
     |
     |
     |
 0.2 +                                    3                    d
     |
     |
DIM2 |
     |
 0.0 +                n
     |              0   2
     |              1
     |                                                          4
     |
     |                                          m
-0.2 +
     --+------------+------------+------------+------------+------------+--
      -1.0         -0.5          0.0          0.5          1.0          1.5
                                      DIM1
NOTE: 1 obs had missing values.
```

El archivo de resultados (**C20-6.LST**) nos muestra primeramente los valores propios, sus inercias, sus valores χ^2 y el porcentaje explicado del χ^2 total.

El χ^2 total vale 513.836 (igual que con los dos métodos anteriores) que es significativo par gl=8 al 0.05, por lo que existe asociación entre las dosis de tóxico y las muertes y malformaciones.

El χ^2 de la primera dimensión, correspondiente al segundo valor propio vale 502.094, que es 97.71% del χ^2 total. Este χ^2 tiene 5+3-3=5 grados de libertad, por lo que es significativo al 0.05. El χ^2 de la segunda dimensión vale 11.742, que constituye el 2.29% del χ^2 total y tiene 5+3-5=3 grados de libertad, por lo que también es significativo y se puede, por tanto, representar las dos dimensiones.

Las siguientes salidas del procedimiento **CORRESP** son los valores de las coordenadas que las tomaran el siguiente procedimiento, **PLOT**, para representar las filas y las columna y observar las causas de la asociación entre ellas

En la salida del procedimiento **PLOT**, que es la representación de las dos primeras dimensiones del análisis de correspondencias, se observa, con respecto a la primera dimensión, que los individuos normales se asocian a las dosis 0, 100 y 200, mientras que los individuos muertos y deformes se asocian a las dosis 300 y 400. Y observando la segunda dimensión, que también es significativa, se observa los individuos deformes están asociados a la dosis 300 y los individuos muertos están asociados a la dosis 400.

Por lo tanto, la conclusión de este análisis es que existe asociación entre las dosis de este compuesto y el número de malformaciones y muertes embrionarias, en el sentido de que las dosis 0, 100 y 200 no producen ni muertes ni deformaciones, la dosis 300 produce deformaciones y la dosis 400 produce muertes del embrión.

Partición de los grados de libertad en tablas de contingencia.-

En el caso de una tabla de 2x2, la interpretación estadística resultante de una prueba χ^2 es clara y fácil de ver observando la tabla ya que no hay más que fijarse en una de las dos proporciones que contiene la tabla. La interpretación, sin embargo, no es tan fácil con tablas de contingencia que tengan más de un grado de libertad pues se trata de comparar más de dos proporciones, y aunque el χ^2 significativo general indica que estas proporciones son heterogéneas, se requiere un análisis más detallado para decidir justamente dónde ocurren las diferencias significativas. Ese análisis más detallado lo provee los modelos lineales, pero aún así existe otra cuestión que no se ha resuelto aún, esta es la siguiente

Si se tiene una tabla mayor de 2x2 puede ocurrir que las pruebas de asociación den significativas pero que no sean significativas las diferencias entres todas las filas o poblaciones de la tabla. O puede ocurrir todo lo contrario, que las pruebas de asociación general den no significativas cuando si puede ser significativa la diferencia entre dos de las varias poblaciones que contiene la tabla.

Para resolver este problema se requiere la subdivisión del valor general del χ^2 en componentes aditivos, o expresado de otra forma, se va a dividir los grados de libertad en que se basa el valor del χ^2 general. Se puede demostrar que el valor general de χ^2 para una tabla de contingencia puede siempre ser dividido en tantos componentes como grados de libertad tiene la tabla. Una consecuencia interesante de estos procedimientos es que frecuentemente puede suceder que un χ^2 general, que es no significativo, da uno o más componentes que son significativos, o viceversa, por tanto se ve incrementada la sensibilidad de la prueba.

Como se puede comprobar, esto es, conceptualmente, semejante a las *pruebas planeadas* vistas en el análisis de la varianza, para ello se van a utilizar los modelos lineales categóricos con el estamento **CONTRASTS** que se utiliza en el mismo sentido que las pruebas planeadas o contrastes ortogonales del análisis de la varianza, vistas en el Capítulo 10.

Ejemplo.-

Cuando el ejemplo anterior se resolvió con el análisis de correspondencias se llego a la conclusión, vista la representación gráfica de los dos primeros factores, de que las dosis 0, 100 y 200 eran iguales y no producen ni muertes ni deformaciones, la dosis 300 produce deformaciones y la dosis 400 produce muertes del embrión. Veamos como se puede probar estadísticamente si esta apreciación gráfica es correcta o no

Archivo del programa *SAS* (C20-7.SAS).-

```
title 'Contrastes Ortogonales';
options ls=75 ps=60;
data chi2;
infile 'c20-2.dat';
input cantidad $ respues $ n @@;
proc catmod order=data;
 weight n;
 response marginals;
 model respues = cantidad ;
  contrast '  0 vs 100' cantidad -1  1  0  0;
  contrast '  0 vs 200' cantidad -1  0  1  0;
  contrast '  0 vs 300' cantidad -1  0  0  1;
  contrast '  0 vs 400' cantidad -2 -1 -1 -1;
  contrast '100 vs 200' cantidad  0 -1  1  0;
  contrast '100 vs 300' cantidad  0 -1  0  1;
  contrast '100 vs 400' cantidad -1 -2 -1 -1;
  contrast '200 vs 300' cantidad  0  0 -1  1;
  contrast '200 vs 400' cantidad -1 -1 -2 -1;
  contrast '300 vs 400' cantidad -1 -1 -1 -2;
run;
```

Los coeficientes de los contrastes se obtienen a partir de la matriz de diseño (**DESIGN MATRIX**) que obtuvimos cuando resolvimos este ejemplo por medio de los modelos lineales. Estos coeficientes es el resultado de restar los coeficientes de dicha matriz para las categorías contrastadas.

Archivo de resultados (C20-7.LST).-

ANALYSIS OF CONTRASTS

Contrast	DF	Chi-Square	Prob
0 vs 100	2	0.47	0.7906
0 vs 200	2	5.55	0.0623
0 vs 300	2	111.05	0.0000
0 vs 400	2	1284.18	0.0000
100 vs 200	2	6.64	0.0362
100 vs 300	2	111.29	0.0000
100 vs 400	2	1246.01	0.0000
200 vs 300	2	70.80	0.0000
200 vs 400	2	769.75	0.0000
300 vs 400	2	153.35	0.0000

En el archivo de resultados (**C20-7.LST**), lo nuevo con respecto al anterior (**C20-5.LST**) es la salida de los contrastes, lo demás del archivo de resultados ya se conoce. Las pruebas de los contrastes nos confirma que las dosis 0 (o control), la dosis 100 y la 200 son estadísticamente iguales, y éstas son diferentes de la dosis 300 y 400, y éstas

dos son diferentes entre si.

Pruebas de independencia o asociación en los muestreos aleatorios simples.-

Como se dijo en un epígrafe anterior, la prueba que se desarrollo específicamente para los muestreos aleatorios simple (Modelo I) es la prueba *G*, si bien la prueba χ^2 es general y no esta desarrollada para algún modelo concreto, por lo que se pueden usar ambas. La mecánica de cálculo e interpretación es exactamente la misma que la vista para el Modelo II, por lo que para no repetir lo mismo con diferentes tablas, remito al lector al epígrafes *Análisis de las tablas de contingencia de dos vías* y sucesivos donde se exponen estas pruebas. Sin embargo la utilización de los modelos no es igual pues para estos muestreos lo indicado es la utilización de los modelos *loglineales.*

Modelos loglineales.-

En los modelos loglineales el punto de vista general es que todas las variables son dependientes por tanto la idea de independencia estadística está implícita. Las ideas básicas de estos modelo son

- Particiona la variabilidad observada en fuentes de variación que miden la dependencia o independencia estadística de las dos (o más) variables categóricas.
- Si la fuente refleja independencia estadística entre las variables, entonces la variación es estadísticamente no significativa, por tanto la variación puede ser puesta en el espacio del error.
- Si la fuente refleja dependencia estadística entre las variables, entonces la variación es estadísticamente significativa, por tanto es retenida en el espacio del modelo.

La manera de medir la variación de la dependencia es mediante el logaritmo de la probabilidad, en lugar de la probabilidad misma, es por ello por lo que se denomina modelo *loglineal.*

Sea dos variables, *A* y *B* categóricas que se representan en una tabla de contingencia con *i* niveles de la variable *A* y *j* niveles de la variable *B*. Sea π_{ij} la probabilidad conjunta de las dos respuestas categóricas (la probabilidad de la casilla *ij*). Estas respuestas serán estadísticamente independientes si

$$\pi_{ij} = \pi_{i.}\ \pi_{.j}$$

es decir, si la probabilidad observada en una casilla es igual al producto de las probabilidades marginales. En ese caso la expresión que nos define las frecuencias observadas en las casillas es

$$m_{ij} = N\,\pi_{ij} = N\,\pi_{i.}\ \pi_{.j}$$

Se puede representar la independencia entre ambas variables en un modelo

lineal (loglineal) aditivo, en escala logarítmica utilizando m_{ij} y de π_{ij}, de la siguiente manera

$$\ln m_{ij} = \ln N + \ln \pi_{i.} + \ln \pi_{.j}$$

Que viene a indicar que el **ln** de la frecuencia esperada en una casilla es función aditiva del efecto de la fila y del efecto de la columna en la que se encuentra dicha casilla.

Este modelo se puede poner de la siguiente manera

$$\ln m_{ij} = \mu + \mu_i^A + \mu_j^B$$

siendo

$$\mu = \frac{\sum_{ij} \ln m_{ij}}{i \quad j}$$

$$\mu_i^A = \frac{\sum_i \ln m_{ij}}{i} - \mu$$

$$\mu_j^B = \frac{\sum_j \ln m_{ij}}{j} - \mu$$

Como se observa en las ecuaciones anteriores, μ es la media total de los $\ln m_{ij}$, mientras que μ^A_i es la desviación, con respecto a la media, del *i-ésimo* nivel de la variable *A*, esto es, es el efecto de la *i-ésima* fila o nivel de la variable *A* y μ^B_j es el efecto de la *j-ésima* columna de la variable *B*.

Este modelo es un loglineal modelo de *independencia* de una tabla de contingencia de dos vías. Pero si existe *dependencia* entre ambas variable, las frecuencias observadas en cada casilla no viene determinada por los totales marginales, de manera que

$$N \pi_{ij} \neq N \pi_{i.} \ \pi_{.j} = m_{ij}$$

por lo que el modelo correcto sería, en este caso,

$$\ln m_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB}$$

siendo el nuevo sumando

$$\mu_{ij}^{AB} = \ln m_{ij} - \mu_i^A - \mu_j^B + \mu$$

Este sumando refleja la dependencia entre ambas variable, *A, B*.

Si las variables *A* y *B* tienen dos niveles, se tiene,

$$\mu_1^A = -\mu_2^A$$
$$\mu_1^B = -\mu_2^B$$
$$\mu_{11}^{AB} = \mu_{22}^{AB} = -\mu_{12}^{AB} = -\mu_{21}^{AB}$$

Por lo que

$$\ln m_{11} = \mu + \mu_1^A + \mu_1^B + \mu_{11}^{AB}$$
$$\ln m_{12} = \mu + \mu_1^A - \mu_1^B - \mu_{11}^{AB}$$
$$\ln m_{21} = \mu - \mu_1^A + \mu_1^B - \mu_{11}^{AB}$$
$$\ln m_{22} = \mu - \mu_1^A - \mu_1^B + \mu_{11}^{AB}$$

Siendo, en este caso, la matriz de transformación o *respuesta*

	1	2	3	4
1	1	1	1	1
2	1	1	-1	-1
3	1	-1	1	-1
4	1	-1	-1	1

Pero, aunque hay cuatro casillas o probabilidades, solo hay tres grados de libertad, pues la última probabilidad viene obligada a que la suma de las cuatro valga uno. Por lo que se pueden referir todas las probabilidades o frecuencias a una de ellas, por ejemplo, se pueden referir a la última, es decir a m_{22}, en este caso se tiene

$$\ln m_{11} - \ln m_{22} = 2\ \mu_1^A + 2\ \mu_1^B + 0\ \mu_{11}^{AB}$$
$$\ln m_{12} - \ln m_{22} = 2\ \mu_1^A + 0\ \mu_1^B - 2\ \mu_{11}^{AB}$$
$$\ln m_{21} - \ln m_{22} = 0\ \mu_1^A + 2\ \mu_1^B - 2\ \mu_{11}^{AB}$$

Siendo, en este caso, la matriz de *diseño*

	1	2	3
1	2	2	0
2	2	0	-2
3	0	2	-2

Por lo que el modelo puede escribirse de forma matricial de la siguiente forma,

$$\begin{bmatrix} \ln m_{11} - \ln m_{22} \\ \ln m_{12} - \ln m_{22} \\ \ln m_{21} - \ln m_{22} \end{bmatrix} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 0 & -2 \\ 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} \mu_1^A \\ \mu_1^B \\ \mu_{11}^{AB} \end{bmatrix}$$

lo que muestra que un modelo loglineal es equivalente a un modelo lineal.

Ejemplo.-

Supóngase que en una población humana cuyos miembros se encuentran igualmente expuestos a la infección de un virus que produce una enfermedad de carácter leve, se ha vacunado un porcentaje significativo de individuos. Pasada la estación del año en la que se manifiesta la epidemia, se toma una muestra al azar de 148 individuos y se registra el número de vacunados y no vacunados que escaparon a la infección, obteniéndose los siguientes valores

	Enfermos	*No enfermos*	*total*
Vacunados	11	35	46
No vacunados	48	54	102
total	59	89	148

Es decir, que de 148 individuos muestreados, 46 estaban vacunados y de éstos, 11 enfermaron; en total enfermaron 59.

Archivo del programa *SAS* (C20-8.SAS).-

```
title 'Modelo loglineal en tabla 2x2';
options ls=75 ps=60;
data mod_line;
infile 'c20-8.dat';
input vacunado $ enfermo $ n @@;
proc catmod order=data;
 weight n;
 model vacunado * enfermo =_response_ / wls ml;
 loglin vacunado enfermo vacunado*enfermo;
run;
```

Como se ve en el programa ***SAS***, se ha realizado el análisis del modelo loglineal por el método de *mínimos cuadrados ponderados* (**WLS**) y por el método de *máxima verosimilitud* (**ML**) si bien, al haber explicitado en el estamento **LOGLIN** el modelo completo (modelo saturado) ambos métodos dan el mismo resultado. Si no se especifica ningún método, el SAS realiza por defecto el método de máxima verosimilitud,.

Como en este tipo de diseño ambas variables son dependiente y en un modelo hay que poner en el miembro de la derecha la variable independiente, se pone como variable independiente la respuesta, esto es **_RESPONSE_** , y se utiliza el estamento **LOGLIN** para especificar los efectos que se quieren estudiar; estos efectos comprenden el efecto de la **_RESPONSE_** en el modelo. Como se ve, en **LOGLIN** se ha especificado el modelo completo, esto es, que se analicen los efectos de los dos factores (vacunado y enfermo) por separado y la asociación entre estos dos factores (vacunado*enfermo).

Archivo de datos (C20-8.DAT).-

```
si si 11    si no 35
no si 48    no no 54
```

Archivo de resultados (C20-8.LST).-

Modelo loglineal en tabla 2x2

CATMOD PROCEDURE

Response: VACUNADO*ENFERMO	Response Levels (R)= 4
Weight Variable: N	Populations (S)= 1
Data Set: MOD_LINE	Total Frequency (N)= 148
Frequency Missing: 0	Observations (Obs)= 4

Sample	Sample Size
1	148

RESPONSE PROFILES

Response	VACUNADO	ENFERMO
1	si	si
2	si	no
3	no	si
4	no	no

RESPONSE MATRIX

	1	2	3
1	1	1	1
2	1	-1	-1
3	-1	1	-1
4	-1	-1	1

Sample	Function Number	Response Function	DESIGN MATRIX 1	2	3
1	1	-1.59109	2	2	0
	2	-0.43364	2	0	-2
	3	-0.11778	0	2	-2

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
VACUNADO	1	22.89	0.0000
ENFERMO	1	10.24	0.0014
VACUNADO*ENFERMO	1	6.81	0.0091
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
VACUNADO	1	-0.4767	0.0996	22.89	0.0000
ENFERMO	2	-0.3188	0.0996	10.24	0.0014
VACUNADO*ENFERMO	3	-0.2599	0.0996	6.81	0.0091

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion
0	0	375.1012	1.0000
1	0	375.1012	0

```
                          Parameter Estimates
      Iteration        1            2            3
      ---------------------------------------------
           0       -0.4767      -0.3188      -0.2599
           1       -0.4767      -0.3188      -0.2599

     MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE
Source                    DF   Chi-Square      Prob
-------------------------------------------------
VACUNADO                   1        22.89    0.0000
ENFERMO                    1        10.24    0.0014
VACUNADO*ENFERMO           1         6.81    0.0091
LIKELIHOOD RATIO           0            .        .

        ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                        Standard    Chi-
Effect            Parameter  Estimate     Error    Square   Prob
-----------------------------------------------------------------
VACUNADO                  1   -0.4767    0.0996     22.89  0.0000
ENFERMO                   2   -0.3188    0.0996     10.24  0.0014
VACUNADO*ENFERMO          3   -0.2599    0.0996      6.81  0.0091
```

El archivo de resultados (**C20-8.LST**) informa que hay una poblaciones y cuatro respuestas que son

si vacunado, *si* enfermo
si vacunado, *no* enfermo
no vacunado, *si* enfermo
no vacunado, *no* enfermo.

También nos presenta la matriz de transformación o *respuesta* y la matriz de *diseño* deducidas anteriormente.

Como se tiene en total solo tres grados, estas respuestas hay que referirlas a una de ellas, tomándose la última como referencia, siendo la función de respuesta de, por ejemplo, la primera casilla

$$\ln \frac{\frac{11}{148}}{\frac{54}{148}} = \ln 11 - \ln 54 = -1.59109$$

Las significaciones de los diferentes efectos se miran en la tabla de ANALYSIS-OF-VARIANCE o en la tabla de MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE.

La significación del efecto principal VACUNADO (P<0.001) indica que el número de vacunados y no vacunados no está uniformemente distribuido, más adelante se verá en que sentido se produce esta falta de uniformidad.

La significación del efecto principal ENFERMO (P<0.01) indica que el número de enfermos y no enfermos no está uniformemente distribuido, más adelante se verá en que sentido se produce esta falta de uniformidad.

La significación de la interacción VACUNADO*ENFERMO (P<0.01) indica que la distribución de las cantidades de las diferentes casillas no están uniforme distribuidas, es decir, que existe asociación entre ambas variables.

Como se ha introducido en el modelo todos los efectos que pueden intervenir no queda residuo.

Para saber el sentido de estas significaciones (en una tabla de 2x2 es fácil de ver a simple vista, pero no en tablas mayores), se estudia la tabla ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATE o la tabla ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES.

La desviación de la primera fila (si vacunados) con respecto a la media es significativa ($P<0.001$) y es menor de cero (-0.4767), lo que indica que la proporción de vacunados es menor que la proporción de no vacunados.

La desviación de la primera columna (enfermos) con respecto a la media es significativa ($P<0.01$) y menor de cero (-0.3188), lo que indica que la proporción de enfermos es menor que la proporción de no enfermos.

La desviación de la primera casilla (vacunados/enfermos) con respecto a la media es significativa ($P<0.01$) y menor de cero (-0.2599), lo que indica que la proporción de esta casilla es menor que la proporción de las demás casillas.

Recuérdese que estos valores se estiman de la siguiente manera

$$\mu = \frac{\ln 11 + \ln 35 + \ln 48 + \ln 54}{4} = 3.4534$$

$$\mu_1^{\text{VACUNADO}} = \frac{\ln 11 + \ln 35}{2} - 3.4534 = -0.4767$$

$$\mu_1^{\text{ENFERMO}} = \frac{\ln 11 + \ln 48}{2} - 3.4534 = -0.3188$$

$$\mu_{11}^{\text{VACUAND*ENFERMO}} = \ln 11 + 0.4767 + 0.3188 - 3.4534 = -0.2599$$

Por lo que las conclusiones son

1) La proporción de vacunados es estadísticamente inferior que la proporción de no vacunados.
2) La proporción de enfermos es estadísticamente inferior que la proporción de no enfermos.
3) La proporción de vacunados/enfermos es estadísticamente inferior que la proporción de no vacunados/no enfermos, es decir, existe asociación negativa entre la vacuna y la enfermedad. Dicho de otro modo, la vacuna protege de la enfermedad.

Ejemplo.-

Se quiere saber si existe asociación entre el tipo y la localización de tumores cerebrales. Los tumores se clasificaron en tres tipos: tumores *benignos*, tumores *malignos* y *otros* tumores cerebrales. Las localizaciones fueron: en lóbulos *frontales*, en

lóbulos *temporales* y en *otras* áreas cerebrales.

La incidencia de los diferentes tumores en las distintas localizaciones para una muestra de 141 pacientes de neurocirugía fue

	Benigno	*Maligno*	*Otro*	*total*
Frontal	23	9	6	38
Temporal	21	4	3	28
Otro	34	24	17	75
total	78	37	26	141

Archivo del programa *SAS* (C20-9.SAS).-

```
title 'Modelos log lineales';
options ls=75 ps=60;
data mod_line;
infile 'c20-9.dat';
input lugar $ tipo $ n @@;
proc catmod order=data;
 weight n;
 model lugar * tipo =_response_ / wls ml;
 loglin lugar tipo lugar*tipo;
run;
```

Como se ve en el programa ***SAS***, se ha realizado el análisis del modelo loglineal por el método de *mínimos cuadrados ponderados* (**WLS**) y por el método de *máxima verosimilitud* (**ML**) si bien, al haber explicitado en el estamento **LOGLIN** el modelo completo (modelo saturado) ambos métodos dan el mismo resultado. Si no se especifica ninguna opción, el SAS hace por defecto el método de máxima verosimilitud,.

Como en este tipo de diseño ambas variables son dependiente y en un modelo hay que poner en el miembro de la derecha la variable independiente, se pone como variable independiente la respuesta, esto es **_RESPONSE_** , y se utiliza el estamento **LOGLIN** para especificar los efectos que se quieren estudiar; estos efectos comprenden el efecto de la **_RESPONSE_** en el modelo. Como se ve, en **LOGLIN** se ha especificado el modelo completo, esto es, que se analicen los efectos de los dos factores (LUGAR y TIPO) por separado y la asociación entre estos dos factores (LUGAR*TIPO).

Archivo de datos (C20-9.DAT).-

```
frontal  benigno 23  frontal  maligno  9  frontal  otro  6
temporal benigno 21  temporal maligno  4  temporal otro  3
otro     benigno 34  otro     maligno 24  otro     otro 17
```

Archivo de resultados (C20-9.LST).-

Modelos log lineales

CATMOD PROCEDURE

Response: LUGAR*TIPO
Weight Variable: N
Data Set: MOD_LINE
Frequency Missing: 0

Response Levels (R)= 9
Populations (S)= 1
Total Frequency (N)= 141
Observations (Obs)= 9

Sample	Sample Size
1	141

RESPONSE PROFILES

Response	LUGAR	TIPO
1	frontal	benigno
2	frontal	maligno
3	frontal	otro
4	temporal	benigno
5	temporal	maligno
6	temporal	otro
7	otro	benigno
8	otro	maligno
9	otro	otro

Sample	Function Number	Response Function
1	1	0.30228
	2	-0.63599
	3	-1.04145
	4	0.21131
	5	-1.44692
	6	-1.73460
	7	0.69315
	8	0.34484

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
LUGAR	2	28.08	0.0000
TIPO	2	33.33	0.0000
LUGAR*TIPO	4	7.50	0.1116
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
LUGAR	1	-0.0909	0.1589	0.33	0.5673
	2	-0.6226	0.1916	10.55	0.0012
TIPO	3	0.7697	0.1335	33.24	0.0000
	4	-0.2119	0.1682	1.59	0.2077
LUGAR*TIPO	5	-0.00906	0.1859	0.00	0.9611
	6	0.0343	0.2292	0.02	0.8811
	7	0.4316	0.2156	4.01	0.0453
	8	-0.2450	0.2818	0.76	0.3847

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2
0	0	556.04607	1.0000	-0.0909	-0.6226
1	2	556.04607	0	-0.0909	-0.6226

```
                                  Parameter Estimates
Iteration       3          4          5          6          7          8
-------------------------------------------------------------------------
     0       0.7697    -0.2119  -0.009065     0.0343     0.4316    -0.2450
     1       0.7697    -0.2119  -0.009065     0.0343     0.4316    -0.2450

          MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE
        Source                       DF    Chi-Square       Prob
        ---------------------------------------------------------
        LUGAR                         2         28.08     0.0000
        TIPO                          2         33.33     0.0000
        LUGAR*TIPO                    4          7.50     0.1116
        LIKELIHOOD RATIO              0             .          .

               ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                                 Standard     Chi-
    Effect              Parameter  Estimate       Error      Square    Prob
    ------------------------------------------------------------------------
    LUGAR                       1   -0.0909      0.1589        0.33  0.5673
                                2   -0.6226      0.1916       10.55  0.0012
    TIPO                        3    0.7697      0.1335       33.24  0.0000
                                4   -0.2119      0.1682        1.59  0.2077
    LUGAR*TIPO                  5  -0.00906      0.1859        0.00  0.9611
                                6    0.0343      0.2292        0.02  0.8811
                                7    0.4316      0.2156        4.01  0.0453
                                8   -0.2450      0.2818        0.76  0.3847
```

El archivo de resultados (**C20-9.LST**) informa que hay una población y nueve respuestas que son

lugar *frontal*,	tipo *benigno*
lugar *frontal*,	tipo *maligno*
lugar *frontal*,	tipo *otro*
lugar *temporal*,	tipo *benigno*
lugar *temporal*,	tipo *maligno*
lugar *temporal*,	tipo *otro*
lugar *otro*,	tipo *benigno*
lugar *otro*,	tipo *maligno*
lugar *otro*,	tipo *otro*

Como se tiene en total solo ocho grados, estas respuestas hay que referirlas a una de ellas, tomándose la última como referencia, siendo la función de respuesta de, por ejemplo, la segunda casilla

$$\ln \frac{\frac{9}{141}}{\frac{17}{141}} = \ln 9 - \ln 17 = -0.63599$$

Las significaciones de los diferentes efectos se miran en la tabla de ANALYSIS-OF-VARIANCE o en la tabla de MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE.

La significación del efecto principal LUGAR (P<0.001) indica que las proporciones de tumores observados en los diferentes lugares no están uniformemente distribuido, más adelante se verá en que sentido se produce esta falta de uniformidad.

La significación del efecto principal TIPO (P<0.001) indica que las proporciones de los diferentes tipos de tumores no están uniformemente distribuido, más adelante se

verá en que sentido se produce esta falta de uniformidad.

La no significación de la interacción LUGAR*TIPO (P>0.05) indica que la distribución de las cantidades de las diferentes casillas están uniforme distribuidas, es decir, que no existe asociación entre ambas variables.

Como se ha introducido en el modelo todos los efectos que pueden intervenir no queda residuo.

Al salir no significativa la interacción o asociación, se puede eliminar esta del estamento LOGLIN y analizar solo, si interesa, los marginales, pero como se vera dentro de un momento al analizar los parámetros estimados, aunque no sea significativa la interacción general, puede ocurrir que el resultado de alguna casilla sea interesante

Para saber el sentido de estas significaciones, se estudia la tabla ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATE o la tabla ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES.

La desviación de la primera fila (localización en lóbulos frontales) con respecto a la media es no significativa (P>0.05), lo que indica que la proporción de tumores en lóbulos frontales es igual que la proporción media de tumores localizada en las demás zonas.

La desviación de la segunda fila (localización en lóbulos temporales) con respecto a la media es significativa (P<0.01) y es menor de cero (-0.6226), lo que indica que la proporción de tumores en lóbulos temporales es menor que la proporción media de tumores localizada en las demás zonas.

La desviación de la primera columna (tumores benignos) con respecto a la media es significativa (P<0.001) y mayor de cero (0.7697), lo que indica que la proporción de tumores benignos es mayor que la proporción de tumores de los demás tipos.

La desviación de la segunda columna (tumores malignos) con respecto a la media es no significativa (P>0.05), lo que indica que la proporción de tumores malignos es igual que la proporción media de tumores de los demás tipos.

Las desviaciones de las diferentes casilla con respecto a la media son no significativa (P>0.05), lo que indica que la proporción de tumores en esta casilla es igual que la proporción media de tumores en las diferentes combinaciones de LUGAR*TIPO.

Sin embargo la desviación de la casilla 2,1 (localización lóbulo temporales, tumores benignos) es significativa (P<0.05) y mayor de cero (0.4316), lo que indica que la proporción de tumores benignos en lóbulos temporales es mayor que la proporción media de tumores en las demás combinaciones.

Recuérdese que estos valores se estiman de la siguiente manera

$$\mu = \frac{\ln 23+\ln 9+\ln 6+\ln 21+\ln 4+\ln 3+\ln 34+\ln 24+\ln 17}{9} =$$

$$= 2.4657$$

$$\mu_1^{LUGAR} = \frac{\ln 23+\ln 9+\ln 6}{3} - 2.4657 = -0.0909$$

$$\mu_2^{LUGAR} = \frac{\ln 21+\ln 4+\ln 3}{3} - 2.4657 = -0.6226$$

$$\mu_1^{TIPO} = \frac{\ln 23+\ln 21+\ln 34}{3} - 2.4657 = 0.7697$$

$$\mu_2^{TIPO} = \frac{\ln 9+\ln 4+\ln 24}{3} - 2.4657 = -0.2119$$

.....

$$\mu_{21}^{LUGAR*TIPO} = \ln 21+0.6226 - 0.7697 - 2.4657 = 0.4350$$

.....

Por lo que las conclusiones son

1) La proporción de tumores de lóbulos frontales no es estadísticamente diferente que la proporción de tumores en otras zonas.
2) La proporción de tumores de lóbulos temporales es estadísticamente inferior que la proporción de tumores en otras zonas.
3) La proporción de tumores benignos es estadísticamente superior que la proporción de tumores de otro tipo.
4) La proporción de tumores malignos no es estadísticamente diferente que la proporción de tumores de otro tipo.
5) No existe asociación entre el tipo de tumor y su localización, es decir, la probabilidad de localizar tumores benignos o maligno u otros es la misma en las tres zonas del cerebro y el que se localice un tumor en los lóbulos frontales o temporales u otras zonas no indica nada sobre la mayor o menor probabilidad de benignidad o malignidad u otro tipo.
6) Sin embargo, hay una mayor localización de tumores benignos en los lóbulos temporales que en las demás zonas.

Prueba exacta de Fisher para experimentos aleatorios o Modelo III.-

Como se ha expuesto anteriormente la *prueba exacta de Fisher* se desarrolló para analizar aquellas tablas en las que los dos marginales son fijos, en la que, por tanto, se tiene una distribución *Hipergeométrica*. Pero ocurre que al diseñar una experiencia con los dos marginales fijo, suele ocurrir que el número de datos sea pequeño, lo que ha llevado a la confusión de utilizar esta prueba siempre que el tamaño de muestra sea pequeño. La realidad es que para tamaños de muestra pequeños esta prueba es más exacta de la prueba χ^2 o que la prueba *G*.

Hay ocasiones en las que sólo es posible obtener cantidades limitadas de datos si, por ejemplo, es preciso destruir las unidades experimentales y estas son costosas o es muy costoso y difícil la obtención de los datos. Cuando en una tabla 2x2 los números son pequeños (los totales de las filas y columnas son inferiores a 15), puede ser mejor calcular probabilidades exactas. Al estar basada en la distribución hipergeométrica, la prueba exacta de Fisher consiste en, dado los dos marginales fijos, calcular la probabilidad de obtener la tabla observada y sumarle las probabilidades de obtener las tablas que presenten una mayor desviación que la desviación de la tabla observada con respecto a lo esperado por los marginales fijos.

El número total de maneras diferentes en que puede obtenerse una tabla bifactorial con totales marginales fijos es el producto del número de combinaciones de *N* elementos de orden $(n_{11}+n_{12})$ por el número de combinaciones de *N* elementos de orden $(n_{11}+n_{21})$.

$$C_N^{(n_{11}+n_{12})} \times C_N^{(n_{11}+n_{21})} = \frac{N!}{(n_{11}+n_{12})!\ (n_{21}+n_{22})!} \times \frac{N!}{(n_{11}+n_{21})!\ (n_{12}+n_{22})!}$$

Mientras que se puede demostrar que existen

$$\frac{N!}{n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!}$$

maneras diferentes de obtener las casillas observadas. Por tanto, la probabilidad de obtener una tabla 2x2 con la frecuencia observada en las casillas se calcula como el cociente de las dos cantidades anteriores, cociente que puede simplificarse a

$$P = \frac{(n_{11}+n_{12})!\,(n_{21}+n_{22})!\,(n_{11}+n_{21})!\,(n_{12}+n_{22})!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!\,N!}$$

Dado que generalmente interesan las pruebas de dos colas, se deberían de considerar todos los casos peores en ambas colas, aunque por comodidad de cálculo (si se realiza a mano) simplemente se calcula la probabilidad para la cola que contiene el resultado observado y, posteriormente, se multiplica por dos esta probabilidad. O bien, en vez de doblar por dos la probabilidad, podríamos comparar la probabilidad exacta con un nivel de significación de $\alpha/2$, es decir, se puede comparar con α=0.025, α=0.005 ó α=0.0005, en lugar de α=0.05, α=0.01 ó α=0.001, respectivamente.

Ejemplo.-

Un investigador aplicó un nuevo tratamiento a un grupo de siete macacos y un tratamiento estándar a un grupo de seis macacos a los que previamente se les había inoculado un virus. De los siete macacos a los que se les aplicó el nuevo tratamiento murieron dos y de los seis macacos con el tratamiento estándar murieron tres. ¿Es más efectivo el nuevo que el viejo tratamiento?

	No murieron	*murieron*	*total*
Nuevo	5	2	7
Estándar	3	3	6
total	8	5	13

$$P = \frac{7!\ 6!\ 8!\ 5!}{5!\ 2!\ 3!\ 3!\ 13!} = 0.3263$$

Ahora se le resta 1 a las casillas n_{12} y n_{21} y se le suma 1 a las casillas n_{11} y n_{22}, dando una nueva tabla en la que las frecuencias de casillas se desvían más de las frecuencias esperadas (suponiendo independencia) que lo hacen los resultados del experimento.

6	1	7
2	4	6
8	5	13

como los totales marginales permanecen constantes, el numerador de la probabilidad no hay que volverlo a calcular.

$$P = \frac{7!\ 6!\ 8!\ 5!}{6!\ 1!\ 2!\ 4!\ 13!} = 0.0816$$

Se repite esta operación hasta que la frecuencia de una de las casillas sea cero, lo que ocurre en el siguiente paso

7	0	7
1	5	6
8	5	13

$$P = \frac{7!\ 6!\ 8!\ 5!}{7!\ 0!\ 1!\ 5!\ 13!} = 0.0047$$

Se suman todas las probabilidades calculadas

$$P_{total} = 2(P_1 + P_2 + P_3) = 2(0.3263 + 0.0816 + 0.00047) = 0.8252$$

como es mayor de 0.05, se concluyen que son independientes el número de muertos del tratamiento. No existe asociación.

Suponiendo independencia entre los tratamientos y el número de muertes, y siendo fijos los marginales, la probabilidad de obtener los resultados de este experimento más los resultados menos probables, es la suma de las probabilidades calculadas multiplicada por dos para una prueba de dos colas. La suma de estas probabilidades es 0.4126 que multiplicada por dos es 0.8252. Es claro que para responder a la pregunta de si es significativa la asociación es suficiente con el cálculo de la primera o última probabilidad solamente. En la práctica, primero se calcula la mayor probabilidad individual, y así sucesivamente si interesa la probabilidad exacta o si ésta aún es significativa.

Archivo del programa *SAS* (C20-10.SAS).-

```
title 'Prueba exacta de Fisher';
options ls=75 ps=60;
data chi2;
infile 'c20-10.dat';
input trata $ respues $ n @@;
proc freq order=data;
weight n;
table trata * respues / chisq;
run;
```

Archivo de datos (C20-10.DAT).-

```
nuevo vivos 5   nuevo muertos 2
viejo vivos 3   viejo muertos 3
```

Archivo de resultados (C20-10.LST).-

```
                    Prueba exacta de Fisher

                TABLE OF TRATA BY RESPUES
TRATA       RESPUES
Frequency|
Percent  |
Row Pct  |
Col Pct  |vivos    |muertos |  Total
---------+--------+--------+
nuevo    |      5 |      2 |      7
         |  38.46 |  15.38 |  53.85
         |  71.43 |  28.57 |
         |  62.50 |  40.00 |
---------+--------+--------+
viejo    |      3 |      3 |      6
         |  23.08 |  23.08 |  46.15
         |  50.00 |  50.00 |
         |  37.50 |  60.00 |
---------+--------+--------+
Total            8        5       13
                61.54    38.46   100.00
```

```
        STATISTICS FOR TABLE OF TRATA BY RESPUES
Statistic                          DF      Value         Prob
-------------------------------------------------------------
Chi-Square                          1      0.627        0.429
Likelihood Ratio Chi-Square         1      0.630        0.427
Continuity Adj. Chi-Square          1      0.048        0.826
Mantel-Haenszel Chi-Square          1      0.579        0.447
Fisher's Exact Test (Left)                                0.914

                        (Right)                           0.413
                        (2-Tail)                          0.592
Phi Coefficient                            0.220
Contingency Coefficient                    0.214
Cramer's V                                 0.220

Sample Size = 13
WARNING: 100% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.
```

Como se ve, el resultado para una cola (**1-Tail**) es el mismo que el obtenido con la calculadora de bolsillo, sin embargo el resultado para dos colas no coincide puesto que se ha realizado una aproximación, al multiplicar por dos, suponiendo que la distribución es simétrica, lo que no es cierto, por lo que es más fiable el resultado del ***SAS*** para dos colas.

También se puede comprobar que si se hubiera hecho la prueba χ^2 o la prueba *G* no hubiera cambiado la conclusión.

Por supuesto que este ejemplo se podía hacer también por medio de un modelo lineal.

Lo explicado en este epígrafe se refieren exclusivamente a tablas 2x2, si se tienen tablas mayores el método sería el mismo pero utilizando al distribución hipergeómetrica múltiple. En este caso la realización de los cálculos para esta prueba con calculadora de bolsillo sería muy tediosos, por lo que se recomienda el utilizar paquetes estadísticos. En el caso concreto del ***SAS***, como se ha visto en el último ejemplo, si la tabla es de 2x2 hace la prueba exacta de *Fisher* por defecto, esto es, sin necesidad de pedírselo en el programa, pero si la tabla es mayor de 2x2 no hace esta prueba por defecto, como se ve en los resultados de ejemplos anteriores, en este caso hay que explicitar en el programa que se quiere la prueba exacta poniendo la opción **EXACT** en el estamento **TABLE**.

Ejemplo.-

Se toma una muestra de individuos con infarto de miocardio y se les pregunta por el número de cigarrillos diarios que fumaban antes del infarto y se clasifican en tres grupos: 0 cigarrillos, de 1 a 24 cigarrillos y más de 25 cigarrillos. Para contrastar se toma una muestra control de individuos que no ha sufrido infarto de miocardio y se les preguntan sobre la misma cuestión. Obteniendo los siguientes resultados

	0	1-24	>25	*total*
Control	25	25	12	62
Infarto	0	1	3	4
total	25	26	15	66

Archivo del programa *SAS* (C20-11.SAS).-

```
title 'Prueba exacta de Fisher para tablas mayores de 2x2';
options ls=75 ps=60;
data chi2;
infile 'c20-11.dat';
input cantidad $ respues $ n @@;
proc freq order=data;
weight n;
table cantidad * respues / exact norow nocol nopercent;
run;
```

Archivo de datos (C20-11.DAT).-

```
  0  control 25       0 infarto 0
1-24 control 25    1-24 infarto 1
 >25 control 12     >25 infarto 3
```

Archivo de resultados (C20-11.LST).-

```
        Prueba exacta de Fisher para tablas mayores de 2x2

                  TABLE OF CANTIDAD BY RESPUES
                CANTIDAD     RESPUES
                Frequency|control |infarto |  Total
                ---------+--------+--------+
                0        |     25 |      0 |     25
                ---------+--------+--------+
                1-24     |     25 |      1 |     26
                ---------+--------+--------+
                >25      |     12 |      3 |     15
                ---------+--------+--------+
                Total          62        4       66

             STATISTICS FOR TABLE OF CANTIDAD BY RESPUES
        Statistic                      DF      Value         Prob
        ------------------------------------------------------------
        Chi-Square                      2      6.956        0.031
        Likelihood Ratio Chi-Square     2      6.690        0.035
        Mantel-Haenszel Chi-Square      1      5.845        0.016
        Fisher's Exact Test (2-Tail)                        0.034
        Phi Coefficient                        0.325
        Contingency Coefficient                0.309
        Cramer's V                             0.325

        Sample Size = 66
        WARNING: 50% of the cells have expected counts less
                 than 5. Chi-Square may not be a valid test.
```

Como, para la prueba exacta, $P<0.05$, se concluye que existe asociación entre

ambas variables.

Diferentes disposiciones de los datos en tablas de 2x2.-

Los datos presentados en una tabla 2x2 pueden ser a veces dispuestos de tal manera que pueden ser probadas dos hipótesis nulas. Con un ejemplo se entenderá mejor esta cuestión.

Ejemplo.-

Se toma una muestra al azar de 41 jamones de cerdo ibérico y otra muestra de 22 jamones de cerdo blanco. Se le pide a un catador que clasifique los jamones en una u otra clase basándose en su degustación. Del total de jamones ibéricos 35 los clasifica correctamente y de los jamones blancos, clasifica correctamente 10. Quedando los datos de esta manera

		Clasificación del catador		
		Ibérico	*Blanco*	*total*
Clasificació n correcta	*Ibérico*	35	6	41
	Blanco	12	10	22
	total	47	16	63

La primera hipótesis nula que se prueba es la de no asociación entre la clasificación del catador y la clasificación real

$$\chi^2 = \frac{(35\times10-6\times12)^2\times63}{41\times22\times47\times16} = 7.1780^{**}$$

$$\chi^2_{adj} = \frac{\left(|35\times10-6\times12|-\frac{63}{2}\right)^2 63}{41\times22\times47\times16} = 5.6435^{*}$$

$$\chi^2_{(1;\,0.05)} = 3.841$$

por lo que se desecha la hipótesis nula y se puede concluir que la clasificación realizada por el catador fue mejor que la que sería de esperar por el azar.

Pero ahora puede surgir la pregunta de si la proporción de jamones clasificados correctamente era la misma para cada tipo. Estas proporciones se obtienen de las casillas n_{11} y n_{22} y de los subtotales de las filas de la tabla de arriba. Estos son 35/41=0.854 para la muestra de ibéricos y 10/22=0.455 para la muestra de blancos. Para probar si estas dos proporciones difieren se hace una nueva distribución de los datos de la siguiente manera

		Clasificación		
		Correcta	*Incorrecta*	*total*
Clasificación	*Ibérico*	35	6	41
	Blanco	10	12	22
	total	45	18	63

$$\chi^2 = \frac{(35\times12-6\times10)^2 63}{41\times22\times45\times18} = 11.1752***$$

$$\chi^2_{adj} = \frac{\left(/35\times12-6\times10/-\frac{63}{2}\right)^2 63}{41\times22\times45\times18} = 9.3051**$$

$$\chi^2_{(1;\,0.01)} = 6.635$$

por lo que se rechaza la hipótesis nula de que la proporción de jamones clasificados correctamente no difieren. El catador ha demostrado ser más eficiente al clasificar correctamente los jamones ibéricos que al clasificar los blancos.

Tablas Multifactoriales y Análisis Múltiple de Correspondencias.-

Al igual que ocurre con el *ANOVA* en el que partiendo de dos factores se generaliza a más de dos factores, también ahora se puede avanzar de las tablas de dos factores hasta aquellas que contienen clasificaciones con más de dos criterios.

El análisis de tablas con más de dos criterios de clasificación no tiene complicación conceptual, si tiene complicación operativa pues el número de cálculos es mayor pero esto no es un obstáculo si se utilizan los paquetes estadísticos.

Tablas Multifactoriales y Análisis Múltiple de Correspondencias en un Modelo Aleatorio simples estratificado o Modelo II (modelo lineal).-

Aunque la base teórica es la misma de la vista anteriormente, explicaremos como se hace el calculo manual, solo en este caso, con objeto de una mejor compresión de los resultado y porque el ejemplo es pequeño. En las tablas multifactoriales del Modelo I el desarrollo numérico es el mismo.

En una prueba de independencia de tres factores, el valor esperado para cada casilla, e_{ijk} es igual al producto de los totales de la fila i, columna j y profundidad k dividido por el cuadrado del tamaño de la muestra total N. Por lo que se puede calcular el χ^2 por la fórmula ya conocida

$$\chi^2 = \Sigma_i \frac{(o - e)^2}{e}$$

o bien se puede calcular la razón de verosimilitud

$$G = 2\sum_{ijk} o_{ijk} \ln\left(\frac{o_{ijk}}{e_{ijk}}\right)$$

aunque se puede desarrollar una expresión más simple, semejante a la empleada en ejemplos anteriores, para cada uno de los múltiples análisis que se pueden hacer con una tabla multifactorial. Para ello es conveniente hallar previamente la expresión $\Sigma\, o \ln o$ para cada una de las tablas, es decir,

1 para la tabla trifactorial: *FxCxP*

$$F \times C \times P \Rightarrow \sum O_{(\text{en cada casilla})} \ln O_{(\text{en cada casilla})}$$

2 para las tres tablas bifactoriales: *FxC*, *FxP* y *CxP*,

$$F \times C \Rightarrow \sum O_{(\text{en cada casilla})} \ln O_{(\text{en cada casilla})}$$
$$F \times P \Rightarrow \sum O_{(\text{en cada casilla})} \ln O_{(\text{en cada casilla})}$$
$$C \times P \Rightarrow \sum O_{(\text{en cada casilla})} \ln O_{(\text{en cada casilla})}$$

3 para las tres tablas de un factor o totales marginales de los tres criterios de clasificación: *F*,*C* y *P*

$$F \Rightarrow \sum O_{(\text{en cada marginal})} \ln O_{(\text{en cada marginal})}$$
$$C \Rightarrow \sum O_{(\text{en cada marginal})} \ln O_{(\text{en cada marginal})}$$
$$P \Rightarrow \sum O_{(\text{en cada marginal})} \ln O_{(\text{en cada marginal})}$$

4 para el gran total: *T*

$$T \Rightarrow N \ln N$$

De modo que si se quiere probar la asociación entre los tres factores la prueba sería

$$G = 2[T \times C \times P - F - C - P + 2T]$$

Los grados de libertad pueden obtenerse, al igual que lo estudiado para las tablas de contingencia de doble entrada, mediante la expresión

$$gl = F \times C \times P - (F-1) - (C-1) - (P-1) - 1 = F \times C \times P - F - C - P + 2$$

dado que se calculan *F*-1 parámetros para las filas, *C*-1 parámetros para las columnas y *P*-1 parámetros para las *P* profundidades.

Tanto en el caso de que la asociación sea significativa como en el caso de que sea no significativa se puede profundizar en el estudio para ver las causas de la falta

de independencia o si existe alguna asociación bifactorial sin que exista la trifactorial pues la falta de independencia de los tres factores no implica necesariamente falta de independencia de dos de ellos. Las tres pruebas posibles serían

$$G = 2[F \times C - F - C + T]; \quad gl = (F-1)(C-1)$$
$$G = 2[F \times P - F - P + T]; \quad gl = (F-1)(P-1)$$
$$G = 2\,[C \times P - C - P + T]; \quad gl = (C-1)(P-1)$$

En el caso de que sea significativa la asociación de alguna de estas tablas y éstas sean de dimensiones mayores a 2x2 se puede descomponer en tablas de 2x2 para hacer un estudio más pormenorizado de los diferentes sumandos del valor de *G*.

Por último, se puede calcular la componente de la interacción que sería

$$G = 2[F \times C \times P - F \times C - F \times P - C \times P + F + C + P - T]$$
$$gl = (F-1)(C-1)(P-1)$$

La interacción trifactorial tiene el mismo significado que en el *ANOVA* de tres factores, es decir, que el grado de asociación entre un par de factores difiere de un nivel o clase de un tercer factor a otro nivel.

Ejemplo.-

Se pretende probar la efectividad de una vacuna, para ello se toman 50 animales a los que se les vacuna y, pasado el tiempo necesario para la incubación y desarrollo de los anticuerpos, se les inocula la enfermedad. A otros 50 animales se les inocula la enfermedad sin haberlos vacunado previamente. Como quiera que la efectividad de la vacuna puede estar influida por la raza, se realiza la experiencia con 50 animales de la raza *A* y 50 animales de la raza *B*, dando los siguientes resultados: de los 25 animales de la raza *A* vacunados, 13 enfermaron y de la raza *B* enfermaron 6. De los no vacunados, enfermaron 12 de la raza *A* y 18 de la raza *B*.

Raza	*Vacuna*	*Morbilidad*		
		Sanos	*Enfermos*	*total*
A	*Si*	12	13	25
	No	13	12	25
	suma	25	25	50
B	*Si*	19	6	25
	No	17	18	25
	suma	26	24	50
	total	51	49	100

¿Es efectiva la vacuna? ¿Influye realmente la raza? ¿Interacciona la efectividad de la vacuna con la raza?

Las tablas bifactoriales *RxV* y *RxM* salen directamente de los marginales de filas

y columnas, respectivamente, de la tabla anterior. La tabla bifactorial *VxM* hay que elaborarla a partir de la tabla general, para mayor comodidad de los cálculos

		Morbilidad		
		Sanos	*Enfermos*	*total*
Vacuna	*Si*	31	19	50
	No	20	30	50
	total	51	49	100

$$R\times V\times M \Rightarrow 12\ln 12+13\ln 13+13\ln 13+12\ln 12+19\ln 19+6\ln 6+7\ln 7+18\ln 18 =$$
$$= 258.6694$$
$$R\times V \Rightarrow 25\ln 25+25\ln 25+25\ln 25+25\ln 25 =$$
$$= 321.8876$$
$$R\times M \Rightarrow 25\ln 25+25\ln 25+26\ln 26+24\ln 24 =$$
$$= 321.9276$$
$$V\times M \Rightarrow 31\ln 31+19\ln 19+20\ln 20+30\ln 30 =$$
$$= 324.3485$$

$$R \Rightarrow 50\ln 50+50\ln 50 = 391.2023$$
$$V \Rightarrow 50\ln 50+50\ln 50 = 391.2023$$
$$M \Rightarrow 51\ln 51+49\ln 49 = 391.2223$$

$$T \Rightarrow 100\ \ln\ 100 = 460.5170$$

La prueba de asociación entre los tres factores es

$$G = 2(258.6694 - 391.2023 - 391.2023 - 391.2223+2\times 460.5170) =$$
$$= 12.153\ *$$

$\chi^2_{(4;\,0.05)} = 9.488$

existe asociación entre los tres criterios de clasificación.

Si se profundiza en el estudio de las causas de la falta de independencia de los tres factores, se pueden ver las pruebas de asociación doble de interés para el problema, estas son la asociación entre la raza y la morbilidad y entre la vacuna y la morbilidad, no tiene sentido hallar la asociación entre la raza y la vacuna, pues estas frecuencias han sido determinadas por el experimentador

Raza × *Morbilidad*

$$G = 2(321.9276 - 391.2023 - 391.2223+460.5170) = 0.04 ns$$

$\chi^2_{(1;\,0.05)} = 3.841$

Vacuna × *Morbilidad*

$$G = 2(324.3487 - 391.2023 - 391.2223 + 460.5170) = 4.8822 *$$

$$\chi_{(1;\,0.05)} = 3.841$$

Se concluye que existe asociación entre la morbilidad y la vacunación como consecuencia de que del total de enfermos un 61% no están vacunados mientras que el 39% restante si están vacunados. Pero no existe asociación entre la morbilidad y la raza. Por supuesto, tampoco existe asociación entre la raza y la vacunación.

La significación de la componente *interacción* es

$$G = 2\ (258.6694 - 321.8876 - 321.9276 - 324.3487 +$$
$$+ 391.2023 + 391.2023 + 391.2223 - 460.5170) =$$
$$= 7.2308 * *$$

$$\chi^2_{(1;\,0.05)} = 6.635$$

Por lo tanto existe interacción, esto es, la morbilidad no es lo mismo en una raza que en la otra, o lo que es lo mismo, la vacuna no es igual de efectiva en una raza que en la otra. Para saber en que sentido se produce dicha interacción hay que analizar las tablas de asociación de *VxM* dentro de las *razas*.

Raza A		*Morbilidad*		
		Sanos	*Enfermos*	*total*
Vacuna	*Si*	12	13	25
	No	13	12	25
	total	25	25	50

$$G = 2\,[(12 \ln 12 + 13 \ln 13 + 13 \ln 13 + 12 \ln 12) -$$
$$-(25 \ln 25 + 25 \ln 25 + 25 \ln 25 + 25 \ln 25) +$$
$$+ 50 \ln 50] =$$
$$= 0.08 ns$$

En la raza *A* no es efectiva la vacuna.

Raza B		*Morbilidad*		
		Sanos	*Enfermos*	*total*
Vacuna	*Si*	19	6	25
	No	7	18	25
	total	26	24	50

$$G = 2\,[(19\ \ln 19 + 6\ \ln 6 + 7\ \ln 7 + 18\ \ln 18) -$$
$$- (25\ \ln 25 + 25\ \ln 25 + 26\ \ln 26 + 24\ \ln 24) +$$
$$+ 50\ \ln 50] =$$
$$= 12.033\ ***$$

Mientras que en la raza *B* la vacuna es muy efectiva.

El cuadro resumen sería

Prueba	*gl*	*G*	*significación*
RxVxM	4	12.153	*
RxM	1	0.040	*ns*
VxM	1	4.883	*
Interacción	1	7.231	**
VxM (R=A)	1	0.080	*ns*
VxM (R=B)	1	12.033	*

Obsérvese que el valor de la *G* total es la suma de las *G* de todas las tablas 2x2 más la interacción. El grado de libertad que falta es el de la tabla *raza* $\times$ *vacuna* que cuya *G* vale cero pues las frecuencias han sido determinadas por el experimentador y éste ha hecho un diseño equilibrado.

Este tipo de problemas se resuelve, si se usan paquetes estadísticos, por medio de modelos lineales

Archivo del programa *SAS* (C20-12.SAS).-

```
title 'Modelos lineales con tablas multifactoriales';
Options ls=75 ps=30;
data modecad;
infile 'c20-12.dat';
input raza $ vacuna $ morbili $ n @@;
proc catmod order=data;
weight n;
model morbili = raza vacuna raza*vacuna ;
response marginals;
run;
title 'Análisis jerárquico para saber el sentido de la interacción';
proc catmod order=data;
weight n;
model morbili = raza vacuna(raza='A') vacuna(raza='B');
response marginals;
run;
title 'Análisis de correspondencias';
proc corresp mca out=corr;
tables raza vacuna morbili;
weight n;
run;
proc plot;
plot dim1 * dim2 = _name_;
run;
```

La opción **MCA** del procedimiento **CORRESP** es para requerir un análisis de correspondencias múltiple.

Archivo de datos (C20-12.DAT).-

```
A si Sano  12  A si Enfermo  13
A no Sano  13  A no Enfermo  12
B si Sano  19  B si Enfermo   6
B no Sano   7  B no Enfermo  18
```

Archivo de resultados (C20-12.LST).-

```
          Modelos lineales con tablas multifactoriales

                        CATMOD PROCEDURE
Response: MORBILI                    Response Levels (R)=      2
Weight Variable: N                   Populations     (S)=      4
Data Set: MODECAD                    Total Frequency (N)=    100
Frequency Missing: 0                 Observations  (Obs)=      8

                      POPULATION PROFILES
                                          Sample
                Sample  RAZA  VACUNA       Size
                ------------------------------
                    1    A       si           25
                    2    A       no           25
                    3    B       si           25
                    4    B       no           25

                      RESPONSE PROFILES
                      Response  MORBILI
                      ------------------
                           1    Sano
                           2    Enfermo

                  ANALYSIS-OF-VARIANCE TABLE
        Source                      DF   Chi-Square       Prob
        ---------------------------------------------------------
        INTERCEPT                    1       117.80     0.0000
        RAZA                         1         0.05     0.8315
        VACUNA                       1         5.48     0.0192
        RAZA*VACUNA                  1         7.65     0.0057
        RESIDUAL                     0            .         .

           ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
                                             Standard     Chi-
Effect              Parameter  Estimate       Error     Square    Prob
-------------------------------------------------------------------------
INTERCEPT                   1    0.5100      0.0470     117.80  0.0000
RAZA                        2   -0.0100      0.0470       0.05  0.8315
VACUNA                      3    0.1100      0.0470       5.48  0.0192
RAZA*VACUNA                 4   -0.1300      0.0470       7.65  0.0057


  Análisis jerárquico para saber el sentido de la interacción

                        CATMOD PROCEDURE

Response: MORBILI                    Response Levels (R)=      2
Weight Variable: N                   Populations     (S)=      4
Data Set: MODECAD                    Total Frequency (N)=    100
Frequency Missing: 0                 Observations  (Obs)=      8
```

POPULATION PROFILES

Sample	RAZA	VACUNA	Sample Size
1	A	si	25
2	A	no	25
3	B	si	25
4	B	no	25

RESPONSE PROFILES

Response	MORBILI
1	Sano
2	Enfermo

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	117.80	0.0000
RAZA	1	0.05	0.8315
VACUNA(RAZA=A)	1	0.08	0.7771
VACUNA(RAZA=B)	1	15.00	0.0001
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.5100	0.0470	117.80	0.0000
RAZA	2	-0.0100	0.0470	0.05	0.8315
VACUNA(RAZA=A)	3	-0.0200	0.0707	0.08	0.7771
VACUNA(RAZA=B)	4	0.2400	0.0620	15.00	0.0001

Análisis de correspondencias
The Correspondence Analysis Procedure
Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	8 16 24 32 40
0.63795	0.40698	126.069	40.70%	************************
0.57735	0.33333	103.255	33.33%	********************
0.50959	0.25968	80.440	25.97%	***************
	1.00000	309.764	(Degrees of Freedom = 25)	

Column Coordinates

	Dim1	Dim2
A	0.070738	0.995893
B	-.070738	-.995893
no	0.778121	-.090536
si	-.778121	0.090536
Enfermo	0.797115	0.000000
Sano	-.765856	0.000000

Summary Statistics for the Column Points

	Quality	Mass	Inertia
A	0.996807	0.166667	0.166667
B	0.996807	0.166667	0.166667
no	0.613669	0.166667	0.166667
si	0.613669	0.166667	0.166667
Enfermo	0.610476	0.163333	0.170000
Sano	0.610476	0.170000	0.163333

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
A	0.002049	0.495902
B	0.002049	0.495902
no	0.247951	0.004098
si	0.247951	0.004098
Enfermo	0.255000	0.000000

```
                Sano          0.245000       0.000000

          Indices of the Coordinates that Contribute
             Most to Inertia for the Column Points
                           Dim1           Dim2           Best
      A                       0              2              2
      B                       0              2              2
      no                      1              0              1
      si                      1              0              1
      Enfermo                 1              0              1
      Sano                    1              0              1

            Squared Cosines for the Column Points
                                 Dim1           Dim2
             A             0.005004       0.991803
             B             0.005004       0.991803
             no            0.605472       0.008197
             si            0.605472       0.008197
             Enfermo       0.610476       0.000000
             Sano          0.610476       0.000000
      Plot of DIM1*DIM2.  Symbol is value of _NAME_.
DIM1 |
 1.0 +
     |
     |                          n  E
     |
 0.5 +
     |
     |                                                    A
 0.0 +
     |  B
     |
-0.5 +
     |
     |                          S  s
     |
-1.0 +
     |
     ---+-------------+-------------+-------------+-------------+--
       -1.0          -0.5           0.0           0.5           1.0
                                   DIM2
NOTE: 1 obs had missing values.
```

En el archivo de resultados (**C20-12.LST**), en la salida del primer procedimiento **CADMOD,** se observa:

- la **RAZA** es no significativa, esto es, no existe asociación entre la morbilidad y las razas como consecuencia de que el número de individuos que enferman son estadísticamente los mismos en las dos razas.
- La **VACUNA** es significativa como consecuencia de que es diferente el número de individuos que enferman entre los vacunados y los no vacunados.
- La interacción **RAZA*VACUNA** es significativa, esto es, la morbilidad no es la misma en una raza que en la otra, o lo que es lo mismo, la vacuna no es igual de efectiva en una raza que en la otra. Para saber en que sentido se produce esta interacción se hace el análisis jerárquico del segundo procedimiento CADMOD del programa.

En la salida del segundo procedimiento **CADMOD**, se observa:

- **VACUNA(RAZA=A)** es no significativa, por lo que la vacuna no es efectiva en la raza A

- **VACUNA(RAZA=B)** es altamente significativa, por lo que la vacuna es muy efectiva en la raza B

En la salida del procedimiento **CORRESP** se observa:
- El valor del χ^2 total es altamente significativo.
- Los dos primeros ejes explican un 74% del χ^2 total

La representación de estos dos ejes de correspondencias dadas por el procedimiento **PLOT** muestra:
- Los individuos **Sanos** están junto con los **si vacunados** y los individuos **Enfermos** están junto a los **no vacunados**, por lo tanto es clara la causa de esta asociación.
- La raza **A** y la raza **B** están junto al cero en el primer eje y equidistantes de los vacunados y no vacunados, esto indica que no existe asociación entre la morbilidad y las razas.
- pero las dos razas están muy separadas en el segundo eje, esto indica que el comportamiento de la vacuna y la morbilidad es diferente en una raza que en la otra, en otras palabras, existe interacción.

Tablas Multifactoriales y Análisis Múltiple de Correspondencias en un Modelo Aleatorio simples o Modelo I (Modelo log-lineal).-

La base teórica es la misma de la vista anteriormente, por lo que se pasa directamente al ejemplo.

Ejemplo.-

Se está haciendo un estudio de prospección parar saber la preferencia de coche de las personas, para ello se pregunta a 2559 personas sobre el tipo de coche que posee (*familiar, deportivo* o de *trabajo*), sobre si la casa en que vive es *propia* o es *alquilada*, sobre su status (*soltero sin hijos, casado sin hijos, soltero con hijos* y *casado con hijos*) y el sexo (*hombre* o *mujer*).

Archivo del programa *SAS* (C20-13.SAS).-

```
title 'Modelo log-lineal en tablas multifactoriales';
Options ls=75 ps=30;
Data modlin;
Infile 'c20-13.dat';
Input coche $ casa $ status $ sexo $ n @@;
Proc catmod order=data;
 Weight n;
 Model coche*casa*status*sexo = _response_ ;
 Loglin coche casa status sexo
   coche*casa coche*status coche*sexo
   casa*status casa*sexo status*sexo
   coche*casa*status coche*casa*sexo
   coche*status*sexo casa*status*sexo
   coche*casa*status*sexo;
run;
proc corresp mca out=corr;
tables coche casa status sexo;
weight n;
run;
proc plot;
```

```
plot dim2 * dim1 = _name_;
run;
```

Como es un Modelo I o muestreo aleatorio simple, se analiza por medio de los modelos loglineales.

Como se ve en el programa ***SAS***, se ha realizado el análisis del modelo loglineal por el método de *máxima verosimilitud* (**ML**), que es el que hace por defecto cuando no se especifica nada en el modelo, si bien, al haber explicitado en el estamento **LOGLIN** el modelo completo (modelo saturado) ambos métodos dan el mismo resultado.

Como en este tipo de diseño ambas variables son dependiente y en un modelo hay que poner en el miembro de la derecha la variable independiente, se pone como variable independiente la respuesta, esto es **_RESPONSE_** , y se utiliza el estamento **LOGLIN** para especificar los efectos que se quieren estudiar; estos efectos comprenden el efecto de la **_RESPONSE_** en el modelo. Como se ve, en **LOGLIN** se ha especificado el modelo completo, esto es, que se analicen los efectos de los cuatro factores (COCHE, CASA, STATUS y SEXO) por separado y toda las posibles interacciones entre estos cuatro factores.

Archivo de datos (C20-13.DAT).-

```
familiar propia   soltero hombre  12  deportiv alquiler soltero hombre  27
familiar propia   soltero mujer   83  deportiv alquiler soltero mujer   16
familiar propia   casado  hombre  29  deportiv alquiler casado  hombre  13
familiar propia   casado  mujer  163  deportiv alquiler casado  mujer   17
familiar propia   Sol_hij hombre  27  deportiv alquiler Sol_hij hombre  13
familiar propia   Sol_hij mujer   85  deportiv alquiler Sol_hij mujer   11
familiar propia   Cas_hij hombre  39  deportiv alquiler Cas_hij hombre  17
familiar propia   Cas_hij mujer  152  deportiv alquiler Cas_hij mujer   18
familiar alquiler soltero hombre  66  trabajo  propia   soltero hombre  44
familiar alquiler soltero mujer   18  trabajo  propia   soltero mujer   63
familiar alquiler casado  hombre  41  trabajo  propia   casado  hombre  52
familiar alquiler casado  mujer   24  trabajo  propia   casado  mujer   69
familiar alquiler Sol_hij hombre  49  trabajo  propia   Sol_hij hombre  53
familiar alquiler Sol_hij mujer   26  trabajo  propia   Sol_hij mujer   85
familiar alquiler Cas_hij hombre  64  trabajo  propia   Cas_hij hombre  57
familiar alquiler Cas_hij mujer   37  trabajo  propia   Cas_hij mujer   83
deportiv propia   soltero hombre  32  trabajo  alquiler soltero hombre 158
deportiv propia   soltero mujer   44  trabajo  alquiler soltero mujer   45
deportiv propia   casado  hombre  39  trabajo  alquiler casado  hombre  83
deportiv propia   casado  mujer   56  trabajo  alquiler casado  mujer   56
deportiv propia   Sol_hij hombre  37  trabajo  alquiler Sol_hij hombre 168
deportiv propia   Sol_hij mujer   38  trabajo  alquiler Sol_hij mujer   52
deportiv propia   Cas_hij hombre  23  trabajo  alquiler Cas_hij hombre  87
deportiv propia   Cas_hij mujer   22  trabajo  alquiler Cas_hij mujer   53
```

Archivo de resultados (C20-13.LST).-

```
          Modelo log-lineal en tablas multifactoriales

                          CATMOD PROCEDURE
Response: COCHE*CASA*STATUS*SEXO        Response Levels (R)=    48
Weight Variable: N                      Populations     (S)=     1
Data Set: MODLIN                        Total Frequency (N)=  2546
Frequency Missing: 0                    Observations  (Obs)=    48

                                 Sample
                       Sample     Size
                       ----------------
                          1       2546
```

RESPONSE PROFILES

Response	COCHE	CASA	STATUS	SEXO
1	familiar	propia	soltero	hombre
2	familiar	propia	soltero	mujer
3	familiar	propia	casado	hombre
4	familiar	propia	casado	mujer
5	familiar	propia	Sol_hij	hombre
6	familiar	propia	Sol_hij	mujer
7	familiar	propia	Cas_hij	hombre
8	familiar	propia	Cas_hij	mujer
9	familiar	alquiler	soltero	hombre
10	familiar	alquiler	soltero	mujer
11	familiar	alquiler	casado	hombre
12	familiar	alquiler	casado	mujer
13	familiar	alquiler	Sol_hij	hombre
14	familiar	alquiler	Sol_hij	mujer
15	familiar	alquiler	Cas_hij	hombre
16	familiar	alquiler	Cas_hij	mujer
17	deportiv	propia	soltero	hombre
18	deportiv	propia	soltero	mujer
19	deportiv	propia	casado	hombre
20	deportiv	propia	casado	mujer
21	deportiv	propia	Sol_hij	hombre
22	deportiv	propia	Sol_hij	mujer
23	deportiv	propia	Cas_hij	hombre
24	deportiv	propia	Cas_hij	mujer
25	deportiv	alquiler	soltero	hombre
26	deportiv	alquiler	soltero	mujer
27	deportiv	alquiler	casado	hombre
28	deportiv	alquiler	casado	mujer
29	deportiv	alquiler	Sol_hij	hombre
30	deportiv	alquiler	Sol_hij	mujer
31	deportiv	alquiler	Cas_hij	hombre
32	deportiv	alquiler	Cas_hij	mujer
33	trabajo	propia	soltero	hombre
34	trabajo	propia	soltero	mujer
35	trabajo	propia	casado	hombre
36	trabajo	propia	casado	mujer
37	trabajo	propia	Sol_hij	hombre
38	trabajo	propia	Sol_hij	mujer
39	trabajo	propia	Cas_hij	hombre
40	trabajo	propia	Cas_hij	mujer
41	trabajo	alquiler	soltero	hombre
42	trabajo	alquiler	soltero	mujer
43	trabajo	alquiler	casado	hombre
44	trabajo	alquiler	casado	mujer
45	trabajo	alquiler	Sol_hij	hombre
46	trabajo	alquiler	Sol_hij	mujer
47	trabajo	alquiler	Cas_hij	hombre
48	trabajo	alquiler	Cas_hij	mujer

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
COCHE	2	309.35	0.0000
CASA	1	36.75	0.0000
STATUS	3	2.35	0.5029
SEXO	1	1.80	0.1797
COCHE*CASA	2	77.11	0.0000
COCHE*STATUS	6	38.84	0.0000
COCHE*SEXO	2	35.47	0.0000
CASA*STATUS	3	17.85	0.0005
CASA*SEXO	1	154.60	0.0000
STATUS*SEXO	3	8.48	0.0370
COCHE*CASA*STATUS	6	16.06	0.0134
COCHE*CASA*SEXO	2	58.12	0.0000
COCHE*STATUS*SEXO	6	1.55	0.9560
CASA*STATUS*SEXO	3	14.62	0.0022
COCHE*CASA*STATUS*SEXO	6	8.45	0.2067
LIKELIHOOD RATIO	0	.	.

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
COCHE	1	0.0622	0.0350	3.15	0.0757
	2	-0.5700	0.0402	200.65	0.0000
CASA	3	0.1526	0.0252	36.75	0.0000
STATUS	4	-0.0556	0.0445	1.56	0.2114
	5	0.0253	0.0430	0.35	0.5556
	6	-0.0148	0.0444	0.11	0.7390
SEXO	7	-0.0338	0.0252	1.80	0.1797
COCHE*CASA	8	0.0286	0.0350	0.67	0.4133
	9	0.2387	0.0402	35.20	0.0000
COCHE*STATUS	10	-0.2411	0.0656	13.53	0.0002
	11	0.0202	0.0596	0.11	0.7345
	12	-0.0557	0.0607	0.84	0.3584
	13	0.2247	0.0675	11.07	0.0009
	14	0.0860	0.0686	1.57	0.2096
	15	-0.0928	0.0725	1.64	0.2006
COCHE*SEXO	16	-0.1632	0.0350	21.71	0.0000
	17	0.0152	0.0402	0.14	0.7062
CASA*STATUS	18	-0.1473	0.0445	10.98	0.0009
	19	0.1476	0.0430	11.79	0.0006
	20	0.0318	0.0444	0.51	0.4740
CASA*SEXO	21	-0.3129	0.0252	154.60	0.0000
STATUS*SEXO	22	0.0727	0.0445	2.67	0.1020
	23	-0.1088	0.0430	6.40	0.0114
	24	0.0611	0.0444	1.89	0.1687
COCHE*CASA*STATUS	25	-0.0780	0.0656	1.41	0.2342
	26	0.0636	0.0596	1.14	0.2859
	27	-0.0658	0.0607	1.18	0.2780
	28	0.0514	0.0675	0.58	0.4470
	29	0.0338	0.0686	0.24	0.6217
	30	0.1483	0.0725	4.18	0.0408
COCHE*CASA*SEXO	31	-0.2611	0.0350	55.60	0.0000
	32	0.2487	0.0402	38.19	0.0000
COCHE*STATUS*SEXO	33	-0.0344	0.0656	0.28	0.5996
	34	0.00797	0.0596	0.02	0.8936
	35	0.00760	0.0607	0.02	0.9004
	36	-0.00290	0.0675	0.00	0.9658
	37	-0.0302	0.0686	0.19	0.6601
	38	-0.00738	0.0725	0.01	0.9190
CASA*STATUS*SEXO	39	-0.1613	0.0445	13.16	0.0003
	40	0.0602	0.0430	1.96	0.1610

	41	0.0113	0.0444	0.06	0.7992
COCHE*CASA*STATUS*SEXO	42	-0.0731	0.0656	1.24	0.2652
	43	-0.0517	0.0596	0.75	0.3853
	44	0.1176	0.0607	3.75	0.0527
	45	0.0150	0.0675	0.05	0.8238
	46	-0.0194	0.0686	0.08	0.7770
	47	0.00449	0.0725	0.00	0.9506

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	5 10 15 20 25
0.63487	0.40306	4443.68	23.03%	**********
0.52475	0.27536	3035.85	15.74%	*******
0.50298	0.25299	2789.21	14.46%	******
0.48687	0.23705	2613.42	13.55%	******
0.47179	0.22259	2453.99	12.72%	******
0.45142	0.20378	2246.62	11.64%	*****
0.39393	0.15518	1710.85	8.87%	****
	1.75000	19293.6	(Degrees of Freedom = 100)	

Column Coordinates

	Dim1	Dim2
deportiv	0.35148	1.66370
familiar	0.65825	-0.59833
trabajo	-0.62167	-0.12937
alquiler	-0.80921	-0.17655
propia	0.67619	0.14753
Cas_hij	0.29404	-1.11218
Sol_hij	-0.39207	0.01988
casado	0.54538	0.53345
soltero	-0.47592	0.60832
hombre	-0.74253	0.04561
mujer	0.69401	-0.04263

Summary Statistics for the Column Points

	Quality	Mass	Inertia
deportiv	0.576110	0.041536	0.119122
familiar	0.443915	0.089847	0.091516
trabajo	0.364029	0.118617	0.075076
alquiler	0.573227	0.113806	0.077825
propia	0.573227	0.136194	0.065032
Cas_hij	0.455571	0.064022	0.106273
Sol_hij	0.052181	0.063236	0.106722
casado	0.196247	0.063040	0.106834
soltero	0.187153	0.059701	0.108742
hombre	0.517267	0.120778	0.073841
mujer	0.517267	0.129222	0.069016

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
deportiv	0.012731	0.417512
familiar	0.096586	0.116809
trabajo	0.113735	0.007209
alquiler	0.184894	0.012882
propia	0.154501	0.010764
Cas_hij	0.013733	0.287588
Sol_hij	0.024117	0.000091
casado	0.046522	0.065149
soltero	0.033549	0.080231

```
                 hombre        0.165215       0.000913
                 mujer         0.154418       0.000853



               Indices of the Coordinates that Contribute
                 Most to Inertia for the Column Points
                            Dim1          Dim2          Best
            deportiv         0             2             2
            familiar         2             2             2
            trabajo          1             0             1
            alquiler         1             0             1
            propia           1             0             1
            Cas_hij          0             2             2
            Sol_hij          0             0             1
            casado           0             0             2
            soltero          0             0             2
            hombre           1             0             1
            mujer            1             0             1

             Plot of DIM2*DIM1.  Symbol is value of _NAME_.
DIM2 |
   2 +
     |                                                 d
     |
     |
   1 +
     |
     |                 s                                      c
   0 +     h              S                                      p
         a      t                                                 m
     |
     |                                                           f
  -1 +                                                 C
     |
     |
  -2 +
     |
   ---+--------+--------+--------+--------+--------+--------+--------+--
    -1.00    -0.75    -0.50    -0.25     0.00     0.25     0.50     0.75
                                  DIM1
NOTE: 1 obs had missing values.
```

El archivo de resultados (**C20-13.LST**), la salida del procedimiento **CADMOD** informa que hay una población y 48 respuestas que son el producto de los niveles de los cuatro factores, por lo que se tienen 47 grados de libertad

Las significaciones de los diferentes efectos se miran en la tabla de MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE.

De los cuatro factores principales son significativos el coche y la casa y no son significativos ni el status ni el sexo. Mientras que son significativas todas las interacciones menos una triple (coche*status*sexo) y la cuadruple.

Como se ha introducido en el modelo todos los efectos que pueden intervenir no queda residuo.

Para saber el sentido de estas significaciones, se estudia la tabla **ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES**. Si se desea obtener otras respuestas, no hay más que cambiar el orden de los datos.

Esta tabla indica:

La desviación del primer efecto (coche familiar) con respecto a la media es no significativa al $P>0.05$ (es si significativa al $P>0.1$ y positiva) lo que indica que la proporción de coches familiares es igual (o ligeramente superior) que la proporción media de los tres tipos de coches.

La desviación del segundo efecto (coches deportivos) con respecto a la media es significativa ($P<0.001$) y es menor de cero (-0.5700), lo que indica que la proporción de coches deportivos es significativamente menor que la proporción media de los tres tipos de coches.

La desviación del tercer efecto (casa propia) con respecto a la media es significativa ($P<0.001$) y positiva, lo que indica que la proporción de casas propias es significativamente superior que la proporción media de los dos tipos de casas.

Las desviaciones de los efectos cuarto al sexto, que se corresponden con los status, son no significativas, esto es, hay el mismo número de solteros, casados, solteros con hijos y casados con hijos.

La desviación del efecto séptimo, que se corresponde con el sexo, también es no significativa, esto es, hay el mismo número de hombres que de mujeres.

Las desviaciones de los efectos 8 y 9, que se corresponde con la interacción tipo de coche x tipo de casa, es no significativo el número de los que tienen coche familiar y casa propia, mientras que si es significativo y positivo el número de los que tienen casa propia y coche deportivo.

Las desviaciones de los efectos 10 a 15, que se corresponden con la interacción tipo de coche x status, es significativamente pequeño el número de individuos solteros con coche familiar y significativamente grande el número de individuos solteros con coche deportivo, las demás combinaciones son no significativas.

Las desviaciones de los efectos 16 a 17, que se corresponden con la interacción tipo de coche x sexo, es significativo y negativo la proporción de hombres con coche familiar y no significativo la proporción de hombres con coche deportivo.

Las desviaciones de los efectos 18 a 20, que se corresponden con la interacción tipo de casa x status, es significativamente negativa la proporción de solteros con casa propia y significativamente positiva la proporción de casados con casa propia.

La desviación de la respuesta 21, que se corresponde con la interacción tipo de casa sexo, es significativamente negativa la proporción de hombres con casa propia.

Las desviaciones de las respuestas 22 a 24, que se corresponden con la interacción status x sexo, solo es significativamente negativo la proporción de hombres

casados.

Del resto de las interacciones (las triples y cuádruples) solo son significativas la de los hombres con coche familiar y cada propia cuya proporción es muy pequeña, la de hombres con coche deportivo y casa propia cuya proporción es muy grande y la de hombres con casa propia y solteros cuya proporción es muy pequeña.

Con esto es fácil extraer las conclusiones de estos resultados

Con respecto a la salida del procedimiento del análisis factorial de correspondencias (**CORRESP**), se observa que hay varios ejes con inercia significativa. Analizando los dos primeros ejes, en la salida del procedimiento **PLOT**, se observa que existe una asociación clara entre *mujer* con *casa propia* y *coche familiar*, por un lado, y *hombre* con *casa alquilada* y *coche de trabajo*, por otro lado. *Casado con hijos* y *coche deportivo* está próximo al primer grupo con respecto al primer eje pero alejados y opuestos con respecto al segundo eje; y *soltero* y *soltero con hijos* están próximo al segundo grupo.

Con respecto al segundo eje, como ya se ha señalado, se observa que todas las variables están próximas al cero excepto *coche deportivo* y *casado con hijos* que, claramente, están el lado positivo y negativo, respectivamente,

Análisis factorial de Correspondencias para el estudio de cualquier matriz de números positivos.-

Como se dijo anteriormente, si bien el análisis de correspondencias se concibió para el estudio de tablas de contingencia, se ha revelado eficaz para el estudio de cualquier matriz de números no negativos, esto permite resumir la información de una matriz de datos describiendo sintéticamente pautas de relaciones entre las filas o los individuos y las columnas o las variables, relación que es imposible de entresacar directamente de una tabla de datos. Para ello se hace una reducción del espacio factorial, tal como se hace con el Análisis de Componentes Principales, condensándose el máximo de información en uno o pocos factores. En la representación gráfica de los resultados, los individuos similares aparecen juntos con las variables que más influyen en ellos por lo que es fácil contestar a preguntas como ¿qué categorías de una variable son más similares y podría unirse en unas sola? ¿Cuáles son las categorías más diferentes? ¿Cuál es la asociación entre las filas y las columnas de la tabla de datos?

Ejemplo.-

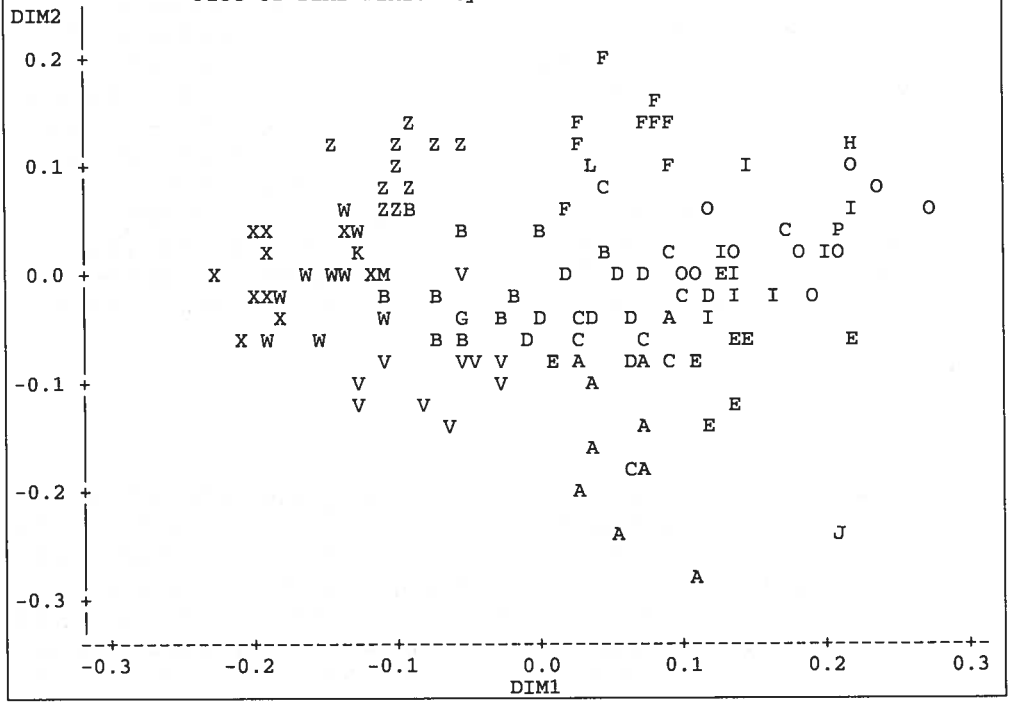
Recuérdese los datos analizados con el análisis de componentes principales y el análisis factorial, se trataba del gasto anual medio que realizan 112 familias en siete productos o categorías de productos alimenticios. Las familias están clasificadas según el nivel profesional del padre y según el número de hijos, habiendo doce tipos en total: **T2** trabajador manual con dos hijos, **O2** empleado de oficina con dos hijos, y **D2** directivo con dos hijos; y los mismos niveles profesionales del padre para 3, 4 y 5 hijos. Las categorías de productos alimenticios son: pan, legumbres, fruta, carne, pollo, leche y vino.

Archivo de programa *SAS* (C20-14.SAS).-

```
title 'Análisis de Correspondencias para una matriz números positivos';
Options ls=75 ps=60;
Data vino;
Infile 'c18-1.dat';
input familia $ simfam $ pan legumbre Kfruta Gcarne Mpollos Hleche Jvino
@@;
proc corresp out=corr;
var pan legumbre Kfruta Gcarne Mpollos Hleche Jvino;
id simfam;
run;
proc plot;
plot dim2 * dim1= simfam;
run;
```

Archivo de resultados (C20-14.LST).-

```
    Análisis de Correspondencias para una matriz números positivos

                The Correspondence Analysis Procedure
                Inertia and Chi-Square Decomposition
     Singular  Principal Chi-
     Values    Inertias  Squares Percents   10   20   30   40   50
                                           ----+----+----+----+----+---
     0.12300   0.01513   9230.83  50.06% **************************
     0.08401   0.00706   4306.01  23.35% ************
     0.05788   0.00335   2043.77  11.08% ******
     0.04523   0.00205   1248.03   6.77% ***
     0.04091   0.00167   1021.05   5.54% ***
     0.03114   0.00097    591.55   3.21% **
               -------   -------
               0.03022   18441.2 (Degrees of Freedom = 714)

             Plot of DIM2*DIM1.  Symbol is value of SIMFAM.
```



```
NOTE: 1 obs had missing values.  8 obs hidden.
```

Analizando la salida del procedimiento PLOT, de las dos primeras dimensiones del procedimiento CORRESP, se observa que en el cuadrante superior izquierdo se asocian la FRUTA y el POLLO con las directivos de muchos hijos. En el cuadrante superior derecho se asocian la LECHE el PAN y las LEGUMBRES con los trabajadores de muchos hijos y los oficinistas de muchos hijos. En el cuadrante inferior izquierdo se asocian la CARNE con los directivos de pocos hijos y los oficinistas de pocos hijos. Y en cuadrante inferior derecha se asocian el VINO con los trabajadores de pocos hijos.

Prueba *U* de Mann-Whitney o prueba *S* de suma de ordenaciones de Wilcoxon.-

Cuando se tiene, por lo menos, una variable ordinal, la prueba de *Wilcoxon-Mann-Whitney* puede usarse para probar si dos muestras independientes proceden de la misma población. Es una de las pruebas no paramétricas más potentes y constituye la alternativa más útil ante la prueba *t* estudiada en el Capítulo 4 cuando no son válidas las suposiciones de la *t* o la medida realizada es menos fuerte que la de intervalos.

La estructura de los datos es idéntica a las presentadas en el Capítulo 4; es decir, se tienen dos muestras aleatorias independientes. La hipótesis nula será que las dos muestras pertenecen a la misma población, es decir, tienen la misma distribución, mientras que la hipótesis alternativa es que una muestra es mayor que la otra. Puesto que la variable es ordinal y pueden calcularse medianas de cada muestra, las hipótesis van a involucrar las medianas de la poblaciones que se desean comparar, siendo, por tanto, las posibles hipótesis

Cola derecha	Cola izquierda	Dos colas
$H_0 : M \leq M_0$	$H_0 : M \geq M_0$	$H_0 : M = M_0$
$H_1 : M > M_0$	$H_1 : M < M_0$	$H_1 : M \neq M_0$

Recuérdese, que la mediana es un parámetro de localización y que la única relación que puede establecerse entre observaciones registradas en valores ordinales es el ordenamiento por magnitud. Si se ordenan juntas las observaciones de ambas muestras, de acuerdo con su valor, pueden obtenerse muy diversos ordenamientos. Sin embargo, si se supone que $M_1 = M_2$, se puede esperar que las observaciones de las dos muestras se encuentren más o menos uniformemente mezcladas. En cambio, si al colocar las n_1+n_2 observaciones en orden ascendente, se encuentran que los elementos de una muestra ocupan las posiciones superiores, es lógico pensar que las poblaciones de las que se extrajeron las dos muestras tienen distribuciones diferentes, es decir, una distribución tiene las frecuencias concentradas en valores mayores (mayor mediana) que la otra distribución.

Toda la lógica de esta prueba está basada en la ordenación de los datos; para tal fin, se definirá el rango de una observación como el lugar que le corresponde en el ordenamiento de todas las observaciones, es decir, es el rango jerárquico numérico. Toda la información que contiene una muestra se sintetiza en la suma de los rangos de sus datos. Es decir, para una muestra, $X_1,\ldots, X_n$, se obtendrá $R(X_1),\ldots, R(X_n)$, que son

los rangos respectivos y además se tendrá R que es la suma de los rangos de dicha muestra.

La prueba *U/S* consta de los siguientes pasos

- Combinar todos los valores muestrales y asignarle rango en orden ascendente. Si dos o más valores coinciden se le asigna a cada uno el valor medio de los rangos que les hubiera correspondido (ver ejemplo).
- Hallar la suma de los rangos para cada muestra. Se denominarán R_1 y R_2, siendo n_1 y n_2 los respectivos tamaños de muestra. Por conveniencia se elige como n_1 el de la muestra de menor tamaño, si fueran desiguales.
- Una diferencia significativa entre la suma de los dos rangos implicaría una diferencia significativa entre las dos muestras. Para ello se usará el estadístico *U/S*

$$U = S = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Hay que hacer notar que este valor de U es el número total de veces que los valores de la muestra primera son mayores que los de la muestra segunda (ver ejemplo).

- La distribución muestral de U es simétrica con media y varianza

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12}$$

- Si n_1 y n_2 son mayores de ocho, la distribución de *U/S* se aproxima a la normal, por lo que la prueba de hipótesis puede ser

$$Z_o = \frac{U - \mu_U}{\sigma_U}$$

Corrección por continuidad.-

La prueba de *Wilcoxon-Mann-Whitney* supone que los datos están tomados con tal precisión que tienen una continuidad básica. Si eso fuera así, la probabilidad de encontrar dos valores iguales sería cero, por lo que no habría que compartir un rango. Sin embargo, esta prueba está recomendada precisamente para datos en los que no se va a encontrar esta continuidad, por lo que se supone que los datos que tienen el mismo valor son en realidad diferentes, pero en un grado que no detecta la medida que se ha realizado.

En el caso de que haya muchos datos repetidos puede alterar ligeramente la variabilidad del conjunto de los rangos y, por tanto, la prueba cuando se realiza la aproximación a la normal, por lo que se recomienda una corrección por continuidad,

consistente en calcular la varianza de la siguiente manera

$$\sigma_U = \left(\frac{n_1 n_2}{N(N-1)} \right) \left(\frac{N^3 - N}{12} - \Sigma T \right)$$

siendo

$$T = \frac{t^3 - t}{12}$$

donde t es el número de observaciones repetidas para un rango dado.

El valor de Z obtenido con esta corrección es un poco mayor que el anterior, en el caso de que haya datos repetidos, por lo que si no se hace la corrección la prueba será conservadora.

Ejemplo.-

Se está estudiando la concentración de jugo gástrico en pacientes de úlcera. Se toman dos grupos, un grupo formado por 29 personas con úlcera peptídica y otro grupo de 30 personas sin úlcera que se toman como control. La finalidad de este estudio es determinar si el nivel medio de lisozima es diferente para ambos grupos

Con Ulcera						*Sin Ulcera*					
0.2	16.2	24.0	50.0	10.4	2.1	0.2	1.5	2.5	4.8	5.4	8.8
4.8	7.5	0.3	17.6	25.4	60.0	16.5	20.7	0.3	1.5	2.8	4.8
10.9	3.3	4.9	9.8	0.4	18.9	5.7	9.1	16.7	33.0	0.4	1.9
40.0	11.3	3.8	5.0	1.1	20.7	3.6	5.8	10.3	20.0	0.7	2.0
42.2	12.4	4.5	5.3	2.0		7.5	15.6	1.2	2.4	8.7	16.1

Si se mezclan ambos grupos de datos y se ordenan por su rango se tiene la siguiente tabla

Ulcera	Normal	rango
0.2		1.5
	0.2	1.5
0.3		3.5
	0.3	3.5
0.4		5.5
	0.4	5.5
	0.7	7.0
1.1		8.0
	1.2	9.0
	1.5	10.5
	1.5	10.5
	1.9	12.0
2.0		13.5
	2.0	13.5
2.1		15.0
	2.4	16.0
	2.5	17.0
	2.8	18.0
3.3		19.0
	3.6	20.0

Ulcera	Normal	rango
3.8		21.0
4.5		22.0
4.8		24.0
	4.8	24.0
	4.8	24.0
4.9		26.0
5.0		27.0
5.3		28.0
	5.4	29.0
	5.7	30.0
	5.8	31.0
7.5		32.5
	7.5	32.5
	8.7	34.0
	8.8	35.0
	9.1	36.0
9.8		37.0
	10.3	38.0
10.4		39.0
10.9		40.0

Ulcera	Normal	rango
11.3		41.0
12.4		42.0
	15.6	43.0
	16.1	44.0
16.2		45.0
	16.5	46.0
	16.7	47.0
17.6		48.0
18.9		49.0
	20.0	50.0
20.7		51.5
	20.7	51.5
24.0		53.0
25.4		54.0
	33.7	55.0
40.0		56.0
42.2		57.0
50.0		58.0
60.0		59.0

Se tiene que el grupo de ulcerosos

$$n_1 = 29 \qquad R_1 = 976$$

y en el grupo de normales

$$n_2 = 30 \qquad R_2 = 794$$

El estadístico U o S

$$S = U = 29 \times 30 + \frac{29x30}{2} - 976 = 329$$

como N_1 y N_2 son mayores de 20 se puede hacer la aproximación a la normal

$$\mu_U = \frac{29 \times 30}{2} = 435$$

$$\sigma_U^2 = \frac{29 \times 30 \times (29 + 30 + 1)}{12} = 4350$$

$$Z_o = \frac{329 - 435}{65.9545} = 1.6072$$

Si se hace la corrección por continuidad se tendría que

el 0.2 se repite 2 veces
el 0.3 se repite 2 veces
el 0.4 se repite 2 veces
el 1.5 se repite 2 veces
el 2.0 se repite 2 veces
el 4.8 se repite 3 veces
el 7.5 se repite 2 veces
el 20.7 se repite 2 veces

por lo que se tiene siete valores que se repiten dos veces y un valor que se repite tres veces

$$\Sigma\ T = 7\times\left(\frac{2^3-2}{12}\right)+\frac{2^3-2}{12} = 5.58333$$

$$\sigma_U^2 = \left(\frac{29\times30}{59\times58}\right)\left(\frac{59^3-59}{12}-5.58333\right) = 4348.5805$$

y por tanto

$$Z_o = \frac{329\ -\ 435}{65.9438} = 1.6074$$

Archivo del programa *SAS* (C20-15.SAS).-

```
title 'Prueba de Mann-Whitney';
option ls=75 ps=60;
data ulcera;
infile 'c20-15.dat';
input grupo $ lisozima @@;
proc npar1way wilcoxon;
class grupo;
var lisozima;
run;
```

Archivo de datos (C20-15.DAT).-

```
U  0.2 U 16.2 U 24.0 U 50.0 U 10.4 U  2.1
U  4.8 U  7.5 U  0.3 U 17.6 U 25.4 U 60.0
U 10.9 U  3.3 U  4.9 U  9.8 U  0.4 U 18.9
U 40.0 U 11.3 U  3.8 U  5.0 U  1.1 U 20.7
U 42.2 U 12.4 U  4.5 U  5.3 U  2.0
N  0.2 N  1.5 N  2.5 N  4.8 N  5.4 N  8.8
N 16.5 N 20.7 N  0.3 N  1.5 N  2.8 N  4.8
N  5.7 N  9.1 N 16.7 N 33.0 N  0.4 N  1.9
N  3.6 N  5.8 N 10.3 N 20.0 N  0.7 N  2.0
N  7.5 N 15.6 N  1.2 N  2.4 N  8.7 N 16.1
```

Archivo de resultados (C20-15.LST).-

```
                    Prueba de Mann-Whitney

                N P A R 1 W A Y  P R O C E D U R E
          Wilcoxon Scores (Rank Sums) for Variable LISOZIMA
                  Classified by Variable GRUPO

                      Sum of      Expected       Std Dev         Mean
GRUPO        N        Scores      Under H0      Under H0        Score
U           29         976.0         870.0    65.9439284   33.6551724
N           30         794.0         900.0    65.9439284   26.4666667
                Average Scores Were Used for Ties

        Wilcoxon 2-Sample Test (Normal Approximation)
        (with Continuity Correction of .5)

        S =  976.000    Z =  1.59984    Prob > |Z| = 0.1096

        T-Test Approx. Significance = 0.1151

        Kruskal-Wallis Test (Chi-Square Approximation)
        CHISQ =  2.5838    DF =  1    Prob > CHISQ = 0.1080
```

Dos muestras apareadas o bloques aleatorios.-

Ya se ha visto que tanto en pruebas paramétricas como no paramétricas, se usan dos muestras cuando se desea comparar dos tratamientos. Si estas muestras son independientes podría ocurrir que la prueba detecte diferencias significativas que no son debidas a los tratamientos sino a la diferencia de los individuos empleados en uno y otro tratamiento. Ya se estudió en los Capítulos 4 y 5, y en este capítulo, que una manera de vencer esta dificultad es el usar muestras apareadas.

Se ha visto anteriormente que muchas pruebas de independencia responden a muestreos aleatorios simples estratificados o Modelo II, que en el caso de tablas 2x2, pueden considerarse como pruebas del efecto de tratamientos sobre el porcentaje de algún atributo. Cuando se estudió esta prueba en páginas anteriores se vio que cada tratamiento se aplicaba sendos grupos de individuos seleccionados independiente y aleatoriamente, individuos que eran diferentes para los diferentes tratamientos. Así se vio el ejemplo en el que se tomaron dos muestras de ratones que fueron sometidos a tratamientos diferentes: un grupo con vacunación y el otro grupo sin vacunación. Estos ratones seguramente fueron seleccionados aleatoriamente del conjunto de animales que estaban a disposición del experimentador y se puede considerar, por lo tanto, como un diseño completamente aleatorio o de muestras independientes.

Algunas veces no es posible ni conveniente recoger los datos en esta forma pues podría ocurrir que por las características del tratamiento el comportamiento de los individuos pudiera ser muy heterogéneo en función de la edad, color, condición fisiológica, *etc.* En tal caso, podría suceder que hubiese más individuos viejos en un grupo que en el otro, o cualquier otro sesgo, de manera que los efectos de tratamiento pueden ser confundidos o solapado con la heterogeneidad de los sujetos experimentales.

Existen tres maneras posibles para solucionar este problema. Si se conocen los factores causantes de la heterogeneidad se puede intentar que permanezcan constantes, lo cual obliga a buscar una población muy grande de donde poder extraer una muestra de animales uniformes con respecto a los factores causantes de la heterogeneidad, lo cual puede ser bastante difícil y además caro. Si los factores causantes de la heterogeneidad de respuesta son desconocidos se pueden intentar otros dos métodos. Uno de ellos consiste en tomar muestras de tamaño suficiente para superar la heterogeneidad; este método no es ni factible ni conveniente.

El método alternativo, es el de utilizar individuos emparejados como individuos del mismo sexo y camada o el mismo individuo para ambos tratamientos (autoemparejamiento). El experimento resultante es semejante al contraste de medias de datos emparejados del Capítulo 4 o de un modelo bifactorial con una medida por casilla del Capítulo 5.

Una segunda situación analítica en la que se recomienda este diseño es cuando existe alguna razón para creer que un tratamiento afectará la respuesta del otro tratamiento. En tales casos, ha de utilizarse sin duda alguna la misma muestra, debido a que los efectos del segundo tratamiento han de compararse con los efectos del primer tratamiento sobre los mismos individuos. Un ejemplo para ilustrar esta situación puede ser el de aprendizaje en animales: se le aplica un estímulo a un grupo de animales al que responderán o no; si más tarde se aplica este mismo estímulo u otro diferente, la presencia o ausencia de una respuesta en estos animales puede estar condicionada por sus respuestas anteriores, y de esta manera puede determinarse si ha habido aprendizaje.

Como en este caso las muestras no son independientes, los procedimientos analíticos vistos hasta ahora en este capítulo no son aplicables. Sin embargo, el problema es de fácil resolución.

Prueba de McNemar para significación de cambios.-

Si las *n* observaciones emparejadas son nominales u ordinales y se puede, por lo tanto, registrar en una tabla de 2x2 de la misma forma que las presentadas anteriormente pero con la diferencia de que es un diseño del tipo de *antes* y *después* en el que cada individuo es su propio control o hay dos tratamientos con individuos emparejados por su características. Se tiene, entonces

		Segundo tratamiento		
		Éxito	*Fracaso*	*total*
Primer tratamiento	*Éxito*	n_{11}	n_{12}	$n_{1.}$
	Fracaso	n_{21}	n_{22}	$n_{2.}$
	total	$n_{.1}$	$n_{.2}$	N

Pero a diferencia de la situación de los diseños completamente aleatorios, aquí la hipótesis nula es

$$H_o: P_{(\text{éxito primer tratamiento})} = P_{(\text{éxito segundo tratamiento})}$$

$$H_o: P_{n_{1.}} = P_{n_{.1}}$$

ó bien

$$H_o: P_{n_{11}} + P_{n_{12}} = P_{n_{11}} + P_{n_{21}}$$

Esto se reduce a

$$H_o: P_{n_{12}} = P_{n_{21}}$$
$$H_1: P_{n_{12}} \neq P_{n_{21}}$$

En consecuencia, sólo hay que tratar las entradas n_{12} y n_{21}, preguntándose si son iguales dentro del muestreo aleatorio. Esto se prueba con la siguiente fórmula simplificada para χ^2 con gl=1

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Si la cantidad $n_{12}+n_{21}$ es menor de 200, se deberá utilizar la corrección de continuidad, siendo la ecuación para la prueba de χ^2 ajustada la siguiente

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}$$

Ejemplo.-

En un estudio de analgésicos se compara un supuesto analgésico con un placebo. El experimento se realiza con 50 individuos, dando un alivio del 70% con el analgésico y del 40% con el placebo. Pero de los 35 individuos que le alivió el dolor el analgésico hubo 17 que también alivió el dolor el placebo, por lo que estos individuos no sirven para la prueba pues no discriminan entre ambos productos (no tenían autentico dolor), sin embargo hubo 18 individuos que no le alivió el placebo (tenían autentico dolor), por lo que estos individuos si sirven para la prueba; y de los 15 que no le alivió el analgésico hubo 3 que sí le alivió el placebo

		Placebo		
		Éxito	*Fracaso*	*total*
Analgésico	*Éxito*	17	18	35
	Fracaso	3	12	15
	total	20	30	50

$$\chi^2 = \frac{(18-3)^2}{18+3} = 10.7143^{***}$$

$$\chi^2_{adj} = \frac{(|18-3|-1)^2}{18+3} = 9.333^{**}$$

$$\chi^2_{(1;\,0.05)} = 3.841$$

Lo que indica que la respuesta ha sido diferente para el analgésico que para el placebo.

Archivo del programa *SAS* (C20-16.SAS).-

```
title 'Prueba de McNemar';
option ls=75 ps=60;
data bloques;
infile 'c20-16.dat';
input paciente analge $ placebo $ @@;
proc freq;
table analge*placebo ;
run;
data jidos;
  set bloques;
  producto='bueno'; exito=analge ; output;
  producto='malo' ; exito=placebo; output;
run;
proc freq;
 table paciente*producto*exito / noprint cmh1;
run;
```

Archivo de datos (C20-16.DAT).-

```
 1 ex ex   11 ex ex   21 ex ex   31 ex ex   41 ex ex
 2 ex ex   12 ex ex   22 ex ex   32 ex ex   42 ex ex
 3 ex ex   13 ex ex   23 ex ex   33 ex ex   43 ex ex
 4 ex ex   14 ex ex   24 ex fr   34 ex fr   44 ex fr
 5 ex fr   15 ex fr   25 ex fr   35 ex fr   45 ex fr
 6 ex fr   16 ex fr   26 ex fr   36 ex fr   46 ex fr
 7 fr ex   17 fr ex   27 fr ex   37 fr fr   47 fr fr
 8 ex fr   18 ex fr   28 ex fr   38 ex fr   48 ex fr
 9 fr fr   19 fr fr   29 fr fr   39 fr fr   49 fr fr
10 fr fr   20 fr fr   30 fr fr   40 fr fr   50 fr fr
```

Archivo de resultados (C20-16.LST).-

Prueba de McNemar

TABLE OF ANALGE BY PLACEBO

ANALGE Frequency Percent Row Pct Col Pct	PLACEBO ex	 fr	Total
ex	17 34.00 48.57 85.00	18 36.00 51.43 60.00	35 70.00
fr	3 6.00 20.00 15.00	12 24.00 80.00 40.00	15 30.00
Total	20 40.00	30 60.00	50 100.00

SUMMARY STATISTICS FOR PRODUCTO BY EXITO
CONTROLLING FOR PACIENTE
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	10.714	0.001

Prueba de los rangos con signos de Wilcoxon.-

Si la variable utilizada tiene, como mínimo, un grado de información de ordinal se puede utilizar para comparar dos muestra apareadas la prueba de los rangos con signos de *Wilcoxon*. Con esta prueba se puede saber cuál de los dos miembros de un par es mayor, es decir, indicar el signo de la diferencia en cualquier par, y clasificar las diferencias por orden de tamaño absoluto.

Los supuestos de esta prueba son, por tanto, que las diferencias entre los valores del par son variables aleatorias y que estas diferencias pueden ordenarse.

Esta prueba es la equivalente a la prueba *t* de diferencias del Capítulo VIII. El procedimiento a seguir en este caso es el siguiente

- Se estima la diferencia de cada par
- Se ordenan estas diferencias sin tener en cuenta el signo, es decir, al valor absoluto más pequeño se le asigna el valor ordinal 1, *etc.* Si hay varias diferencias con el mismo valor se les asignan el valor promedio de los rangos que le corresponderían.
- Al valor de orden o rango de cada diferencia se le añade ahora el signo (- ó +).
- Se suman los valores absolutos de los rangos del signo (- ó +) que sumen menos. Este es el estadístico *T*.
- Las diferencias igual a cero no se tienen en cuenta, por tanto el tamaño de muestra (*N*) será igual al de diferencias diferentes de cero.

- Si N es mayor de 25 esta distribución se aproxima a una normal de

$$\mu_T = \frac{N(N+1)}{4}$$

$$\sigma_T^2 = \frac{N(N+1)\ (2N+1)}{24}$$

- Por tanto se puede hacer la prueba de hipótesis

$$Z_o = \frac{T + \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)\ (2N+1)}{24}}}$$

La prueba de hipótesis puede ser tanto de una como de dos colas.

Ejemplo.-

Para comparar el efecto preventivo de dos fungicidas en plantas de algodón se tomaron 20 plantas, cada una de las cuales se dividen en dos partes iguales y a cada mitad se le aplica uno u otro fungicida al azar. Se expone cada planta a la acción de las esporas de un hongo y, al cabo de cierto tiempo se cuenta el número de pústulas en cada una de las partes de cada planta, obteniéndose

Planta	*Fungi A*	*Fungi B*	*A-B*	*|A-B|*	*rango*	*con sig*
1	8	16	- 8	8	12.5	-12.5
2	12	13	- 1	1	1	- 1
3	14	7	7	7	10.5	10.5
4	16	16	0	0	-	-
5	9	4	5	5	7	7
6	5	8	- 3	3	4	- 4
7	2	12	-10	10	14.5	-14.5
8	7	3	4	4	6	6
9	6	9	- 3	3	4	- 4
10	6	12	- 6	6	8.5	- 8.5
11	2	9	- 7	7	10.5	-10.5
12	4	1	3	3	4	4
13	5	15	-10	10	14.5	-14.5
14	0	2	- 2	2	2	- 2
15	2	13	-11	11	16	-16
16	18	10	8	8	12.5	12.5
17	1	7	- 6	6	8.5	- 8.5
18	2	19	-17	17	19	-19
19	4	20	-16	16	18	-18
20	1	15	-14	14	17	-17

Como hay una diferencia igual a cero (la cuarta planta), se opera con N=19.

El signo menos abundante es el positivo, y la suma total de dichos rangos es T = 40.

Se puede hacer la prueba por aproximación a al normal, por lo que se tiene

$$\mu_T = \frac{19\times 20}{4} = 95$$

$$\sigma_T^2 = \frac{19\times 20\times 39}{24} = 617.5$$

y la prueba es

$$Z_o = \frac{40-95}{\sqrt{617.5}} = -2.21^*$$

El *SAS* provee como estadístico de prueba el numerador de la anterior expresión, así como la significación.

Archivo del programa *SAS* (C20-17.SAS).-

```
title 'Prueba de Wilcoxon';
option ls =75 ps=60;
data fungi;
infile 'c20-17.dat';
input A B @@;
dif=A-B;
proc univariate;
var dif;
run;
```

Archivo de datos (C20-17.DAT).-

```
 8 16 12 13 14  7 16 16
 9  4  5  8  2 12  7  3
 6  9  6 12  2  9  4  1
 5 15  0  2  2 13 18 10
 1  7  2 19  4 20  1 15
```

Archivo de resultados (C20-17.LST).-

```
                          Prueba de Wilcoxon

                          Univariate Procedure

Variable=DIF

                                Moments
                N                 20    Sum Wgts          20
                Mean           -4.35    Sum              -87
                Std Dev     7.450009    Variance    55.50263
                Skewness    0.006383    Kurtosis    -0.89551
                USS             1433    CSS          1054.55
                CV          -171.265    Std Mean    1.665873
                T:Mean=0    -2.61124    Pr>|T|        0.0172
                Num ^= 0          19    Num > 0             5
                M(Sign)         -4.5    Pr>=|M|        0.0636
                Sgn Rank         -55    Pr>=|S|        0.0249
```

Pruebas de hipótesis no paramétricas para más de dos muestras.-

Siguiendo con el paralelismo con las pruebas anteriormente estudiadas, ahora toca estudiar el caso en que se tengan más de dos muestras, independientes o relacionadas.

Prueba de *Kruskal-Wallis* para más de dos muestras independientes.-

Esta prueba permite realizar pruebas unifactoriales o de una vía o de un diseño completamente aleatorio con datos que no cumplen las condiciones paramétricas de normalidad estudiadas en el Capítulo 6.

Las suposiciones para esta prueba son

- Las t muestras son muestras aleatorias de sus respectivas poblaciones y además son independientes entre sí.
- La variable medida es al menos ordinal.

Las hipótesis probadas son

H_o: Los efectos de los t tratamientos son iguales.
H_1: Al menos hay dos efectos diferentes.

El procedimiento para realizar esta prueba es el mismo del estudiado para la prueba de *Mann-Whitney*, es decir

Se ordenan todos los datos de las t muestras en una sola serie, asignando rangos del uno al N.
Se determina el valor de R (la suma de rangos) para cada una de las t muestras.
Si hay una gran cantidad de datos con el mismo valor ordinal hay que hacer una corrección.

Las prueba es

$$\chi^2 = \left(\frac{12}{N(N+1)} \Sigma \frac{R^2}{n_i} \right) - 3(N+1)$$

que se contrasta con el tabular para t-1 grados de libertad.

En el caso de muchos empates de la misma ordenación, la corrección de este estadístico es

$$\chi^2_{adj} = \frac{\chi^2}{1 - \frac{\Sigma (r_i^2 - r_i)}{N^3 - N}}$$

siendo r_i el número de observaciones corregidas en cada grupo de repeticiones.

Esta prueba es como un *ANOVA* pero con los rangos, en lugar de con los datos originales.

El *ANOVA* de los rangos da los mismos resultados que la prueba de *Kruskal-Wallis*. Si se quiere obtener el valor del χ^2 a partir del *ANOVA* de los rangos no se tiene más que multiplicar

$$(N - 1)R^2$$

Por otro lado, el *ANOVA* de los rangos permite, además, realizar cualquier contraste múltiple de medias para separar/agrupar tratamientos diferentes/iguales.

Ejemplo.-

Se está estudiando si los nacimientos humanos están influidos por la fase de la luna, para ello se contabilizan el número de nacimientos que se han producido durante un periodo de cuatro años en un hospital, siendo los cuatro *tratamientos* las cuatro fases de la luna tomadas de la siguiente manera: *fase A*, los cinco primeros días del ciclo; *fase B*, los cinco días siguientes (día 6 al 10); *fase C*, los diez días siguientes (día 11 al 20); y *fase D* los diez últimos días (día 21 al último).

Fase	*nacimientos por día*
A	273 273 291 287 270
B	282 293 276 260 271
C	243 292 261 279 268 270 276 290 278 303
D	241 252 268 294 265 305 280 285 264 287

Se le asigna rangos a las 30 observaciones sin tener en cuenta el grupo o tratamiento

Fase *rango*	*nacimientos por día* *rango*										*R*	R^2/n_i
A	273	273	291	287	270							
rango	13.5	13.5	25	22.5	10.5						85	1445
B	282	293	276	260	271							
rango	20	27	15.5	4	12						78.5	1232.45
C	243	292	261	279	268	270	276	290	278	303		
rango	2	26	5	18	8.5	10.5	15.5	24	17	29	155.5	2418.025
D	241	252	268	294	265	305	280	285	264	287		
rango	1	3	8.5	28	7	30	19	21	6	22.5	146	2131.6

$$\chi^2 = \left(\frac{12}{30x31}7227.075\right) - 3\times 31 = 0.25258 ns$$

$$\chi^2_{(3;\,0.05)} = 7.815$$

que es no significativo por tanto no influye la fase lunar en el número de nacimientos.

Si se necesitara hacer la corrección por repeticiones se tendría que se repiten dos veces cinco valores, el 273, 287, 270, 276 y el 268, por tanto

$$\chi^2_{adj} = \frac{0.25258}{1 - \frac{5(2(4-1))}{900-30}} = 0.2616ns$$

Archivo del programa *SAS* (C20-18.SAS).-

```
title 'Prueba de Kuskal-Wallis';
option ls=75 ps=60;
data unavia;
infile 'c20-18.dat';
input fase $ nacimien @@;
proc npar1way wilcoxon;
class fase;
var nacimien;
run;
title 'ANOVA de los ordinales o rangos';
proc rank out=rangos;
var nacimien;
ranks rnacimi;
run;
proc anova data=rangos;
class fase;
model rnacimi = fase;
means fase / snk;
run;
```

Archivo de datos (C20-18.DAT).-

```
A 273 A 273 A 291 A 287 A 270
B 282 B 293 B 276 B 260 B 271
C 243 C 292 C 261 C 279 C 268 C 270 C 276 C 290 C 278 C 303
D 241 D 252 D 268 D 294 D 265 D 305 D 280 D 285 D 264 D 287
```

Archivo de resultados (C20-18.LST).-

```
                        Prueba de Kuskal-Wallis

                  N P A R 1 W A Y   P R O C E D U R E
            Wilcoxon Scores (Rank Sums) for Variable NACIMIEN
                      Classified by Variable FASE

                       Sum of      Expected      Std Dev         Mean
   FASE          N      Scores      Under H0      Under H0        Score
   A             5   85.000000     77.500000   17.9598851   17.0000000
   B             5   78.500000     77.500000   17.9598851   15.7000000
   C            10  155.500000    155.000000   22.7176573   15.5500000
   D            10  146.000000    155.000000   22.7176573   14.6000000
                     Average Scores Were Used for Ties

          Kruskal-Wallis Test (Chi-Square Approximation)
          CHISQ = 0.25286     DF =  3     Prob > CHISQ = 0.9686

                      ANOVA de los ordinales o rangos

                        Analysis of Variance Procedure
Dependent Variable: RNACIMI    RANK FOR VARIABLE NACIMIEN
                                    Sum of          Mean
Source                    DF        Squares        Square  F Value     Pr > F
Model                      3     19.5750000     6.5250000     0.08     0.9723
Error                     26   2225.4250000    85.5932692
```

```
Corrected Total          29   2245.0000000

                  R-Square             C.V.      Root MSE        RNACIMI Mean
                  0.008719         59.68815       9.25166            15.5000

Source                   DF       Anova SS   Mean Square  F Value     Pr > F
FASE                      3    19.5750000     6.5250000     0.08     0.9723

              Student-Newman-Keuls test for variable: RNACIMI

        NOTE: This test controls the type I experimentwise error rate
              under the complete null hypothesis but not under partial
              null hypotheses.
                    Alpha= 0.05  df= 26  MSE= 85.59327
                    WARNING: Cell sizes are.not equal.
                  Harmonic Mean of cell sizes= 6.666667
               Number of Means          2          3          4
               Critical Range  10.416103 12.591814 13.901353
        Means with the same letter are not significantly different.

                SNK Grouping                 Mean      N  FASE

                           A               17.000      5  A
                           A               15.700      5  B
                           A               15.550     10  C
                           A               14.600     10  D
```

Prueba de *Bartlett* para homogeneidad de varianzas.-

Cuando se tienen varias poblaciones puede ser de interés el contrastar si las varianzas de estas poblaciones son iguales o no. Esta prueba se ha estudiado en el Capítulo 6.

Bloques aleatorios o muestras relacionadas para más de dos niveles del tratamiento.-

Al igual que ocurría con dos muestras, muchas veces se va a tener *t* (más de dos) muestras de igual tamaño igualadas o emparejadas de acuerdo con criterios susceptibles de afectar los valores de las observaciones. La igualación más clara es cuando se hace comparando los mismos individuos o casos bajo todas las *t* condiciones, o cada uno de los *n* grupos se miden en la *t* condiciones.

Prueba *Q* de *Cochran* para *t* muestras relacionadas.-

La prueba de *McNemar* estudiada anteriormente, puede extenderse para más de dos muestras. Esta prueba es muy adecuada cuando los datos son o se han dicotomizado. Por tanto, cuando los individuos han sido analizados tres o más veces para la *presencia/éxito* o *ausencia/fracaso* de algún atributo, se puede aplicar también en pruebas análogas al diseño de bloques aleatorios. Esta prueba se denomina *Prueba Q de Cochran.*

Para aplicar la prueba se construye una tabla de doble entrada con tantas columnas (*C*) como niveles presente el factor, y tantas filas (*F*) como individuos hayan sido sometidos a los niveles. A cada individuo le asignamos un 1 o un 0, para cada nivel, según que su respuesta al nivel haya sido *favorable* o *desfavorable*, *sí* o *no*, *éxito*

o *fracaso, etc.* Para cada columna se suma el número de éxitos (C_j), así como para cada fila (F_i); t es el número de niveles del factor (columnas) y n el número de individuos (filas). La fórmula es

$$Q = \frac{(t-1)\left[t\Sigma_j C_j^2 - (\Sigma_j C_j)^2\right]}{t\Sigma_i F_i - \Sigma_i F_i^2}$$

la cual se ajusta a una distribución χ^2 con $gl = t\text{-}1$.

Ejemplo.-

Una empresa láctea quiere saber si cuatro aditivos edulcorantes de la leche son igualmente aceptados por los niños; para contrastar esta hipótesis se toma una muestra de seis niños y se les da sucesivamente un vaso de leche con cada tipo de edulcorante, debiendo contestar cada niño si le ha gustado o no. El número de niveles es t=4 (columnas) y el de individuos es 6 (filas). Los datos son

Niño	*Aditivo I*	*Aditivo II*	*Aditivo III*	*Aditivo IV*	F_i
1	0	0	1	0	1
2	0	0	1	1	2
3	0	1	1	1	3
4	1	1	0	0	2
5	1	0	0	1	2
6	0	1	1	0	2
C_j	2	3	4	3	

$$\Sigma_i F_i = 12$$
$$\Sigma_i F_i^2 = 26$$
$$\Sigma_j C_j^2 = 38$$
$$\Sigma_j (C_j)^2 = 144$$
$$Q = \frac{(4-1)(4\times 38 - 144)}{4\times 12 - 26} = 1.091 ns$$
$$\chi^2_{(3;\,0.05)} = 7.815$$

Por lo que no hay evidencia que permita concluir que existe preferencia por alguno de los cuatro tipos de edulcorantes.

Este estadístico es el número 3 que ofrece el *SAS* en la salida de la opción ***cmh*** del procedimiento ***FREQ***.

Archivo del programa *SAS* (C20-19.SAS).-

```
title 'Prueba de Cochran';
option ls=75 ps=60;
data qcochran;
infile 'c20-19.dat';
input nino aditivo $ gusto @@;
proc freq;
table nino*aditivo*gusto/noprint cmh;
run;
```

Archivo de datos (C20-19.DAT).-

```
1 I 0  1 II 0  1 III 1  1 IV 0
2 I 0  2 II 0  2 III 1  2 IV 1
3 I 0  3 II 1  3 III 1  3 IV 1
4 I 1  4 II 1  4 III 0  4 IV 0
5 I 1  5 II 0  5 III 0  5 IV 1
6 I 0  6 II 1  6 III 1  6 IV 0
```

Archivo de resultados (C20-19.LST).-

```
                        Prueba de Cochran
                SUMMARY STATISTICS FOR ADITIVO BY GUSTO
                        CONTROLLING FOR NINO

      Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

    Statistic   Alternative Hypothesis     DF       Value      Prob
    ------------------------------------------------------------------
       1        Nonzero Correlation         1       0.436     0.509
       2        Row Mean Scores Differ      3       1.091     0.779
       3        General Association         3       1.091     0.779
     Total Sample Size = 24
```

Prueba de *Friedman* para diseños de bloques aleatorios.-

Cuando los datos de las t muestras están, por lo menos, en una escala ordinal, se puede hacer un análisis con los rangos, esta prueba se denomina de *Friedman*. El procedimiento consiste en

- Darle el valor ordinal a cada medida pero dentro de los bloques e independientemente de los demás bloques
- Se calcula la suma de los rangos de cada tratamiento.
- El estadístico de prueba es

$$\chi^2 = \frac{12}{rt(t+1)} \sum R_i^2 - 3r(t+1)$$

- Se contrasta para $gl=t-1$

Ejemplo.-

En un experimento para comparar la duración del dibujo de cuatro tipos de neumáticos se toman cuatro coches, y a cada uno se le asigna, aleatoriamente, un tipo

de neumático para cada rueda. Los resultados en kilómetros recorridos son

		Neumático			
		A	*B*	*C*	*D*
Coche	1	42000	40800	49000	38000
	2	43613	44000	51000	36000
	3	42830	51600	50700	39318
	4	49800	39720	49900	40642

se les asigna los rangos dentro de cada bloque (coche).

		Neumático			
		A	*B*	*C*	*D*
Coche	1	3	2	4	1
	2	2	3	4	1
	3	2	4	3	1
	4	3	1	4	2
	R_i	10	10	15	5

$$\chi^2 = \frac{12}{4\times4\times5}\,(10^2 + 10^2 + 15^2 + 5^2) - 3\times4\times5 = 7.5ns$$

$$\chi^2_{(3;\,0.05)} = 7.815$$

Está en el borde del nivel crítico, pero no lo supera, por lo que se concluyen que los cuatro neumáticos duran los mismos kilómetros.

Esta prueba la da la salida 2 de la opción **CMH** del procedimiento **FREQ** del *SAS*.

Archivo del programa *SAS* (C20-20.SAS).-

```
title 'Prueba de Friedman';
option ls=75 ps=60;
data dosvias;
infile 'c20-20.dat';
input coche llanta $  kilome @@;
proc sort;
by coche;
run;
proc rank;
by coche;
var kilome;
ranks rkilome;
run;
proc freq;
 table coche*llanta*rkilome /noprint cmh;
run;
```

Archivo de datos (C20-20.DAT).-

```
1 A 42000  1 B 40800  1 C 49000  1 D 38000
2 A 43613  2 B 44000  2 C 51000  2 D 36000
3 A 42830  3 B 51600  3 C 50700  3 D 39318
4 A 49800  4 B 39720  4 C 49900  4 D 40642
```

Archivo de resultados (C20-20.LST).-

```
                        Prueba de Friedman

                SUMMARY STATISTICS FOR LLANTA BY RKILOME
                        CONTROLLING FOR COCHE

        Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

      Statistic   Alternative Hypothesis     DF        Value       Prob
      ---------------------------------------------------------------
         1        Nonzero Correlation         1        0.750      0.386
         2        Row Mean Scores Differ      3        7.500      0.058
         3        General Association         9       12.000      0.213
       Total Sample Size = 16
```

Pruebas no paramétricas con dos variables. Correlación no paramétrica.-

Muchas veces una población bivariante se sabe que no se distribuye normalmente por lo que en este caso no tiene sentido el cálculo de *r* como una estima del parámetro ρ. En este caso se utiliza una correlación no paramétrica. Estas correlaciones son

Coeficiente de correlación de rangos de *Spearman*.-
Coeficiente de correlación de clasificación de *Kendall*.-
Coeficiente de correlación serial.-

El coeficiente de correlación de Spearman y el de Kendall se estudiaron en el Capítulo 15 y el coeficiente de correlación serial se ha estudiado en el Capítulo 6

Bibliografía

Afifi, A.A. y Clark, V. 1984. COMPUTER-AIDED MULIVARIATE ANALYSIS. Ed: Lifetime Learning Publications. Belmont (USA).

Agresti, A. 1990. CATEGORICAL DATA ANALYSIS. Ed. John Wiley & Sons, Inc. New York.

Bisquerra Alzina, R. 1989. INTRODUCCIÓN CONCEPTUAL AL ANÁLISIS MULTIVARIABLE. Ed: PPU. Barcelona (España).

Cuadras, C.M. 1981. MÉTODOS DE ANÁLISIS MULTIVARIANTE. Ed:EUNIBAR. Barcelona (España).

Dagnelie, P. 1982. ANALYSE STATISTIQUE À PLUSIEURS VARIABLES. Ed: Les Presses Agronomiques De Gembloux. Gembloux (Belgique).

González López-Valcárcel, B. 1991. ANÁLISIS MULTIVARIANTE: APLICACIÓN AL ÁMBITO SANITARIO. Ed: SG Editores. Barcelona (España).

Judez Asensio, L. 1989. TÉCNICAS DE ANÁLISIS DE DATOS MULTIDIMENSIONALES. Ed: MAPA. Madrid (España).

Lebart, L., Morineau, A. Y Fénelon, J.P. 1979. TRAITEMENT DES DONNÉES STATISTIQUES. Ed: Dunod. Paris (France).

Lefebvre, J. 1980. INRODUCTION AUX ANALYSES STATISTIQUES MULTIDIMENSIONNELLES. Ed: Masson. Paris (France)

Milton, J.S. 1994. ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD. Ed. Interamericana-McGraw-Hill. México.

Siegel, S. 1986. ESTADÍSTICA NO PARAMÉTRICA. Ed. TRILLAS. México.

Sokal, R.R. y Rohlf, F.J. 1994. BIOMETY. Ed. W.H.FREEMAN. San Francisco.

Snedecor, G.W. y Cochran, W.G. 1971. MÉTODOS ESTADÍSTICOS. Ed. C.E.C.S.A. México.

Spiegel M.R. 1990. ESTADÍSTICA. Ed. McGraw-Hill. Madrid.

Stanish, W.M. y Stokes, M. 1991. CATEGORICAL DATA ANALYSIS. COURSE NOTES. *SAS* Institute Inc. *SAS* Campus Drive. Cary, NC, USA.

Steel, R. 1996. PRINCIPLES AND PROCEDURES OF STATISTICS. Ed. McGRAW-HILL Education. New York .

*Srivastava, M.S. y Carter, E.M.*1983. AN INTRODUCTION TO APPLIED MULTIVARIATE STATISTICS. Ed:Elsevier Sciencie Publishing. New York (USA).

SAS Institute Inc. 1990. *SAS PROCEDURE* GUIDE. Cary, NC, USA.

SAS Institute Inc. 1990. *SAS/STAT* USER'S GUIDE. Volume 1 and 2. Cary, NC, USA.

Tablas Estadísticas

TABLAS ESTADÍSTICAS

INDICE

TABLA 1

Áreas bajo la curva Normal tipificada de 0 a *Z*.

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1627	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2122	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2356	0.2389	0.2421	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2703	0.2734	0.2764	0.2793	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3079	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3414	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3622
1.1	0.3643	0.3665	0.3687	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4083	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4193	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4485	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4648	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4874	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4895	0.4898	0.4901	0.4903	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4924	0.4926	0.4928	0.4930	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4944	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4958	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4986	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Límites de significación de la distribución de Student (t)

2α	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
α	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
gl 1	0.7270	1.0000	1.3760	1.9630	3.0780	6.3138	12.706	31.821	63.657	636.62
2	0.6172	0.8165	1.0610	1.3860	1.8860	2.9200	4.3027	6.9650	9.9248	31.598
3	0.5840	0.7649	0.9780	1.2500	1.6380	2.3534	3.1825	4.5410	5.8409	12.924
4	0.5692	0.7407	0.9410	1.1900	1.5330	2.1318	2.7764	3.7470	4.6041	8.6100
5	0.5598	0.7267	0.9200	1.1560	1.4760	2.0150	2.5706	3.3650	4.0321	6.8690
6	0.5536	0.7176	0.9060	1.1340	1.4400	1.9432	2.4469	3.1430	3.7074	5.9590
7	0.5493	0.7111	0.8960	1.1190	1.4150	1.8946	2.3646	2.9980	3.4995	5.4080
8	0.5461	0.7064	0.8890	1.1080	1.3970	1.8595	2.3060	2.8960	3.3554	5.0410
9	0.5436	0.7027	0.8830	1.1000	1.3830	1.8331	2.2622	2.8210	3.2498	4.7810
10	0.5416	0.6998	0.8790	1.0930	1.3720	1.8125	2.2281	2.7640	3.1693	4.5870
11	0.5400	0.6975	0.8760	1.0880	1.3630	1.7959	2.2010	2.7180	3.1058	4.4370
12	0.5387	0.6955	0.8730	1.0830	1.3560	1.7823	2.1788	2.6810	3.0545	4.3180
13	0.5375	0.6938	0.8700	1.0790	1.3500	1.7709	2.1604	2.6500	3.0123	4.2210
14	0.5366	0.6924	0.8680	1.0760	1.3450	1.7613	2.1448	2.6240	2.9768	4.1400
15	0.5358	0.6912	0.8660	1.0740	1.3410	1.7530	2.1315	2.6020	2.9467	4.0730
16	0.5350	0.6901	0.8650	1.0710	1.3370	1.7459	2.1199	2.5830	2.9208	4.0150
17	0.5344	0.6892	0.8630	1.0690	1.3330	1.7396	2.1098	2.5670	2.8982	3.9650
18	0.5338	0.6884	0.8620	1.0670	1.3300	1.7341	2.1009	2.5520	2.8784	3.9220
19	0.5333	0.6876	0.8610	1.0660	1.3280	1.7291	2.0930	2.5390	2.8609	3.8830
20	0.5329	0.6870	0.8600	1.0640	1.3250	1.7247	2.0860	2.5280	2.8453	3.8500
21	0.5325	0.6864	0.8590	1.0630	1.3230	1.7207	2.0796	2.5180	2.8314	3.8190
22	0.5321	0.6858	0.8580	1.0610	1.3210	1.7171	2.0739	2.5080	2.8188	3.7920
23	0.5318	0.6853	0.8580	1.0600	1.3190	1.7139	2.0687	2.5000	2.8073	3.7670
24	0.5315	0.6849	0.8570	1.0590	1.3180	1.7109	2.0639	2.4920	2.7969	3.7450
25	0.5312	0.6844	0.8560	1.0580	1.3160	1.7081	2.0595	2.4850	2.7874	3.7250
26	0.5309	0.6841	0.8560	1.0580	1.3150	1.7056	2.0555	2.4790	2.7787	3.7070
27	0.5307	0.6837	0.8550	1.0570	1.3140	1.7033	2.0518	2.4730	2.7707	3.6900
28	0.5304	0.6834	0.8550	1.0560	1.3130	1.7011	2.0484	2.4670	2.7633	3.6740
29	0.5302	0.6830	0.8540	1.0550	1.3110	1.6991	2.0452	2.4620	2.7564	3.6590
30	0.5300	0.6828	0.8540	1.0550	1.3100	1.6973	2.0423	2.4570	2.7500	3.6460
31	0.5298	0.6825	0.8535	1.0541	1.3095	1.6955	2.0395	2.4530	2.7441	3.6338
32	0.5297	0.6823	0.8531	1.0536	1.3086	1.6939	2.0370	2.4490	2.7385	3.6221
33	0.5295	0.6820	0.8527	1.0531	1.3078	1.6924	2.0345	2.4450	2.7333	3.6111
34	0.5294	0.6818	0.8524	1.0526	1.3070	1.6909	2.0323	2.4410	2.7284	3.6011

TABLA 2 (Cont.)

Límites de significación de la distribución de Student (t)

2α	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
α	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
gl 35	0.5292	0.6816	0.8521	1.0521	1.3062	1.6896	2.0301	2.4380	2.7239	3.5915
36	0.5291	0.6814	0.8518	1.0516	1.3055	1.6883	2.0281	2.4340	2.7195	3.5824
37	0.5290	0.6812	0.8515	1.0512	1.3049	1.6871	2.0262	2.4310	2.7155	3.5741
38	0.5288	0.6810	0.8512	1.0508	1.3042	1.6860	2.0244	2.4280	2.7116	3.5661
39	0.5287	0.6808	0.8510	1.0504	1.3037	1.6849	2.0227	2.4260	2.7079	3.5586
40	0.5286	0.6807	0.8507	1.0501	1.3031	1.6839	2.0211	2.4230	2.7045	3.5511
41	0.5285	0.6805	0.8505	1.0498	1.3026	1.6829	2.0196	2.4210	2.7012	3.5446
42	0.5284	0.6804	0.8503	1.0494	1.3020	1.6820	2.0181	2.4180	2.6981	3.5383
43	0.5283	0.6803	0.8501	1.0491	1.3016	1.6811	2.0167	2.4160	2.6952	3.5323
44	0.5282	0.6801	0.8499	1.0488	1.3011	1.6802	2.0154	2.4140	2.6923	3.5264
45	0.5281	0.6800	0.8497	1.0485	1.3007	1.6794	2.0141	2.4120	2.6896	3.5207
46	0.5281	0.6799	0.8495	1.0483	1.3002	1.6787	2.0129	2.4100	2.6870	3.5153
47	0.5280	0.6798	0.8494	1.0480	1.2998	1.6779	2.0118	2.4080	2.6846	3.5104
48	0.5279	0.6796	0.8492	1.0478	1.2994	1.6772	2.0106	2.4060	2.6822	3.5053
49	0.5278	0.6795	0.8490	1.0476	1.2991	1.6766	2.0096	2.4050	2.6800	3.5010
50	0.5278	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4030	2.6878	3.4965
51	0.5277	0.6793	0.8488	1.0471	1.2984	1.6753	2.0077	2.4020	2.6758	3.4924
52	0.5276	0.6792	0.8486	1.0469	1.2981	1.6747	2.0067	2.4000	2.6738	3.4883
53	0.5276	0.6792	0.8485	1.0467	1.2978	1.6742	2.0058	2.3990	2.6719	3.4845
54	0.5275	0.6791	0.8484	1.0465	1.2975	1.6736	2.0049	2.3970	2.6700	3.4807
55	0.5275	0.6790	0.8483	1.0463	1.2972	1.6731	2.0041	2.3960	2.6683	3.4770
56	0.5274	0.6789	0.8481	1.0461	1.2969	1.6725	2.0033	2.3950	2.6666	3.4733
57	0.5274	0.6789	0.8480	1.0460	1.2967	1.6721	2.0025	2.3930	2.6650	3.4702
58	0.5273	0.6788	0.8479	1.0458	1.2964	1.6716	2.0017	2.3920	2.6633	3.4670
59	0.5273	0.6787	0.8478	1.0457	1.2962	1.6712	2.0010	2.3910	2.6618	3.4638
60	0.5272	0.6786	0.8477	1.0455	1.2959	1.6707	2.0003	2.3900	2.6603	3.4606
61	0.5272	0.6786	0.8476	1.0454	1.2957	1.6703	1.9997	2.3890	2.6590	3.4577
62	0.5271	0.6785	0.8475	1.0452	1.2954	1.6698	1.9990	2.3880	2.6576	3.4548
63	0.5271	0.6785	0.8474	1.0451	1.2952	1.6694	1.9984	2.3870	2.6563	3.4521
64	0.5270	0.6784	0.8473	1.0449	1.2950	1.6690	1.9977	2.3860	2.6549	3.4494

Límites de significación de la distribución de Student (*t*)

2α	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
α	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
gl 65	0.5270	0.6783	0.8472	1.0448	1.2948	1.6687	1.9972	2.3850	2.6537	3.4470
66	0.5270	0.6782	0.8471	1.0447	1.2945	1.6683	1.9966	2.3840	2.6525	3.4445
67	0.5270	0.6782	0.8471	1.0446	1.2944	1.6680	1.9961	2.3830	2.6513	3.4423
68	0.5269	0.6781	0.8470	1.0444	1.2942	1.6676	1.9955	2.3820	2.6501	3.4400
69	0.5269	0.6781	0.8469	1.0443	1.2940	1.6673	1.9950	2.3810	2.6491	3.4378
70	0.5268	0.6780	0.8468	1.0442	1.2938	1.6669	1.9945	2.3810	2.6480	3.4355
71	0.5268	0.6780	0.8468	1.0441	1.2936	1.6666	1.9940	2.3800	2.6470	3.4333
72	0.5267	0.6779	0.8467	1.0440	1.2934	1.6663	1.9935	2.3790	2.6459	3.4310
73	0.5267	0.6779	0.8466	1.0439	1.2933	1.6660	1.9931	2.3780	2.6450	3.4291
74	0.5267	0.6778	0.8465	1.0438	1.2931	1.6657	1.9926	2.3780	2.6440	3.4272
75	0.5267	0.6778	0.8465	1.0437	1.2930	1.6655	1.9922	2.3770	2.6431	3.4253
76	0.5266	0.6777	0.8464	1.0436	1.2928	1.6652	1.9917	2.3760	2.6421	3.4234
77	0.5266	0.6777	0.8464	1.0435	1.2927	1.6649	1.9913	2.3760	2.6413	3.4217
78	0.5266	0.6777	0.8463	1.0434	1.2925	1.6646	1.9909	2.3750	2.6404	3.4200
79	0.5266	0.6777	0.8463	1.0433	1.2924	1.6644	1.9905	2.3740	2.6396	3.4185
80	0.5265	0.6776	0.8462	1.0432	1.2922	1.6641	1.9901	2.3740	2.6388	3.4169
81	0.5265	0.6776	0.8461	1.0431	1.2921	1.6639	1.9897	2.3730	2.6380	3.4152
82	0.5265	0.6775	0.8460	1.0430	1.2920	1.6637	1.9893	2.3720	2.6372	3.4135
83	0.5265	0.6775	0.8460	1.0430	1.2919	1.6635	1.9890	2.3720	2.6365	3.4121
84	0.5264	0.6774	0.8459	1.0429	1.2917	1.6632	1.9886	2.3710	2.6357	3.4106
85	0.5264	0.6774	0.8459	1.0428	1.2916	1.6630	1.9883	2.3710	2.6350	3.4091
86	0.5264	0.6774	0.8458	1.0427	1.2915	1.6628	1.9880	2.3700	2.6343	3.4076
87	0.5264	0.6774	0.8458	1.0427	1.2914	1.6626	1.9877	2.3700	2.6336	3.4063
88	0.5263	0.6773	0.8457	1.0426	1.2913	1.6624	1.9873	2.3690	2.6329	3.4050
89	0.5263	0.6773	0.8457	1.0426	1.2912	1.6622	1.9870	2.3690	2.6323	3.4036
90	0.5263	0.6772	0.8457	1.0425	1.2910	1.6620	1.9867	2.3680	2.6316	3.4022
91	0.5263	0.6772	0.8457	1.0424	1.2909	1.6618	1.9864	2.3680	2.6310	3.4010
92	0.5262	0.6772	0.8456	1.0423	1.2908	1.6616	1.9861	2.3670	2.6303	3.3997
93	0.5262	0.6772	0.8456	1.0423	1.2907	1.6614	1.9859	2.3670	2.6298	3.3986
94	0.5262	0.6771	0.8456	1.0422	1.2906	1.6612	1.9856	2.3660	2.6292	3.3975
95	0.5262	0.6771	0.8454	1.0422	1.2905	1.6611	1.9853	2.3660	2.6286	3.3964
96	0.5262	0.6771	0.8454	1.0421	1.2904	1.6609	1.9850	2.3660	2.6280	3.3952
97	0.5262	0.6771	0.8454	1.0421	1.2904	1.6608	1.9848	2.3650	2.6275	3.3940
98	0.5261	0.6770	0.8453	1.0420	1.2903	1.6606	1.9845	2.3650	2.6270	3.3928
99	0.5261	0.6770	0.8453	1.0419	1.2902	1.6604	1.9843	2.3640	2.6265	3.3919
100	0.5261	0.6770	0.8452	1.0418	1.2901	1.6602	1.9840	2.3640	2.6260	3.3909

TABLA 3

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 1	2	3	4	5	6	7	8	9	10
	0.05	161	199	216	225	230	234	237	239	241	241
	0.025	648	800	864	900	922	937	948	957	963	969
1	0.01	4050	5000	5400	5620	5760	5860	5930	5980	6020	6060
	0.005	16200	20000	21600	22500	23100	23400	23700	23900	24100	24200
	0.001	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*
	0.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
	0.025	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4
2	0.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
	0.005	198	199	199	199	199	199	199	199	199	199
	0.001	999	999	999	999	999	999	999	999	999	999
	0.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	0.025	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4
3	0.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
	0.005	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7
	0.001	167	149	141	137	135	133	132	131	130	129
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	0.025	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
4	0.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
	0.005	31.3	26.3	24.3	23.2	22.4	22.0	21.6	21.4	21.1	21.0
	0.001	74.1	61.3	56.2	53.4	51.7	50.5	49.7	49.0	48.5	48.1
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	0.025	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
5	0.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
	0.005	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6
	0.001	47.2	37.1	33.2	31.1	29.7	28.8	28.2	27.6	27.2	26.9
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
6	0.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	0.005	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3
	0.001	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7	18.4
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.77	3.73	3.68	3.64
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.89	4.82	4.76
7	0.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	0.005	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.52	8.38
	0.001	29.3	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.3	14.1

* = x 1000

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 11	12	15	20	30	40	50	60	120	∞
1	0.05	243	244	246	248	250	251	252	252	253	254
	0.025	973	977	985	993	1000	1010	1010	1010	1010	1020
	0.01	6080	6110	6160	6210	6260	6290	6300	6310	6340	6370
	0.005	24300	24400	24630	24836	25440	25148	25211	25253	25359	25465
	0.001	6084*	6107*	6158*	6209*	6261*	6287*	6303*	6313*	6340*	6366*
2	0.05	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
	0.025	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5
	0.01	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
	0.005	199	199	199	199	199	199	199	199	199	200
	0.001	999	999	999	999	1000	1000	1000	1000	1000	1000
3	0.05	8.76	8.74	8.70	8.66	8.62	8.59	8.58	8.57	8.55	8.53
	0.025	14.3	14.3	14.3	14.2	14.1	14.0	14.0	14.0	13.9	13.9
	0.01	27.1	27.1	26.9	26.7	26.5	26.4	26.3	26.3	26.2	26.1
	0.005	43.5	43.4	43.1	42.8	42.5	42.3	42.2	42.1	42.0	41.8
	0.001	128	128	127	126	125	125	125	124	124	124
4	0.05	5.93	5.91	5.86	5.80	5.75	5.72	5.70	5.69	5.66	5.63
	0.025	8.79	8.75	8.66	8.56	8.46	8.41	8.38	8.36	8.31	8.26
	0.01	14.4	14.4	14.2	14.0	13.8	13.7	13.7	13.7	13.6	13.5
	0.005	20.8	20.7	20.4	20.2	19.9	19.8	19.7	19.6	19.5	19.3
	0.001	47.7	47.4	46.8	46.1	45.4	45.1	44.9	44.8	44.4	44.0
5	0.05	4.71	4.68	4.62	4.56	4.50	4.46	4.44	4.43	4.40	4.36
	0.025	6.57	6.52	6.43	6.33	6.23	6.18	6.14	6.12	6.07	6.02
	0.01	9.99	9.89	9.72	9.55	9.38	9.29	9.24	9.20	9.11	9.02
	0.005	13.5	13.4	13.1	12.9	12.7	12.5	12.4	12.4	12.3	12.1
	0.001	26.6	26.4	25.9	25.4	24.9	24.6	24.4	24.3	24.1	23.7
6	0.05	4.03	4.00	3.94	3.87	3.81	3.77	3.75	3.74	3.70	3.67
	0.025	5.41	5.37	5.27	5.17	5.07	5.01	4.98	4.96	4.90	4.85
	0.01	7.79	7.72	7.56	7.40	7.23	7.14	7.09	7.06	6.97	6.88
	0.005	10.1	10.0	9.81	9.59	9.36	9.24	9.17	9.12	9.00	8.88
	0.001	18.2	18.0	17.6	17.1	16.7	16.4	16.3	16.2	16.0	15.8
7	0.05	3.60	3.57	3.51	3.44	3.38	3.34	3.32	3.30	3.27	3.23
	0.025	4.71	4.67	4.57	4.47	4.36	4.31	4.27	4.25	4.20	4.14
	0.01	6.54	6.47	6.31	6.16	5.99	5.91	5.86	5.82	5.74	5.65
	0.005	8.27	8.18	7.97	7.75	7.53	7.42	7.35	7.31	7.19	7.08
	0.001	13.9	13.7	13.3	12.9	12.5	12.3	12.2	12.1	11.9	11.7

* = x 1000

TABLA 3 (Cont.)

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 1	2	3	4	5	6	7	8	9	10
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
8	0.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	0.005	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21
	0.001	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8	11.5
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
9	0.01	10.5	8.02	6.99	6.42	6.06	5.87	5.61	5.47	5.35	5.26
	0.005	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42
	0.001	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1	9.79
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
10	0.01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	0.005	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85
	0.001	21.0	14.9	12.5	11.3	10.5	9.92	9.52	9.20	8.96	8.75
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
11	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	0.005	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42
	0.001	19.7	13.8	11.6	10.3	9.58	9.05	8.66	8.35	8.12	7.92
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
12	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	0.005	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09
	0.001	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48	7.29
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
13	0.01	9.07	6.70	5.74	4.21	4.86	4.62	4.44	4.30	4.19	4.10
	0.005	11.4	8.19	6.93	5.23	5.79	5.48	5.25	5.08	4.94	4.82
	0.001	17.8	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.98	6.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.20	3.15
14	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	0.005	11.1	7.92	6.68	6.00	5.53	5.26	5.03	4.86	4.72	4.60
	0.001	17.1	11.8	9.73	8.62	7.92	7.43	7.08	6.80	6.58	6.40

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 11	12	15	20	30	40	50	60	120	∞
	0.05	3.31	3.28	3.22	3.15	3.08	3.04	3.02	3.01	2.97	2.93
	0.025	4.25	4.20	4.10	4.00	3.89	3.84	3.80	3.78	3.73	3.67
8	0.01	5.73	5.67	5.52	5.36	5.20	5.12	5.70	5.03	4.95	4.86
	0.005	7.10	7.01	6.81	6.61	6.40	6.29	6.22	6.18	6.06	5.95
	0.001	11.3	11.2	10.8	10.5	10.1	9.9	9.8	9.7	9.5	9.3
	0.05	3.10	3.07	3.01	2.94	2.86	2.83	2.81	2.79	2.75	2.71
	0.025	3.91	3.87	3.77	3.67	3.56	3.51	3.47	3.45	3.39	3.33
9	0.01	5.18	5.11	4.96	4.81	4.65	4.57	4.52	4.48	4.40	4.31
	0.005	6.32	6.23	6.03	5.83	5.62	5.52	5.45	5.41	5.30	5.19
	0.001	9.72	9.57	9.24	8.90	8.55	8.37	8.26	8.19	8.00	7.81
	0.05	2.94	2.91	2.85	2.77	2.70	2.66	2.64	2.62	2.58	2.54
	0.025	3.67	3.62	3.52	3.42	3.31	3.26	3.22	3.20	3.14	3.08
10	0.01	4.77	4.71	4.56	4.41	4.25	4.17	4.12	4.08	4.00	3.91
	0.005	5.75	5.66	5.47	5.27	5.07	4.97	4.90	4.86	4.75	4.64
	0.001	8.59	8.45	8.13	7.80	7.47	7.30	7.19	7.12	6.94	6.76
	0.05	2.82	2.79	2.72	2.65	2.57	2.53	2.51	2.49	2.45	2.40
	0.025	3.48	3.43	3.33	3.23	3.12	3.06	3.02	3.00	2.94	2.88
11	0.01	4.46	4.40	4.25	4.10	3.94	3.86	3.81	3.78	3.69	3.60
	0.005	5.32	5.24	5.05	4.86	4.65	4.55	4.49	4.45	4.34	4.23
	0.001	7.76	7.63	7.32	7.01	6.68	6.52	6.42	6.35	6.17	6.00
	0.05	2.72	2.69	2.62	2.54	2.47	2.43	2.40	2.38	2.34	2.30
	0.025	3.32	3.28	3.18	3.07	2.96	2.91	2.87	2.85	2.79	2.72
12	0.01	4.22	4.16	4.01	3.86	3.70	3.62	3.57	3.54	3.45	3.36
	0.005	4.99	4.91	4.72	4.53	4.33	4.23	4.16	4.12	4.01	3.90
	0.001	7.14	7.00	6.71	6.40	6.09	5.93	5.83	5.76	5.59	5.42
	0.05	2.64	2.60	2.53	2.46	2.38	2.34	2.31	2.30	2.25	2.21
	0.025	3.20	3.15	3.05	2.95	2.84	2.78	2.74	2.72	2.66	2.60
13	0.01	4.03	3.96	3.82	3.66	3.51	3.43	3.37	3.34	3.25	3.17
	0.005	4.73	4.64	4.46	4.27	4.07	3.97	3.91	3.87	3.76	2.65
	0.001	6.65	6.52	6.23	5.93	5.63	5.47	5.37	5.30	5.14	4.97
	0.05	2.57	2.53	2.46	2.39	2.31	2.27	2.24	2.22	2.18	2.13
	0.025	3.10	3.05	2.95	2.84	2.73	2.67	2.64	2.61	2.55	2.49
14	0.01	3.86	3.80	3.66	3.51	3.35	3.27	3.21	3.18	3.09	3.00
	0.005	4.51	4.43	4.25	4.06	3.86	3.76	3.70	3.66	3.55	3.44
	0.001	6.26	6.13	5.85	5.56	5.25	5.10	5.00	4.94	4.77	4.60

TABLA 3 (Cont.)

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	*gl del numerador* 1	2	3	4	5	6	7	8	9	10
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.06
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
15	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	0.005	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42
	0.001	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
16	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	0.005	10.6	7.51	6.30	5.64	5.21	4.91	4.69	4.52	5.38	4.27
	0.001	16.1	11.0	9.00	7.94	7.27	6.81	6.46	6.19	5.98	5.81
	0.05	4.45	3.59	3.20	2.96	2.91	2.70	3.61	2.55	2.49	2.45
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
17	0.01	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	0.005	10.4	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14
	0.001	15.7	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
18	0.01	8.28	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	0.005	10.2	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03
	0.001	15.4	10.4	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.76	2.88	2.82
19	0.01	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	0.005	10.1	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93
	0.001	15.1	10.2	8.28	7.26	6.62	6.18	5.85	5.59	5.39	5.22
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
20	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	0.005	9.94	6.99	5.82	5.17	4.75	4.47	4.26	4.09	3.96	3.85
	0.001	14.8	9.95	8.10	7.10	6.45	6.02	5.69	5.44	5.24	5.08
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
21	0.01	8.02	5.75	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	0.005	9.83	6.89	5.73	5.09	4.99	4.39	4.18	4.01	3.88	3.77
	0.001	14.6	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 11	12	15	20	30	40	50	60	120	∞
	0.05	2.04	2.02	2.40	2.33	2.25	2.20	2.18	2.16	2.11	2.07
	0.025	3.01	2.96	2.86	2.76	2.64	2.59	2.55	2.52	2.46	2.40
15	0.01	3.73	3.67	3.52	3.37	3.21	3.13	3.08	3.05	2.96	2.87
	0.005	4.33	4.25	4.07	3.88	3.69	3.59	3.52	3.48	3.37	3.26
	0.001	5.94	5.81	5.54	5.25	4.95	4.80	4.70	4.64	4.47	3.31
	0.05	2.46	2.42	2.35	2.28	2.19	2.15	2.12	2.11	2.06	2.01
	0.025	2.93	2.89	2.79	2.68	2.57	2.51	2.47	2.45	2.38	2.32
16	0.01	3.62	3.55	3.41	3.26	3.10	3.02	2.97	2.93	2.84	2.75
	0.005	4.18	4.10	3.92	3.73	3.54	3.44	3.37	3.33	3.22	3.11
	0.001	5.67	5.55	5.27	4.99	4.70	4.54	4.45	4.39	4.23	4.06
	0.05	2.41	2.38	2.31	2.23	2.15	2.10	2.08	2.06	2.01	1.96
	0.025	2.87	2.82	2.72	2.62	2.50	2.44	2.41	2.38	2.32	2.25
17	0.01	3.52	3.46	3.31	3.16	3.00	2.92	2.87	2.83	2.75	2.65
	0.005	4.05	3.97	3.79	3.61	3.41	3.31	3.25	3.21	3.10	2.98
	0.001	5.44	5.32	5.05	4.78	4.48	4.33	4.24	4.18	4.02	3.85
	0.05	2.37	2.34	2.27	2.19	2.11	2.06	2.04	2.02	1.97	1.92
	0.025	2.81	2.77	2.67	2.56	2.44	2.38	2.35	2.32	2.26	2.19
18	0.01	3.43	3.37	3.23	3.08	2.92	2.84	2.78	2.75	2.66	2.57
	0.005	3.94	3.86	3.68	3.50	3.30	3.20	3.14	3.10	2.99	2.87
	0.001	5.25	5.13	4.87	4.59	4.30	4.15	4.06	4.00	3.84	3.67
	0.05	2.34	2.31	2.23	2.16	2.07	2.03	2.00	1.98	1.93	1.88
	0.025	2.76	2.72	2.62	2.51	2.39	2.33	2.30	2.27	2.20	2.13
19	0.01	3.56	3.30	3.15	3.00	2.84	2.76	2.71	2.67	2.58	2.78
	0.005	3.84	3.76	3.59	3.40	3.21	3.11	3.04	3.00	2.89	2.78
	0.001	5.08	4.97	4.70	4.43	4.14	3.99	3.90	3.84	3.68	3.51
	0.05	2.31	2.28	2.20	2.12	2.04	1.99	1.97	1.95	1.90	1.84
	0.025	2.72	2.68	2.57	2.46	2.35	2.29	2.25	2.22	2.16	2.09
20	0.01	3.29	3.23	3.09	2.94	2.78	2.69	2.64	2.61	2.52	2.42
	0.005	3.76	3.68	3.50	3.32	3.12	3.02	2.96	2.92	2.81	2.69
	0.001	4.94	4.82	4.56	4.29	4.00	3.86	3.76	3.70	3.54	3.38
	0.05	2.28	2.25	2.18	2.10	2.01	1.96	1.94	1.92	1.87	1.81
	0.025	2.68	2.64	2.53	2.42	2.31	2.25	2.21	2.18	2.11	2.04
21	0.01	3.24	3.17	3.03	2.88	2.72	2.64	2.58	2.55	2.46	2.36
	0.005	3.68	3.60	3.43	3.24	3.05	2.95	2.88	2.84	2.73	2.61
	0.001	4.81	4.70	4.44	4.17	3.88	3.74	3.64	3.58	3.42	3.26

T A B L A 3 (Cont.)

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 1	2	3	4	5	6	7	8	9	10
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.39	2.30
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
22	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	0.005	6.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70
	0.001	14.4	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
23	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	0.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64
	0.001	14.2	9.47	7.67	6.69	6.08	5.65	5.33	5.09	4.89	4.73
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
24	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	0.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59
	0.001	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
25	0.01	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
	0.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54
	0.001	13.9	9.22	7.45	6.49	5.88	5.46	5.15	4.91	4.71	4.56
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	0.025	5.56	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
26	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	0.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49
	0.001	13.7	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
27	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.45	3.06
	0.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45
	0.001	13.6	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 11	12	15	20	30	40	50	60	120	∞
	0.05	2.26	2.23	2.15	2.07	1.98	1.94	1.91	1.89	1.84	1.78
	0.025	2.65	2.60	2.50	2.39	2.27	2.21	2.17	2.14	2.08	2.00
22	0.01	3.18	3.12	2.98	2.83	2.67	2.58	2.53	2.50	2.40	2.31
	0.005	3.61	3.53	3.36	3.18	2.98	2.88	2.82	2.77	2.66	2.55
	0.001	4.70	4.58	4.33	4.06	3.78	3.63	3.54	3.48	3.32	3.15
	0.05	2.24	2.20	2.13	2.05	1.96	1.91	1.89	1.86	1.81	1.76
	0.025	2.62	2.57	2.47	2.36	2.24	2.18	2.14	2.11	2.04	1.97
23	0.01	3.14	3.07	2.93	2.78	2.62	2.54	2.48	2.45	2.35	2.26
	0.005	3.55	3.47	3.30	3.12	2.92	2.82	2.76	2.71	2.60	2.48
	0.001	4.59	4.48	4.23	3.96	3.68	3.53	3.44	3.38	3.22	3.05
	0.05	2.22	2.18	2.11	2.03	1.94	1.89	1.86	1.84	1.79	1.73
	0.025	2.59	2.54	2.44	2.33	2.21	2.15	2.11	2.08	2.01	1.94
24	0.01	3.09	3.03	2.89	2.74	2.58	2.49	2.44	2.40	2.31	2.21
	0.005	3.50	3.42	3.25	3.06	2.87	2.77	2.70	2.66	2.55	2.43
	0.001	4.50	4.39	4.14	3.87	3.59	3.45	3.36	3.29	3.14	2.97
	0.05	2.20	2.16	2.09	2.01	1.92	1.87	1.84	1.82	1.77	1.71
	0.025	2.56	2.51	2.41	2.30	2.18	2.12	2.08	2.05	1.98	1.91
25	0.01	3.06	2.99	2.85	2.70	2.54	2.45	2.40	2.36	2.27	2.17
	0.005	3.44	3.37	3.20	3.01	2.82	2.72	2.65	2.61	2.50	2.38
	0.001	4.42	4.31	4.06	3.97	3.52	3.37	3.28	3.22	3.06	2.89
	0.05	2.18	2.15	2.07	1.99	1.90	1.85	1.82	1.80	1.75	1.69
	0.025	2.54	2.49	2.39	2.28	2.16	2.09	2.05	2.03	1.95	1.88
26	0.01	3.02	2.96	2.81	2.66	2.50	2.42	2.36	2.33	2.23	2.13
	0.005	3.40	3.33	3.15	2.97	2.77	2.67	2.61	2.56	2.45	2.33
	0.001	4.35	4.24	3.99	3.72	3.44	3.30	3.21	3.15	2.99	2.82
	0.05	2.17	2.13	2.06	1.97	1.88	1.84	1.81	1.79	1.73	1.67
	0.025	2.51	2.47	2.36	2.25	2.13	2.07	2.03	2.00	1.93	1.85
27	0.01	2.99	2.93	2.78	2.63	2.47	2.38	2.33	2.29	2.20	2.10
	0.005	3.36	3.28	3.11	2.94	2.73	2.63	2.57	2.52	2.41	2.29
	0.001	4.28	4.17	3.92	3.66	3.38	3.23	3.14	3.08	2.92	2.75

TABLA 3 (Cont.)

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 1	2	3	4	5	6	7	8	9	10
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
28	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	0.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41
	0.001	13.5	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	0.025	5.59	4.20	3.60	3.27	3.04	2.88	2.76	2.67	2.59	2.51
29	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	0.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38
	0.001	13.4	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
30	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	0.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34
	0.001	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
40	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	0.005	8.83	6.07	4.98	4.37	3.99	3.71	3.61	3.35	3.22	3.12
	0.001	12.6	8.25	6.60	5.70	5.13	4.73	4.44	4.21	4.02	3.87
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.61	2.41	2.33	2.27
60	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	0.005	8.49	5.79	4.73	4.14	3.76	3.69	3.29	3.13	3.01	2.90
	0.001	12.0	7.76	6.17	5.31	4.76	4.37	4.09	3.87	3.69	3.54
	0.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	0.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
120	0.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	0.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71
	0.001	11.4	7.32	5.79	4.95	4.42	4.04	3.77	3.55	3.38	3.24
	0.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	0.025	5.02	3.69	3.11	2.79	2.57	2.41	2.29	2.19	2.11	2.05
∞	0.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	0.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52
	0.001	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

Límites de significación de la distribución de Snedecor (*F*).

gl del denominador	α	gl del numerador 11	12	15	20	30	40	50	60	120	∞
28	0.05	2.15	2.12	2.04	1.96	1.87	1.82	1.79	1.77	1.71	1.65
	0.025	2.49	2.45	2.34	2.23	2.11	2.05	2.01	1.98	1.91	1.83
	0.01	2.96	2.90	2.75	2.60	2.44	2.35	2.30	2.26	2.17	2.06
	0.005	3.32	3.25	3.07	2.89	2.69	2.59	2.53	2.48	2.37	2.25
	0.001	4.22	4.11	3.86	3.60	3.32	3.18	3.09	3.02	2.86	2.69
29	0.05	2.14	2.10	2.03	1.94	1.85	1.81	1.78	1.75	1.70	1.64
	0.025	2.47	2.43	2.32	2.21	2.09	2.03	1.99	1.96	1.89	1.81
	0.01	2.93	2.87	2.73	2.57	2.41	2.33	2.27	2.23	2.14	2.03
	0.005	3.29	3.21	3.04	2.86	2.66	2.56	2.49	2.45	2.33	2.21
	0.001	4.16	4.05	3.80	3.54	3.27	3.12	3.03	2.97	2.81	2.64
30	0.05	2.13	2.09	2.01	1.93	1.84	1.79	1.76	1.74	1.68	1.62
	0.025	2.46	2.41	2031	2.20	2.07	2.01	1.97	1.94	1.87	1.79
	0.01	2.90	2.84	2.70	2.55	2.39	2.30	2.25	2.21	2.11	2.01
	0.005	3.25	3.18	3.01	2.82	2.63	2.52	2.46	2.42	2.30	2.18
	0.001	4.11	4.00	3.75	3.49	3.22	3.07	2.98	2.92	2.76	2.59
40	0.05	2.04	2.04	1.92	1.84	1.74	1.69	1.66	1.64	1.58	1.51
	0.025	2.33	2.29	2.18	2.07	1.94	1.88	1.83	1.80	1.72	1.64
	0.01	2.73	2.66	2.52	2.37	2.20	3.29	3.12	2.99	2.89	2.80
	0.005	8.83	6.07	4.98	4.37	3.99	2.11	2.06	2.02	1.92	1.80
	0.001	3.74	3.64	3.40	2.15	2.87	2.73	2.64	2.57	2.41	2.23
60	0.05	1.95	1.92	1.84	1.75	1.65	1.59	1.56	1.53	1.47	1.39
	0.025	2.22	2.17	2.06	1.94	1.82	1.74	1.70	1.67	1.58	1.48
	0.01	2.56	2.50	2.35	2.20	2.03	1.94	1.88	1.84	1.73	1.60
	0.005	2.81	2.74	2.57	2.39	2.19	2.08	2.01	1.96	1.83	1.69
	0.001	3.41	3.31	3.08	2.83	2.55	2.41	2.32	2.25	2.08	1.89
120	0.05	1.87	1.83	1.75	1.56	1.55	1.50	1.46	1.43	1.35	1.25
	0.025	2.10	2.05	1.95	1.82	1.69	1.61	1.56	1.53	1.43	1.31
	0.01	2.40	2.34	2.19	2.03	1.86	1.76	1.70	1.66	1.53	1.38
	0.005	2.62	2.54	2.37	2.19	1.98	1.87	1.80	1.75	1.61	1.43
	0.001	3.11	3.02	2.78	2.53	2.26	2.11	2.02	1.95	1.76	1.54
∞	0.05	1.79	1.75	1.67	1.57	1.46	1.39	1.35	1.32	1.22	1.00
	0.025	1.99	1.94	1.83	1.71	1.57	1.48	1.43	1.39	1.27	1.00
	0.01	2.25	2.18	2.04	1.88	1.70	1.59	1.52	1.47	1.32	1.00
	0.005	2.43	2.36	2.19	2.00	1.79	1.67	1.59	1.53	1.36	1.00
	0.001	2.84	2.74	2.51	2.27	1.99	1.84	1.73	1.66	1.45	1.00

TABLA 4

Límites de significación de la distribución χ^2.

2α	0.9990	0.990	0.9750	0.950	0.90	0.80	0.70	0.60
α	0.4995	0.495	0.4875	0.475	0.45	0.40	0.35	0.30
gl 1	0.00000	0.00016	0.00098	0.00393	0.0158	0.0642	0.148	0.275
2	0.0020	0.0201	0.0506	0.103	0.211	0.446	0.713	1.022
3	0.0243	0.115	0.216	0.352	0.584	1.005	1.424	1.869
4	0.0908	0.297	0.484	0.711	1.064	1.649	2.195	2.743
5	0.210	0.554	0.831	1.145	1.610	2.343	3.000	3.655
6	0.381	0.872	1.237	1.635	2.204	3.070	3.828	4.570
7	0.598	1.239	1.690	2.167	2.833	3.822	4.671	5.493
8	0.857	1.646	2.180	2.733	3.490	4.594	5.527	6.423
9	1.153	2.088	2.700	3.325	4.168	5.380	6.393	7.357
10	1.479	2.558	3.247	3.940	4.865	6.179	7.267	8.295
11	1.834	3.053	3.816	4.575	5.578	6.989	8.148	9.237
12	2.214	3.571	4.404	5.226	6.304	7.807	9.034	10.182
13	2.617	4.107	5.009	5.892	7.042	8.634	9.926	11.129
14	3.041	4.660	5.629	6.571	7.790	9.467	10.821	12.079
15	3.483	5.229	6.262	7.261	8.547	10.307	11.721	13.030
16	3.942	5.812	6.908	7.962	9.312	11.152	12.624	13.983
17	4.416	6.408	7.564	8.672	10.085	12.002	13.531	14.937
18	4.905	7.015	8.231	9.390	10.865	12.857	14.440	15.893
19	5.407	7.633	8.907	10.117	11.651	13.716	15.352	16.850
20	5.921	8.260	9.591	10.851	12.443	14.578	16.266	17.809
21	6.447	8.897	10.283	11.591	13.240	15.445	17.182	18.768
22	6.983	9.542	10.982	12.338	14.041	16.314	18.101	19.729
23	7.529	10.196	11.688	13.091	14.848	17.187	19.021	20.690
24	8.085	10.856	12.401	13.848	15.659	18.062	19.943	21.652
25	8.649	11.524	13.120	14.611	16.473	18.940	20.867	22.616
26	9.222	12.198	13.884	15.379	17.292	19.820	21.792	23.579
27	9.803	12.879	14.573	16.151	18.114	20.703	22.719	24.544
28	10.391	13.565	15.308	16.928	18.939	21.588	23.647	25.509
29	10.986	14.256	16.047	17.708	19.768	22.475	24.577	26.475
30	11.588	14.953	16.791	18.493	20.599	23.364	25.508	27.442
31	12.196	15.655	17.539	19.281	21.434	24.255	26.440	28.409
32	12.811	16.362	18.291	20.072	22.271	25.148	27.373	29.376
33	13.431	17.073	19.047	20.867	23.110	26.042	28.307	30.344
34	14.057	17.789	19.806	21.664	23.952	26.938	29.242	31.313
35	14.688	18.509	20.569	22.465	24.797	27.836	30.178	32.282
36	15.324	19.233	21.336	23.269	25.643	28.735	31.115	33.252
37	15.965	19.960	22.106	24.075	26.492	29.635	32.053	34.222
38	16.611	20.691	22.878	24.884	27.343	30.538	32.992	35.192
39	17.261	21.426	23.654	25.695	28.196	31.441	33.932	36.163
40	17.916	22.164	24.433	26.509	29.041	32.345	34.872	37.134
50	24.674	29.707	32.357	34.764	37.689	41.449	44.313	46.864
60	31.739	37.485	40.482	43.188	46.459	50.641	53.809	56.620
70	39.036	45.442	48.758	51.739	55.329	59.989	63.346	66.396

Límites de significación de la distribución χ^2.

2α	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.001
α	0.20	0.15	0.10	0.05	0.025	0.0125	0.005	0.00051
gl 1	0.708	1.074	1.642	2.706	3.841	5.024	6.635	10.828
2	1.833	2.408	3.219	4.605	5.991	7.378	9.210	13.816
3	2.946	3.665	4.642	6.251	7.815	9.348	11.345	16.266
4	4.045	4.878	5.989	7.779	9.488	11.143	13.277	18.467
5	5.132	6.064	7.289	9.236	11.070	12.832	15.086	20.515
6	6.211	7.231	8.558	10.645	12.592	14.449	16.812	22.458
7	7.283	8.383	9.803	12.017	14.067	16.013	18.475	24.322
8	8.351	9.524	11.030	13.362	15.507	17.535	20.090	26.125
9	9.414	10.656	12.242	14.684	16.919	19.023	21.666	27.877
10	10.473	11.781	13.442	15.987	18.307	20.483	23.209	29.588
11	11.530	12.899	14.631	17.275	19.675	21.920	24.725	31.264
12	12.584	14.011	15.812	18.549	21.026	23.336	26.217	32.909
13	13.636	15.119	16.985	19.812	22.362	24.736	27.688	34.528
14	14.685	16.222	18.151	21.064	23.685	26.119	29.141	36.123
15	15.733	17.322	19.311	22.307	24.996	27.488	30.578	37.697
16	16.780	18.418	20.465	23.542	26.296	28.845	32.000	39.252
17	17.824	19.511	21.615	24.769	27.587	30.191	33.409	40.790
18	18.868	20.601	22.760	25.989	28.869	31.526	34.805	42.312
19	19.910	21.689	23.900	27.204	30.144	32.852	36.191	43.820
20	20.951	22.775	25.038	28.412	31.410	34.170	37.566	45.315
21	21.991	23.858	26.171	29.615	32.671	35.479	38.932	46.797
22	23.031	24.939	27.301	30.813	33.924	36.781	40.289	48.268
23	24.069	26.018	28.429	32.007	35.172	38.076	41.638	49.728
24	25.106	27.096	29.553	33.196	36.415	39.364	42.980	51.179
25	26.143	28.172	30.675	34.382	37.652	40.646	44.314	52.620
26	27.179	29.246	31.795	35.563	38.885	41.923	45.642	54.052
27	28.214	30.319	32.912	36.741	40.113	43.194	46.963	55.476
28	29.249	31.391	34.027	37.916	41.337	44.461	48.278	56.892
29	30.283	32.461	35.139	39.087	42.557	45.722	49.588	58.302
30	31.316	33.530	36.250	40.256	43.773	46.979	50.892	59.703
31	32.349	34.598	37.359	41.422	44.985	48.232	52.191	61.098
32	33.381	35.665	38.466	42.585	46.194	49.480	53.486	62.487
33	34.413	36.731	39.572	43.745	47.400	50.725	54.776	63.870
34	35.444	37.795	40.676	44.903	48.602	51.966	56.061	65.247
35	36.475	38.859	41.779	46.059	49.802	53.203	57.342	66.619
36	37.505	39.922	42.879	47.212	50.998	54.437	58.619	67.985
37	38.535	40.984	43.978	48.363	52.192	55.668	59.892	69.346
38	39.564	42.045	45.076	49.513	53.384	56.895	61.162	70.703
39	40.593	43.105	46.173	50.660	54.572	58.120	62.428	72.055
40	41.622	44.165	47.269	51.805	55.758	59.342	63.691	73.402
50	51.892	54.723	58.164	63.167	67.505	71.420	76.154	86.661
60	62.135	65.226	68.972	74.397	79.082	83.298	88.379	99.952
70	72.358	75.689	79.715	85.527	90.531	95.023	100.425	112.317

TABLA 5

Factores de tolerancia de las distribuciones normales.

	$\gamma = 0.75$					$\gamma = 0.90$				
N \ P	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
2	4.498	6.301	7.414	9.531	11.920	11.407	15.978	18.800	24.167	30.227
3	2.501	3.538	4.187	5.431	6.844	4.132	5.847	6.919	8.974	11.309
4	2.035	2.892	3.431	4.471	5.657	2.932	4.166	4.943	6.440	8.149
5	1.825	2.599	3.088	4.033	5.117	2.454	3.494	3.152	5.423	6.879
6	1.704	2.429	2.889	3.779	4.802	2.196	3.131	3.723	4.870	6.188
7	1.624	2.318	2.757	3.611	4.593	2.034	2.902	3.452	4.521	5.750
8	1.568	2.238	2.663	3.491	4.444	1.921	2.743	2.264	4.278	5.446
9	1.525	2.178	2.593	3.400	4.330	1.839	2.626	2.125	4.098	5.220
10	1.492	2.131	2.537	3.328	4.241	1.775	2.535	3.018	3.959	5.046
11	1.465	2.093	2.493	3.271	4.169	1.724	2.463	2.933	3.849	4.906
12	1.443	2.062	2.456	3.223	4.110	1.683	2.404	2.863	3.758	4.792
13	1.425	2.036	2.424	3.183	4.059	1.648	2.355	2.805	3.682	4.697
14	1.409	2.013	2.398	3.148	4.016	1.619	2.314	2.756	3.618	4.615
15	1.395	1.994	2.375	3.118	3.979	1.594	2.278	2.713	3.562	4.545
16	1.383	1.977	2.355	3.092	3.946	1.572	2.246	2.676	3.514	4.484
17	1.372	1.962	2.337	3.069	3.917	1.552	2.219	2.643	3.471	4.430
18	1.363	1.948	2.321	3.048	3.891	1.535	2.194	2.614	3.433	4.382
19	1.355	1.936	2.307	3.030	3.867	1.520	2.172	2.588	3.399	4.339
20	1.347	1.925	2.294	3.013	3.846	1.506	2.152	2.564	3.368	4.300
21	1.340	1.915	2.282	2.998	3.827	1.493	2.135	2.543	3.340	4.264
22	1.334	1.906	2.271	2.984	3.809	1.482	2.118	2.524	3.315	4.232
23	1.328	1.898	2.261	2.971	3.793	1.471	2.103	2.506	3.292	4.203
24	1.322	1.891	2.252	2.959	3.778	1.462	2.089	2.489	3.270	4.176
25	1.317	1.883	2.244	2.948	3.764	1.453	2.077	2.474	3.251	4.151
26	1.313	1.877	2.236	2.938	3.751	1.444	2.065	2.460	3.232	4.127
27	1.309	1.871	2.229	2.929	3.740	1.437	2.054	2.447	3.215	4.106
30	1.297	1.855	2.210	2.904	3.708	1.417	2.025	2.413	3.170	4.049
40	1.271	1.818	2.166	2.846	3.635	1.370	1.959	2.334	3.066	3.917
50	1.255	1.794	2.138	2.809	3.588	1.340	1.916	2.284	3.001	3.833
60	1.243	1.778	2.118	2.784	3.556	1.320	1.887	2.248	2.955	3.774
70	1.235	1.765	2.104	2.764	3.531	1.304	1.865	2.222	2.920	3.730
80	1.228	1.756	2.092	2.749	3.512	1.292	1.848	2.202	2.894	3.696
90	1.223	1.748	2.083	2.737	3.497	1.283	1.834	2.185	2.872	3.669
100	1.218	1.742	2.075	2.727	3.484	1.275	1.822	2.172	2.854	3.646
200	1.195	1.709	2.037	2.677	3.419	1.234	1.764	2.102	2.762	3.529
300	1.186	1.696	2.021	2.656	3.393	1.217	1.740	2.073	2.725	3.481
400	1.181	1.688	2.012	2.644	3.378	1.207	1.726	2.057	2.703	3.453
500	1.177	1.683	2.006	2.636	3.368	1.201	1.717	2.046	2.689	3.434
600	1.175	1.680	2.002	2.631	3.360	1.196	1.710	2.038	2.678	3.421
700	1.173	1.677	1.998	2.626	3.355	1.192	1.705	2.032	2.670	3.411
800	1.171	1.675	1.996	2.623	3.350	1.189	1.701	2.027	2.663	3.402
900	1.170	1.673	1.993	2.620	3.347	1.187	1.697	2.023	2.658	3.396
∞	1.150	1.645	1.960	2.576	3.291	1.150	1.645	1.960	2.576	3.291

TABLA 5 (Cont.)

Factores de tolerancia de las distribuciones normales.

N \ P	γ = 0.95					γ = 0.99				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
2	22.858	32.019	37.674	48.430	60.573	114.36	160.19	188.49	242.30	303.05
3	5.922	8.380	9.916	12.861	16.208	13.378	18.930	22.401	29.055	36.616
4	3.779	5.369	6.370	8.299	10.502	6.614	9.398	11.150	14.527	18.383
5	3.002	4.275	5.079	6.634	8.415	4.643	6.612	9.855	10.260	13.015
6	2.604	3.712	4.414	5.775	7.337	3.743	5.337	6.345	8.301	10.548
7	2.361	3.369	4.007	5.248	6.676	3.233	4.613	5.488	7.187	9.142
8	2.197	3.136	3.732	4.891	6.226	2.905	4.147	4.936	6.468	8.234
9	2.078	2.967	3.532	4.631	5.899	2.677	3.822	4.550	5.966	7.600
10	1.987	2.839	3.379	4.433	5.649	2.508	3.582	4.265	5.594	7.129
11	1.916	2.737	3.259	4.277	5.452	2.378	3.397	4.045	5.308	6.766
12	1.858	2.655	3.162	4.150	5.291	2.274	3.250	3.870	5.079	6.477
13	1.810	2.587	3.081	4.044	5.158	2.190	3.130	3.727	4.893	6.240
14	1.770	2.529	3.012	3.955	5.045	2.120	3.029	3.608	4.737	6.043
15	1.735	2.480	2.954	3.878	4.949	2.060	2.945	3.507	4.605	5.876
16	1.705	2.437	2.903	3.812	4.865	2.009	2.872	3.421	4.492	5.732
17	1.679	2.400	2.858	3.754	4.791	1.965	2.808	3.345	4.393	5.607
18	1.655	2.366	2.819	3.702	4.725	1.926	2.753	3.279	4.307	5.497
19	1.635	2.337	2.784	3.656	4.667	1.891	2.703	3.221	4.230	5.399
20	1.616	2.310	2.752	3.615	4.614	1.860	2.659	3.168	4.161	5.312
21	1.599	2.286	2.723	3.577	4.567	1.833	2.620	3.121	4.100	5.234
22	1.584	2.264	2.697	3.543	4.523	1.808	2.584	3.078	4.044	5.163
23	1.570	2.244	2.673	3.512	4.484	1.785	2.551	3.040	3.993	5.098
24	1.557	2.225	2.651	3.483	4.447	1.764	2.522	3.004	3.947	5.039
25	1.545	2.208	2.631	3.457	4.413	1.745	2.494	2.972	3.904	4.985
26	1.534	2.193	2.612	3.432	4.382	1.727	2.469	2.941	3.865	4.935
27	1.523	2.178	2.595	3.409	4.353	1.711	2.446	2.914	3.828	4.888
30	1.497	2.140	2.549	3.350	4.278	1.668	2.385	2.841	3.733	4.768
40	1.435	2.052	2.445	3.213	4.104	1.571	2.247	2.677	3.518	4.493
50	1.396	1.996	2.379	3.126	3.993	1.512	2.162	2.576	3.385	4.323
60	1.369	1.958	2.333	3.066	3.916	1.471	2.103	2.506	3.293	4.206
70	1.349	1.929	2.299	3.021	3.859	1.440	2.060	2.454	3.225	4.120
80	1.334	1.907	2.272	2.986	3.814	1.417	2.026	2.414	3.173	4.053
90	1.321	1.889	2.251	2.958	3.778	1.398	1.999	2.382	3.130	3.999
100	1.311	1.874	2.233	2.934	3.748	1.383	1.977	2.355	3.096	3.954
200	1.258	1.798	2.143	2.816	3.597	1.304	1.865	2.222	2.921	3.731
300	1.236	1.767	2.106	2.767	3.535	1.273	1.820	2.169	2.850	3.641
400	1.223	1.749	2.084	2.739	3.499	1.255	1.794	2.138	2.809	3.589
500	1.215	1.737	2.070	2.721	3.475	1.243	1.777	2.117	2.783	3.555
600	1.209	1.729	2.060	2.707	3.458	1.234	1.764	2.102	2.763	3.530
700	1.204	1.722	2.052	2.697	3.445	1.227	1.755	2.091	2.748	3.511
800	1.201	1.717	2.046	2.688	3.434	1.222	1.747	2.082	2.736	3.495
900	1.198	1.712	2.040	2.682	3.426	1.218	1.741	2.075	2.726	3.483
∞	1.150	1.645	1.960	2.576	3.291	1.150	1.645	1.960	2.576	3.291

TABLA 6

Coeficientes para polinomios ortogonales.

n	3		4			5				6			
	c_1	c_2	c_1	c_2	c_3	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
	-1	1	-3	1	-1	-2	2	-1	1	-5	5	-5	1
	0	-2	-1	-1	3	-1	-1	2	-4	-3	-1	7	-3
	1	1	1	-1	-3	0	-2	0	6	-1	-4	4	2
			3	1	1	1	-1	-2	-4	1	-4	-4	2
						2	2	1	1	3	-1	-7	-3
										5	5	5	1
Divisores													
	2	6	20	4	20	10	14	10	70	70	84	180	28
K_1		1/3			5/16				1/7				5/96
K_2		1/2			1/20				1/10				1/70
K_3					41/240				17/60				101/4320
K_4		1/2			1/16				1/14				1/224
K_5					1/48				1/12				1/864
K_6									1/24				1/768
K_7									31/168				95/2688
K_8									3/35				27/256

n	7				8				9			
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
	-3	5	-1	3	-7	7	-7	7	-4	28	-14	14
	-2	0	1	-7	-5	1	5	-13	-3	7	7	-21
	-1	-3	1	1	-3	-3	7	-3	-2	-8	13	-11
	0	-4	0	6	-1	-5	3	9	-1	-17	9	9
	1	-3	-1	1	1	-5	-3	9	0	-20	0	18
	2	0	-1	-7	3	-3	-7	-3	1	-17	-9	9
	3	5	1	3	5	1	-5	-13	2	-8	-13	-11
					7	7	7	7	3	7	-7	-21
									4	28	14	14
Divisores												
	28	84	6	154	68	168	264	616	60	2772	990	2002
K_1			1/21				1/32				5/693	
K_2			1/28				1/168				1/60	
K_3			7/36				37/3168				59/5940	
K_4			1/84				1/672				1/924	
K_5			1/36				1/3168				1/1188	
K_6			1/264				1/16896				1/3432	
K_7			67/1848				179/59136				115/24024	
K_8			3/77				9/512				9/1001	

Coeficientes para polinomios ortogonales.

n		10		11				12			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
-9	6	-42	18	-5	15	-30	6	-11	55	-33	33
-7	2	14	-22	-4	6	6	-6	-9	25	3	-27
-5	-1	35	-17	-3	-1	22	-6	-7	1	21	-33
-3	-3	31	3	-2	-6	23	-1	-5	-17	25	-13
-1	-4	12	18	-1	-9	14	4	-3	-29	19	12
1	-4	-12	18	0	-10	0	6	-1	-35	7	28
3	-3	-31	3	1	-9	-14	4	1	-35	-7	28
5	-1	-35	-17	2	-6	-23	-1	3	-29	-19	12
7	2	-14	-22	3	-1	-22	-6	5	-17	-25	-13
9	6	42	18	4	6	-6	-6	7	1	-21	-33
				5	15	30	6	9	25	-3	-27
								11	55	33	33
Divisores											
330	132	8580	2860	110	858	4290	286	572	12012	5148	8008
K_1		1/32									
K_2		1/330				5/429				1/336	
K_3		293/205920				1/110				1/572	
K_4		1/1056				89/25740				85/61776	
K_5		1/41184				1/858				1/16016	
K_6		1/109824				1/5148				1/61776	
K_7		41/54912				1/3432				1/439296	
K_8		9/1280				25/3432				419/1537536	
						3/143				27/7168	

TABLA 6 (Cont.)

Coeficientes para polinomios ortogonales.

n	13			14			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
-6	22	-11	99	-13	13	-143	143
-5	11	0	-66	-11	7	-11	-77
-4	2	6	-96	-9	2	66	-132
-3	-5	8	-54	-7	-2	98	-92
-2	-10	7	11	-5	-5	95	-13
-1	-13	4	64	-3	-7	63	63
0	-14	0	84	-1	-8	24	108
1	-13	-4	64	1	-8	-24	108
2	-10	-7	11	3	-7	-67	63
3	-5	-8	-54	5	-5	-95	-13
4	2	-6	-96	7	-2	-98	-92
5	11	0	-66	9	2	-66	-132
6	22	11	99	11	7	11	-77
				13	13	143	143
Divisores							
182	2002	572	68068	910	728	97240	136136
K_1			1/143				5/448
K_2			1/182				1/910
K_3			25/3432				581/2333760
K_4			1/2002				1/5824
K_5			1/3432				1/466752
K_6			1/16688				1/3734016
K_7			19/62832				575/13069056
K_8			3/2431				3/3584

Coeficientes para polinomios ortogonales.

n		15		16			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
-7	91	-91	1001	-15	35	-455	273
-6	52	-13	-429	-13	21	-91	-91
-5	19	35	-869	-11	9	143	-221
-4	-8	58	-704	-9	-1	267	-201
-3	-29	61	-249	-7	-9	301	-101
-2	-44	49	251	-5	-15	265	23
-1	-53	27	621	-3	-19	179	129
0	-56	0	756	-1	-21	63	189
1	-53	-27	621	1	-21	-63	189
2	-44	-49	251	3	-19	-179	129
3	-29	-61	-249	5	-15	-265	23
4	-8	-58	-704	7	-9	-301	-101
5	19	-35	-869	9	-1	-267	-201
6	52	13	-429	11	9	-143	-221
7	91	91	1001	13	21	91	-91
				15	35	455	273
Divisores							
280	37128	39780	6466460	1360	5712	1007760	470288
K_1		1/663					
K_2		1/280				5/1344	
K_3		167/238680				1/1360	
K_4		1/12376				761/12093120	
K_5		1/47736				1/22848	
K_6		1/2217072				1/2418624	
K_7		331/1551950				1/12899328	
K_8		27/230945				755/45147648	
						3/7168	

TABLA 7

Amplitudes studentizadas (r_p).

gl		*P = número de medias para la amplitud a probar*									
error	α	2	3	4	5	6	7	8	9	10	11
1	0.05	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0
	0.01	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
2	0.05	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09
	0.01	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
3	0.05	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
	0.01	8.26	8.5	8.6	8.7	8.8	8.9	8.9	9.0	9.0	9.0
4	0.05	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
	0.01	6.51	6.8	6.9	7.0	7.1	7.1	7.2	7.2	7.3	7.3
5	0.05	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83
	0.01	5.70	5.96	6.11	6.18	6.26	6.33	6.40	6.44	6.5	6.6
6	0.05	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68
	0.01	5.24	5.51	5.65	5.73	5.81	5.88	5.95	6.00	6.00	6.1
7	0.05	3.35	3.47	3.54	3.58	3.60	3.61	3.61	3.61	3.61	3.61
	0.01	4.95	5.22	5.37	5.45	5.53	5.61	5.69	5.73	5.80	5.80
8	0.05	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56
	0.01	4.74	5.00	5.14	5.23	5.32	5.40	5.47	5.51	5.5	5.6
9	0.05	3.20	3.34	3.41	3.47	3.50	3.52	3.52	3.52	3.52	3.52
	0.01	4.60	4.86	4.99	5.08	5.17	5.25	5.32	5.36	5.4	5.5
10	0.05	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.47	3.47
	0.01	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.36
11	0.05	3.11	3.27	3.35	3.39	3.43	3.44	3.45	3.46	3.46	3.46
	0.01	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.24
12	0.05	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.46
	0.01	4.32	4.55	4.68	4.76	4.84	4.92	4.96	5.02	5.07	5.13
13	0.05	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.45
	0.01	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.04
14	0.05	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.45
	0.01	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.37	4.91	4.96
15	0.05	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.44
	0.01	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	4.90

Amplitudes studentizadas (r_p).

gl		*P = número de medias para la amplitud a probar*									
error	α	2	3	4	5	6	7	8	9	10	11
16	0.05	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.44
	0.01	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84
17	0.05	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.44
	0.01	4.10	4.30	4.41	4.40	4.56	4.63	4.68	4.72	4.75	4.80
18	0.05	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.43
	0.01	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.76
19	0.05	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.43
	0.01	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72
20	0.05	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.43
	0.01	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69
24	0.05	2.92	3.07	3.15	3.22	3.28	3.31	3.34	3.37	3.38	3.41
	0.01	3.96	4.14	4.24	4.33	4.39	4.44	4.49	4.53	4.57	4.62
30	0.05	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.40
	0.01	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.54
40	0.05	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.39
	0.01	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.46
60	0.05	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.37
	0.01	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.39
100	0.05	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.36
	0.01	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.35
∞	0.05	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.26	3.29	3.34
	0.01	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.26

TABLA 8

Puntos porcentuales superiores de la amplitud studentizada (q_p).

gl		*P = número de medias para la amplitud a probar*									
error	α	2	3	4	5	6	7	8	9	10	11
5	0.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	0.01	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	0.05	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65
	0.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	0.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	0.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	0.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	0.01	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03
9	0.05	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87
	0.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65
10	0.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	0.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	0.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	0.01	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.69	7.13
12	0.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51
	0.01	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	0.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	0.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	0.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	0.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	0.05	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.80	5.20	5.31
	0.01	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55

Puntos porcentuales superiores de la amplitud studentizada (q_p).

gl		*P = número de medias para la amplitud a probar*									
error	α	2	3	4	5	6	7	8	9	10	11
16	0.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	0.01	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	0.05	2.98	3.63	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21
	0.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	0.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	0.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	0.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	0.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	0.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	0.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	0.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	0.01	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	0.05	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92
	0.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	0.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82
	0.01	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69
60	0.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	0.01	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53
100	0.05	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64
	0.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38
∞	0.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	0.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

TABLA 9

Coeficientes a_{n-i+1} de la prueba W (w).

i \ N	1	2	3	4	5	6	7	8	9	10
1	0.0000	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2			0.0000	0.1667	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3					0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4							0.0000	0.0561	0.0947	0.1224
5									0.0000	0.0399

i \ N	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

i \ N	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12			0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13					0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14							0.0000	0.0084	0.0159	0.0227
15									0.0000	0.0076

Prueba W para desviaciones de la Normalidad.

N \ α	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.919	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991

TABLA 11

Valores críticos de *r* para la prueba de rachas (*r*).

	$n_1 = 2$							
n_2	U0.005	U0.01	U0.025	U0.05	U0.95	U0.975	U0.99	U0.995
2					4	4	4	4
3					5	5	5	5
4					5	5	5	5
5					5	5	5	5
6					5	5	5	5
7					5	5	5	5
8				2	3	5	5	5
9				2	5	5	5	5
10				2	5	5	5	5
11				2	5	5	5	5
12			2	2	5	5	5	5
13			2	2	5	5	5	5
14			2	2	5	5	5	5
15			2	2	5	5	5	5
16			2	2	5	5	5	5
17			2	2	5	5	5	5
18			2	2	5	5	5	5
19		2	2	2	5	5	5	5
20		2	2	2	5	5	5	5

	$n_1 = 3$							
n_2	U0.005	U0.01	U0.025	U0.05	U0.95	U0.975	U0.99	U0.995
3					6	6	6	6
4					6	7	7	7
5				2	7	7	7	7
6			2	2	7	7	7	7
7			2	2	7	7	7	7
8			2	2	7	7	7	7
9		2	2	2	7	7	7	7
10		2	2	3	7	7	7	7
11		2	2	3	7	7	7	7
12	2	2	2	3	7	7	7	7
13	2	2	2	3	7	7	7	7
14	2	2	2	3	7	7	7	7
15	2	2	3	3	7	7	7	7
16	2	2	3	3	7	7	7	7
17	2	2	3	3	7	7	7	7
18	2	2	3	3	7	7	7	7
19	2	2	3	3	7	7	7	7
20	2	2	3	3	7	7	7	7

Valores críticos de *r* para la prueba de rachas (*r*).

$n_1 = 4$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
4				2	7	8	8	8
5			2	2	8	8	8	9
6		2	2	3	8	8	9	9
7		2	2	3	8	8	9	9
8	2	2	3	3	9	9	9	9
9	2	2	3	3	9	9	9	9
10	2	2	3	3	9	9	9	9
11	2	2	3	3	9	9	9	9
12	2	3	3	4	9	9	9	9
13	2	3	3	4	9	9	9	9
14	2	3	3	4	9	9	9	9
15	2	3	3	4	9	9	9	9
16	2	2	4	4	9	9	9	9
17	2	3	4	4	9	9	9	9
18	2	3	4	4	9	9	9	9
19	2	3	4	4	9	9	9	9
20	2	3	4	4	9	9	9	9

$n_1 = 5$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
5		2	2	3	8	9	9	10
6	2	2	3	3	9	9	10	10
7	2	2	3	3	9	10	10	11
8	2	2	3	3	10	10	11	11
9	2	3	3	4	10	11	11	11
10	3	3	3	4	10	11	11	11
11	3	3	4	4	11	11	11	11
12	3	3	4	4	11	11	11	11
13	3	3	4	4	11	11	11	11
14	3	3	4	5	11	11	11	11
15	3	4	4	5	11	11	11	11
16	3	4	4	5	11	11	11	11
17	3	4	4	5	11	11	11	11
18	4	4	5	5	11	11	11	11
19	4	4	5	5	11	11	11	11
20	4	4	5	5	11	11	11	11

TABLA 11 (Cont.)

Valores críticos de *r* para la prueba de rachas (*r*).

$n_1 = 6$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
6	2	2	3	3	10	10	11	11
7	2	3	3	4	10	11	11	12
8	3	3	3	4	11	11	12	12
9	3	3	4	4	11	12	12	13
10	3	3	4	5	11	12	13	13
11	3	4	4	5	12	12	13	13
12	3	4	4	5	12	12	13	13
13	3	4	5	5	12	13	13	13
14	4	4	5	5	12	13	13	13
15	4	4	5	6	13	13	13	13
16	4	4	5	6	13	13	13	13
17	4	5	5	6	13	13	13	13
18	4	5	5	6	13	13	13	13
19	4	5	6	6	13	13	13	13
20	4	5	6	6	13	13	13	13
$n_1 = 7$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
7	3	3	3	4	11	12	12	12
8	3	3	4	4	12	12	13	13
9	3	4	4	5	12	13	13	14
10	3	4	5	5	12	13	14	14
11	4	4	5	5	13	13	14	14
12	4	4	5	6	13	13	14	15
13	4	5	5	6	13	14	15	15
14	4	5	5	6	13	14	15	15
15	4	5	6	6	14	14	15	15
16	5	5	6	6	14	15	15	15
17	5	5	6	7	14	15	15	15
18	5	5	6	7	14	15	15	15
19	5	6	6	7	14	15	15	15
20	5	6	6	7	14	15	15	15
$n_1 = 8$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
8	3	4	4	5	12	13	13	14
9	3	4	5	5	13	13	14	14
10	4	4	5	6	13	14	14	15
11	4	5	5	6	14	14	15	15
12	4	5	6	6	14	15	15	16
13	5	5	6	6	14	15	16	16
14	5	5	6	7	15	15	16	16
15	5	5	6	7	15	15	16	17
16	5	6	6	7	15	16	16	17
17	5	6	7	7	15	16	17	17
18	6	6	7	8	15	16	17	17
19	6	6	7	8	15	16	17	17
20	6	6	7	8	15	16	17	17

Valores críticos de *r* para la prueba de rachas (*r*).

	$n_1 = 9$							
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
9	4	4	5	6	13	14	15	15
10	4	5	5	6	14	15	15	16
11	5	5	6	6	14	15	16	16
12	5	5	6	7	15	15	16	17
13	5	6	6	7	15	16	17	17
14	5	6	7	7	16	16	17	17
15	6	6	7	8	16	17	17	18
16	6	6	7	8	16	17	17	18
17	6	7	7	8	16	17	18	18
18	6	7	8	8	17	17	18	19
19	6	7	8	8	17	17	18	19
20	7	7	8	9	17	17	18	19

	$n_1 = 10$							
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
10	5	5	6	6	15	15	16	16
11	5	5	6	7	15	16	17	17
12	5	6	7	7	16	16	17	18
13	5	6	7	8	16	17	18	18
14	6	6	7	8	16	17	18	18
15	6	7	7	8	17	17	18	19
16	6	7	8	8	17	18	19	19
17	7	7	8	9	17	18	19	19
18	7	7	8	9	18	18	19	20
19	7	8	8	9	18	19	19	20
20	7	8	9	9	18	19	19	20

	$n_1 = 11$							
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
11	5	6	7	7	16	116	17	18
12	6	6	7	8	16	17	18	18
13	6	6	7	8	17	18	18	19
14	6	7	8	8	17	18	19	19
15	7	7	8	9	18	18	19	20
16	7	7	8	9	18	19	20	20
17	7	8	9	9	18	19	20	21
18	7	8	9	10	19	19	20	21
19	8	8	9	10	19	20	21	21
20	8	8	9	10	19	20	21	21

TABLA 11 (Cont.)

Valores críticos de *r* para la prueba de rachas (*r*).

$n_1 = 12$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
12	6	7	7	8	17	18	18	19
13	6	7	8	9	17	18	19	20
14	7	7	8	9	18	19	20	20
15	7	8	8	9	18	19	20	21
16	7	8	9	10	19	20	21	21
17	8	8	9	10	19	20	21	21
18	8	8	9	10	20	20	21	22
19	8	9	10	10	20	21	22	22
20	8	9	10	11	20	21	22	22

$n_1 = 13$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
13	7	7	8	9	18	19	20	20
14	7	8	9	9	19	19	20	21
15	7	8	9	10	19	20	21	21
16	8	8	9	10	20	20	21	22
17	8	9	10	10	20	21	22	22
18	8	9	10	11	20	21	22	23
19	9	9	10	11	21	22	23	23
20	9	10	10	11	21	22	23	23

$n_1 = 14$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
14	7	8	9	10	19	20	21	22
15	8	8	9	10	20	21	22	22
16	8	9	10	11	20	21	22	23
17	8	9	10	11	21	22	23	23
18	9	9	10	11	21	22	23	24
19	9	10	11	12	22	22	23	24
20	9	10	11	12	22	22	24	24

$n_1 = 15$								
n_2	$U0.005$	$U0.01$	$U0.025$	$U0.05$	$U0.95$	$U0.975$	$U0.99$	$U0.995$
15	8	9	10	11	20	21	22	23
16	9	9	10	11	21	22	23	23
17	9	10	11	11	21	22	23	24
18	9	10	11	12	22	23	24	24
19	10	10	11	12	22	23	24	25
20	10	11	12	12	22	24	25	25