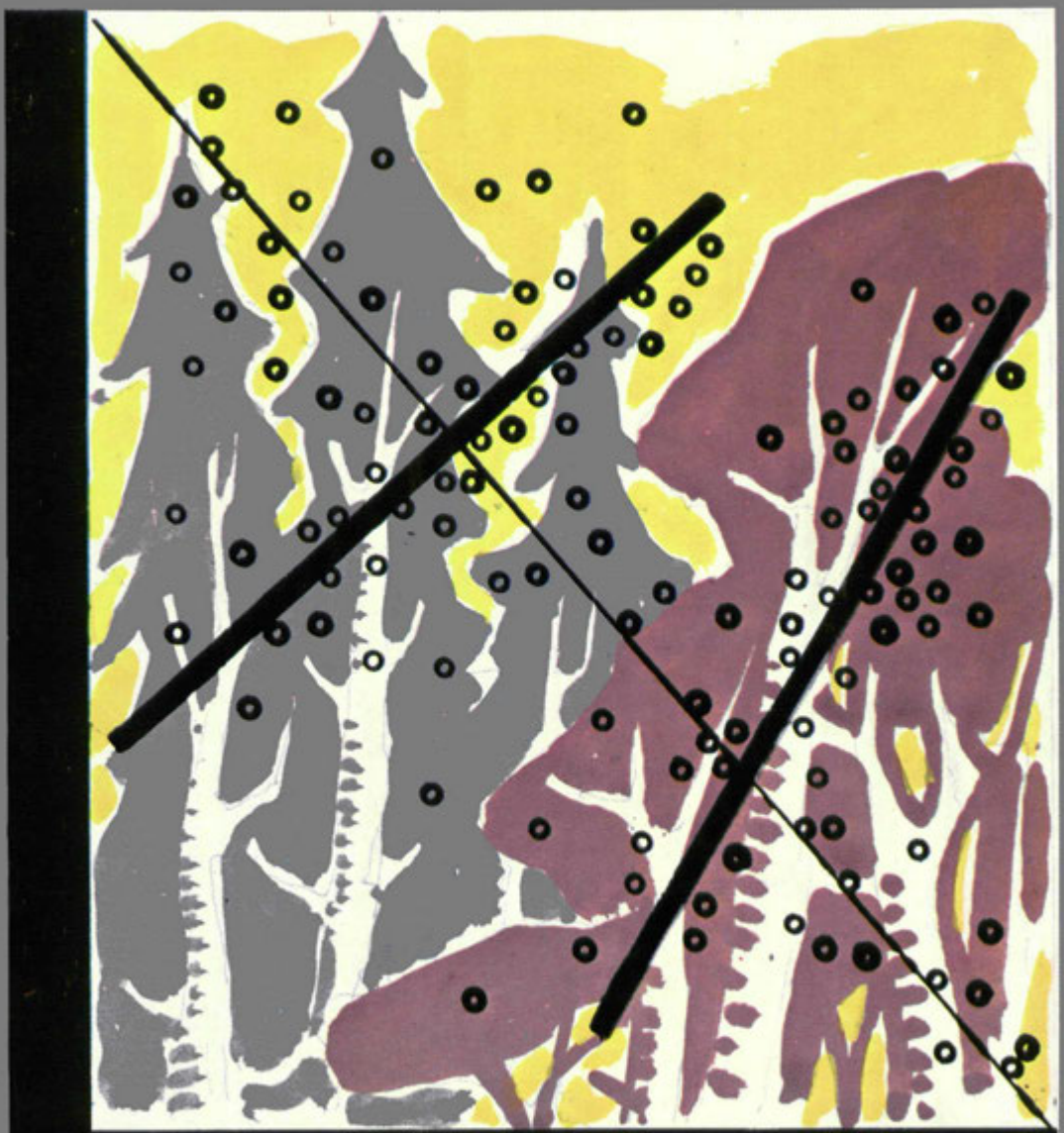
REGRESION LINEAL

UN ENFOQUE CONCEPTUAL Y PRACTICO



E. CARBONELL, J. B. DENIS, R. CALVO, F. GONZALEZ y V. PRUÑONOSA

INSTITUTO NACIONAL DE INVESTIGACIONES AGRARIAS

MINISTERIO DE AGRICULTURA, PESCA Y ALIMENTACION

Colección: MONOGRAFIAS I.N.I.A., N.º 43

MINISTERIO DE AGRICULTURA, PESCA Y ALIMENTACION
INSTITUTO NACIONAL DE INVESTIGACIONES AGRARIAS

REGRESION LINEAL

UN ENFOQUE CONCEPTUAL Y PRACTICO

E. Carbonell
J. B. Denis
R. Calvo
F. González
V. Pruñonosa

Sección de Proceso de Datos del INIA
Carretera de La Coruña km 7. Apartado 8.111
MADRID - 35

Instituto Nacional de Investigaciones Agrarias
José Abascal, 56. Tlfo.: 441.31.93. Telex 48989 INIA E
Madrid - 3 (España)

MADRID – 1983

Depósito Legal: M - 22386 - 1983
I. S. B. N.: 84 - 7498 - 152 - 2 — I. S. S. N.: 0210 - 3354
Imprime: Gráf. Maravillas, S. L. — Madrid - 29

INDICE GENERAL

Páginas

PROLOGO

Es una opinión muy extendida que el Análisis de Regresión es la técnica estadística más utilizada en la experimentación agraria y frecuentemente de forma incorrecta. Desde hace tiempo se dispone de algún buen tratado sobre regresión como el de Draper y, recientemente, se han publicado varias monografías como las de Daniel, Chatterjee o Mosteller y Tukey, entre otras, pero que suelen estar orientados a aplicaciones industriales desde una óptica anglosajona. Cualquier aportación a la bibliografía estadística en castellano, que trate de profundizar y aclarar las cuestiones que se presentan a nuestros investigadores en la aplicación de esta técnica, debe ser, en principio, bienvenida. (Igualmente lo sería, si bien por razones distintas, el desarrollo y publicación en España de «software» de análisis estadístico.)

Podría, no obstante, atribuirse redundancia a algunas partes del texto, pero es importante resaltar que su orientación y desarrollo general me parecen muy acertados para que sea de gran utilidad en la praxis de la investigación y en la enseñanza de la Regresión en las Facultades y Escuelas Técnicas estudiosas de las Ciencias Biológicas.

Los autores han procurado centrarse en cuestiones de interés práctico sin derivar hacia disquisiciones teóricas pero, al mismo tiempo, sin perder un nivel de rigor adecuado y han conseguido capítulos muy buenos, como el dedicado a la tan manida pero siempre escurridiza colinealidad, el menos usual de modelos inclusivos o el dedicado a la discusión e interpretación de ejemplos prácticos. Creo que la lectura de estos capítulos debe resultar interesante y útil para cualquier estadístico aplicado.

Es inevitable en libros como éste, encontrar alguna ausencia: Robustez, poblaciones finitas, técnicas de cálculo, modelos con autocorrelación, etc. Pero creo que los autores han acertado también en la selección de temas.

Me satisface, por todo ello, prologar este libro y además porque muestra, otra vez, la posibilidad de una colaboración estrecha y fructífera entre investigadores de países distintos, en este caso franceses y españoles pertenecientes a los respectivos institutos de investigación agraria, INRA e INIA. Esta colaboración ha podido llevarse a la práctica gracias al apoyo institucional derivado del convenio existente entre ambos Institutos, dentro del programa de Biometría en que cooperan la Sección de Procesos de Datos del INIA, y el

excelente Departamento de Biometría del INRA, pero sobre todo gracias a la gran motivación, entusiasmo y competencia de sus autores.

Mi felicitación a todos ellos y, en particular, a J. B. Denis, de quien salió la propuesta inicial de este trabajo, y a E. Carbonell, sobre quien ha caído la ardua labor de supervisión de la edición del texto.

Javier Moro
Jefe de la Sección de Proceso de Datos del INIA

PREFACIO

El análisis de Regresión es una de las técnicas más utilizadas en el análisis e interpretación de los datos de experimentos agrícolas y biológicos cuando se estudia la variación conjunta de un grupo de variables.

Por su popularidad y facilidad de cálculo el uso de la regresión está ampliamente difundido entre los investigadores, pero es necesario reconocer que no siempre se utiliza de una manera adecuada.

A lo largo de esta publicación se pretende presentar de una forma más bien intuitiva los conceptos relacionados con el análisis de la regresión. El objetivo principal es proporcionar un manual de base para investigadores y estudiantes en ciencias aplicadas a la investigación en áreas relacionadas con la Agricultura, Biología, Medicina, etc., que necesitan conocer el lenguaje básico, lo que puede, y principalmente, lo que no puede resolver la regresión. Se es consciente de la dificultad que entraña el hecho de tener que mantenerse a un nivel de claridad y comprensión determinado para lograr los objetivos fijados, sin perder por ello rigurosidad en la exposición.

La obra se encuentra dividida en ocho capítulos más unos anejos. Se comienza con un capítulo que sirve de presentación del tema; en él se incluyen los tipos de modelos que se estudiarán en capítulos posteriores, así como las suposiciones de base que deben cumplir dichos modelos. Este capítulo prepara el escenario introduciendo al lector en la problemática que se tratará má adelante y por otro lado advierte de los aspectos que es necesario tener en cuenta, de tal modo que su no verificación invalida o al menos cuestiona gravemente la aplicabilidad científica de todos los cálculos matemáticos descritos. A continuación, en el capítulo segundo, se describe el modelo clásico de regresión simple y múltiple en sus aspectos de estimación y pruebas de hipótesis de los parámetros del modelo. Un enfoque matricial de la regresión simple sirve para introducir las fórmulas de la regresión múltiple sin problemas. Por ello, en el apartado dedicado a la regresión múltiple solamente se incluyen aquellos aspectos novedosos o de interpretación distinta a la regresión simple. Al estudias dos o más variables regresoras, puede que se presenten problemas que no aparecían en regresión simple; uno de ellos, la colinearidad, se estudia en el capítulo tercero con indicaciones de su detección, efectos y estrategias a seguir para su tratamiento. Detectada y tratada la colinearidad, el investigador puede estar interesado en encontrar un subconjunto de variables regresoras que sean suficientes para describir adecuadamente el comportamiento de la variable dependiente. Por tanto, en los dos capítulos siguientes se describen los métodos de selección de variables reservando el segundo de ellos para el método apriorístico de selección dirigida, mediante los modelos inclusivos; una aplicación de los modelos inclusivos que se trata con más detalle corresponde al estudio de la heterogeneidad de las pendientes de diversas ecuaciones de regresión simple. Una vez obtenida la ecuación

de regresión es necesario preguntarse si el modelo elegido puede considerarse como «correcto». Así pues, el capítulo sexto está dedicado al importante tema de poner en duda el modelo de regresión obtenido; es decir, se pretende revisar y validar el modelo a través del estudio de las hipótesis de base y del propio modelo. Para la verificación del modelo estimado se pone bastante énfasis en el estudio de los residuos. En el capítulo séptimo se presenta un enfoque geométrico a la regresión como una presentación alternativa a la descripción algebraica realizada en capítulos anteriores. Este es un capítulo complementario para introducir al lector en otros aspectos de la regresión que por su facilidad de visualización pueden resultar muy interesantes. Por último, el capítulo octavo consiste en una serie de ejemplos con objeto de resumir todo lo presentado y mostrar un modo de interpretación de los resultados obtenidos por el análisis de regresión.

A lo largo de toda la publicación se han mantenido dos tipos de letra. La letra que se podría denominar «normal» está dirigida a aquellos lectores que sólo están interesados en estudiar superficialmente el problema de la regresión o para una primera lectura de la publicación. La justificación matemática o conceptos algo más especializados se han incluido en el texto en otro tipo de letra.

Esta obra sale a la luz gracias a la colaboración entre dos Institutos de investigación homónimos; uno, por parte española (INIA) y otro, de parte francesa (INRA), y será, por tanto, publicada en ambos idiomas. Los autores desean expresar su agradecimiento a aquellas personas de ambos institutos que han allanado el camino para que esta colaboración se haya llevado a buen término y sea base para futuras nuevas colaboraciones. Muy particularmente, al Dr. Javier Moro, Jefe de la Sección de Proceso de Datos del INIA y al Dr. R. Tomassone, Jefe del Departamento de Biometría del INRA por el constante apoyo no sólo en el aspecto burocrático, sino también por las sugerencias aportadas para mejorar el contenido de la publicación. Asimismo, los comentarios de doña Susana Pérez y don Ignacio Ibáñez, han sido de gran utilidad. Los errores o imprecisiones que todavía puedan existir no son imputables a ellos, sino que son plena responsabilidad de los autores. La ayuda técnica de las señoritas Rosa M. Adrada y Nuria Martín ha sido realmente inestimable al tener que soportar la dura tarea de escribir a máquina las muchas versiones que de esta publicación se han efectuado antes de llegar a su formato final.

CAPITULO 1

MODELOS PREDICTIVOS LINEALES

1.1. Introducción

Cuando se estudian conjuntamente dos o más variables, es lógico que el inevstigador trate de conocer y evaluar cuál es la relación que existe entre ellas. El *grado* de asociación o relación lineal entre variables viene medido por el coeficiente de correlación simple o múltiple; sin embargo, cuando se postula una relación entre variables que implique una relación funcional (es decir, expresable según una función $y=f(x_1, x_2...x_k)+\varepsilon$) debe buscarse un medio que exprese la *forma* de esa relación. Además sería deseable no sólo encontrar la función matemática $f(x_1, x_2, ..., x_k)$ que ligue a las variables sino también saber con qué precisión se puede predecir el valor que toma una variable y, para valores dados del resto de variables; en definitiva, tener una idea de la magnitud del componente aleatorio ε. Estos dos aspectos se engloban dentro de los objetivos del análisis de regresión. Por tanto, la regresión pertende determinar la «mejor» relación funcional entre variables. El término «mejor» debe interpretarse en el sentido de encontrar aquella ecuación que, según un criterio predeterminado, mejor se ajuste a los datos experimentales observados. Naturalmente, dependiendo del criterio elegido, los resultados obtenidos pueden ser diferentes.

En cualquier análisis de regresión se espera que la supuesta función o ecuación, represente algún mecanismo causal o básico asociado a las unidades experimentales y a los factores que se están estudiando. Por ejemplo, se conoce con toda certeza que el mecanismo que relaciona al voltaje V, con la intensidad, I, que pasa por un conductor se expresa mediante una función matemática lineal $V=RI$ en donde R es el cambio producido en V al variar I en una unidad. Ahora bien, generalmente la Ciencia no está tan avanzada de tal modo que se conozcan con certeza las variables y los mecanismos básicos de un proceso cualquiera. Sin embargo, incluso en estos casos, las técnicas de regresión son útiles como herramientas analíticas y predictivas. En el supuesto de conocer la relación, simplemente se estimará, por ejemplo, el parámetro desconocido R que relaciona V con I, intentando encontrar la mejor relación funcional entre variables a base de utilizar un modelo y unas técnicas de optimización. Obviamente, la bondad del resultado dependerá de la adecuación del modelo propuesto y del método de optimización empleado. Por tanto, debido a ello debe tenerse muy en cuenta que, como señalan Ostle y Mensing (1975, pág. 166), simplemente porque se haya supuesto una relación funcional particular entre variables (modelo) y se haya seguido un procedimiento de cálculo determinado (optimización), no se debe pretender que siempre exista tal relación causal entre las variables. Es decir, sólo porque se haya encontrado una función matemática que se ajuste bien a los datos, no se está necesariamente en situación de inferir con certeza que un cambio en una variable, *causa* un cambio en otra variable. En resumen, la única persona capaz de asegurar que las variables básicas son las realmente usadas y

que el mecnismo básico opera de acuerdo con la función matemática elegida es quien conoce el tema en el que se ha desarrollado el experimento, es decir, el propio investigador.

Cuando se desconoce la relación básica, se puede escoger una relación funcional determinada en base a consideraciones analíticas en relación con el fenómeno estudiado o bien a posteriori a través del examen de gráficos o diagramas dibujando los datos experimentales en ejes coordenados. Una vez que se tome una decisión sobre el tipo de función matemática a emplear, es necesario estimar los parámetros que definen a esa función. Por ejemplo, si esta relación es del tipo lineal $Y=\beta_0+\beta_1 X$, es necesario obtener los valores β_0 y β_1 que definen la ecuación de la recta.

En esta publicación se considerarán modelos que son lineales, es decir, que la función matemática más sencilla, implicando un cambio en la variable dependiente «y», al variar la independiente «x», corresponde a la ecuación de una recta $y=b_0+b_1x$. En esta ecuación b_0 y b_1 son los parámetros o cantidades fijas pero desconocidas, en donde b_0 representa la ordenada en el origen y b_1 la pendiente de la recta (o cambio producido en la «y» al variar la «x» en una unidad). El término lineal debe interpretarse en el sentido de que el modelo es lineal en los parámetros que lo definen. Según esto, el modelo

$$y = b_0 + b_1 x + b_{11} x^2$$

es, para todos los efectos, lineal aunque la «x» se encuentre elevada al cuadrado.

En correspondencia con estos modelos matemáticos en los que dados los valores de los parámetros, para cada valor fijo de «x» se obtiene un *valor fijo* de «y», existen los modelos estadísticos que implican la presencia de una perturbación aleatoria y unas distribuciones impuestas sobre la variable dependiente Y y, a veces, sobre ésta y la independiente X (que a partir de ahora denominaremos regresora) de forma conjunta. Como norma, se denotarán con mayúsculas o con letras griegas los parámetros y aquellas variables que sean aleatorias, reservando las corerspondientes minúsculas para las estimas de los parámetros, las realizaciones de éstas a través de una muestra, o para las variables matemáticas (aquellas que no llevan asociadas ninguna distribución estadística). Por tanto, el modelo estadístico que corresponde a una variación lineal de la Y se expresa como:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

en donde ε representa la perturbación aleatoria.

Igualmente, se incluirán aquí aquel grupo de modelos que, aun siendo no lineales, puedan considerarse como intrínsecamente lineales. Estos modelos intrínsecamente lineales o linearizables, son aquellos que pueden expresarse en forma lineal por medio de transformaciones adecuadas de las variables.

Por ejemplo, sean los modelos siguientes:

$$\text{I.} \quad Y = \gamma_0 X_1^{\gamma_1} X_2^{\gamma_2} \varepsilon$$

que se transforma en

$$\log Y = \log \gamma_0 + \gamma_1 \log X_2 + \gamma_2 \log X_2 + \log \varepsilon$$

Este modelo se utiliza fundamentalmente en problemas económicos y es la conocida función de Cobb-Douglas en donde los valores de γ_1 y γ_2 representan las elasticidades o porcentajes de cambio en la Y al variar en un 1 por 100 la variable regresora, permaneciendo constante la otra variable regresora.

$$\text{II.} \quad Y = \gamma_0 e^{\gamma_1 X_1} \varepsilon$$

que se transforma en

$$\text{Ln } Y = \text{Ln } \gamma_0 + \gamma_1 X_1 + \text{Ln}\varepsilon$$

El modelo representado por la ecuación anterior es el clásico del crecimiento exponencial de microorganismos antes de alcanzarse la fase de latencia.

$$\text{III.} \quad Y = \frac{1}{\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon}$$

que se transforma en

$$\frac{1}{Y} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon$$

Existen una serie de cuestiones que deben de tenerse muy presentes en la interpretación de los resultados obtenidos al aplicar un análisis de regresión sobre modelos intrínsecamente lineales.

Por ejemplo, con respecto al modelo I debe resaltarse el hecho de que el error experimental (perturbación aleatoria) actúa de forma multiplicativa sobre el modelo y no aditivamente. Por tanto, ese modelo es totalmente diferente al siguiente:

$$Y = \gamma_0 X_1^{\gamma_1} X_2^{\gamma_2} + \varepsilon$$

que es un modelo no lineal, ni tampoco linearizable por transformaciones.

Como más adelante se explicará, existen unos supuestos previos que deben cumplir algunos elementos del modelo para que las conclusiones obtenidas con este análisis sean válidas. Para efectuar pruebas de hipótesis con respecto a los parámetros, se exige que la variable aleatoria Y, y por tanto ε, se distribuya normalmente.

Por tanto, en los modelos I y II se debe suponer que la nueva variable $\varepsilon^* = \log\varepsilon$ se distribuye normalmente. Es decir, si se supone que es el error experimental del modelo, ε, quien se distribuye normalmente, las pruebas de significación serán erróneas.

Igualmente, la estimación mínima cuadrática solamente se aplicará a la ecuación transformada. Por lo tanto, al efectuar la transformación inversa para recuperar el modelo inicial, el valor estimado de los parámetros iniciales del modelo ya no será mínimo cuadrático, si la transformación efectuada no es lineal.

Los modelos no lineales son aquellos en los que los parámetros del modelo se encuentran en una función no lineal.

Dentro de los modelos no lineales, los más conocidos y estudiados en agricultura se refieren a las curvas de crecimiento. Uno de los más clásicos es el modelo de Brody, representado por la ecuación.

$$Y = P_1(1 + P_2 e^{-P_3 X}) + \varepsilon$$

Este modelo no es susceptible de linearización por lo que su estudio debe efectuarse dentro de las técnicas no lineales. Dado el contexto y el enfoque de la presente publicación, no se entrará en la descripción de estas técnicas.

1.2. Tipos de modelos según la distribución

A continuación, se describen unos modelos lineales estadísticos que se diferencian entre sí según la suposición efectuada sobre la distribución de las variables del modelo. Tal como se dijo anteriormente, los modelos estadísticos se diferencian de los matemáticos determinísticos precisamente en la parte de aleatoriedad que sobre ellos se impone. La existencia de la aleatoriedad, introducida generalmente por la presencia en el modelo del término aleatorio representado por ε, implica la necesidad de definir una función de distribución de las variables del modelo.

En la mayoría de los textos sobre regresión (ver, por ejemplo, SPRENT, 1969) se tratan dos tipos de modelos lineales diferentes: modelos aleatorios y modelos fijos. Su descripción se centrará en dos variables.

1.2.1. *Modelo Aleatorio*

En este primer tipo, X e Y son dos variables aleatorias con función de densidad conjunta f(x, y). La esperanza de Y condicionada a que X tome el valor x, se define como

$$E(Y/X=x) = \frac{\int_{-\infty}^{\infty} y f(x,y)\,dy}{\int_{-\infty}^{\infty} f(x,y)\,dy}$$

Esta expresión es una función de x por lo que al variar x obtendremos una curva que es la regresión de la variable aleatoria Y, sobre la variable aleatoria X.

Este tipo de modelo aparece cuando, de una distribución bivariante (X, Y), se extraen aleatoriamente parejas de valores (x_i, y_i). Precisamente, el término «regresión» lo usó por primera vez en un problema de este tipo F. Galton, en una serie de artículos (quizá el más conocido fue el publicado en 1886). Describió en términos matemáticos la tendencia de que padres altos tengan hijos altos y padres bajos, hijos igualmente bajos. Aunque existía esta tendencia, la distribución de alturas en la población no cambiaba de una generación a la siguiente, pues la altura media de los hijos de unos padres de una altura determinada, «regresaba», se movía, hacia la altura media de la población. El punto importante a resaltar aquí es que tanto la altura de los padres como la de los hijos se tomaba de forma aleatoria para representar una muestra de una serie de parejas de valores de una población con distribución conjunta bivariante. En este tipo de modelos, lo que se pretende es conocer algo del comportamiento de la altura media de los hijos a partir de la de sus padres. Sin embargo, no se utilizan para predecir la altura de un hijo de una familia determinada.

Si X e Y se distribuyen normalmente con medias μ_x y μ_y, respectivamente; desviaciones típicas σ_x, σ_y; y coeficiente de correlación ρ, la distribución condicionada de Y dado X=x, es normal con media

$$E(Y/X=x) = \mu_Y + \frac{\rho \sigma_Y}{\sigma_X} (x-\mu_X)$$

y varianza $\sigma_y^2 (1-\rho^2)$. Para su demostración puede consultarse ANDERSON (1958).

Si se hace variar x, el lugar geométrico es $y=E(Y/X=x)$, que corresponde a una recta que pasa por (μ_x, μ_y) y pendiente $\frac{\rho \sigma_y}{\sigma_x}$, siendo la regresión de Y sobre X tal como se ha indicado anteriormente. Es necesario destacar que «y» no es un valor obtenido por una variable aleatoria Y sino que es una variable matemática que, para un x dado, toma el valor exacto $E(Y/X=x)$.

Dado que en el modelo aleatorio, la distribución es bivariante, tiene sentido el preguntarse cuál es el valor medio de la variable aleatoria X cuando la Y toma un valor y; es decir,

$$E(X/Y=y) = \mu_X + \frac{\rho \sigma_X}{\sigma_Y} (y-\mu_Y)$$

por lo que se pueden obtener ambos modelos recíprocos.

Conviene indicar que la regresión obtenida según el modelo $E(X/Y=y)$ no equivale a la que se hallaría al despejar la x en el modelo $E(Y/X=x)$.

1.2.2. *Modelo fijo*

En este modelo se estudia el valor que toma una variable aleatoria Y para valores predeterminados de una (o unas) variable matemática x. Si el modelo es lineal, para una sola variable matemática, el comportamiento de la variable aleatoria Y viene representada por

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

en donde ε también es una variable aleatoria.

En el modelo fijo, en contraposición con el aleatorio, la distribución es univariada, ya que la Y es variable aleatoria, pero no la x. Siguiendo con el ejemplo de la distribución de alturas, x representa las alturas *predeterminadas* de unos padres e Y, la de sus hijos. Entonces, la altura de los padres representa una variable matemática ya que se ha seleccionado un conjunto de valores midiendo las alturas de los hijos cuyos padres tenían específicamente las alturas elegidas.

En una situación de modelo fijo resulta totalmente incorrecto intentar ajustar la ecuación x dado Y, puesto que aquí x no es una variable aleatoria. En todo lo que sigue se centrará el estudio en modelos fijos.

1.3. Suposiciones del modelo

1.3.1. *Especificación*

A lo largo de esta exposición se han mencionado una serie de suposiciones o hipótesis de base que deben cumplir las variables para que las conclusiones obtenidas con el uso de estas técnicas sean válidas. Estas hipótesis son:

i. Linearidad del modelo.

La variable dependiente se expresa como función lineal de los parámetros. Por ejemplo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Se supone también (hipótesis iv) que ε es una variable aleatoria de media cero; por tanto, esta expresión implica que la esperanza matemática de la variable Y es:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = \beta_0 + \sum_{h=1}^{k} \beta_h x_n$$

ii. Las variables regresoras son «matemáticas» o, en el caso de modelo aleatorio, aleatorias de valores conocidos.

iii. Independencia.

Las variables ε_i, para valores diferentes de las regresoras, son estadísticamente independientes. Es decir, el error experimental de una obsrvación no influye en el de otra.

iv. Normalidad y homoescedasticidad.

La variable ε se distribuye normalmente (esta suposición no es necesaria para la estimación de los parámetros del modelo). Su media es cero y su varianza constante independientemente de lo que valgan las regresoras. Como consecuencia, la Y se distribuye normalmente para valores fijos de las regresoras, con mdias E(Y) y varianza igual a la de ε que denotaremos por σ^2.

Si el término de error ε (o la variable aleatoria Y) tiene varianza constante se llama homoescedástico; por el contrario, cuando dicha varianza es diferente según el valor de x, se denomina heteroescedástico. Dos casos de heteroescedasticidad están ilustrados por la figura 1.1. y la homoescedasticidad se presenta en la figura 1.2.

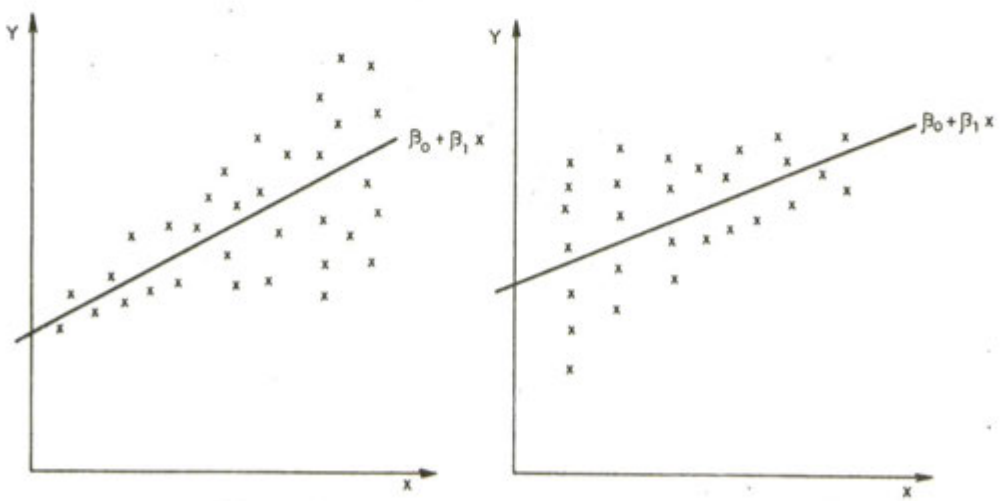


Figura 1.1.—Dos casos de heteroescedasticidad.

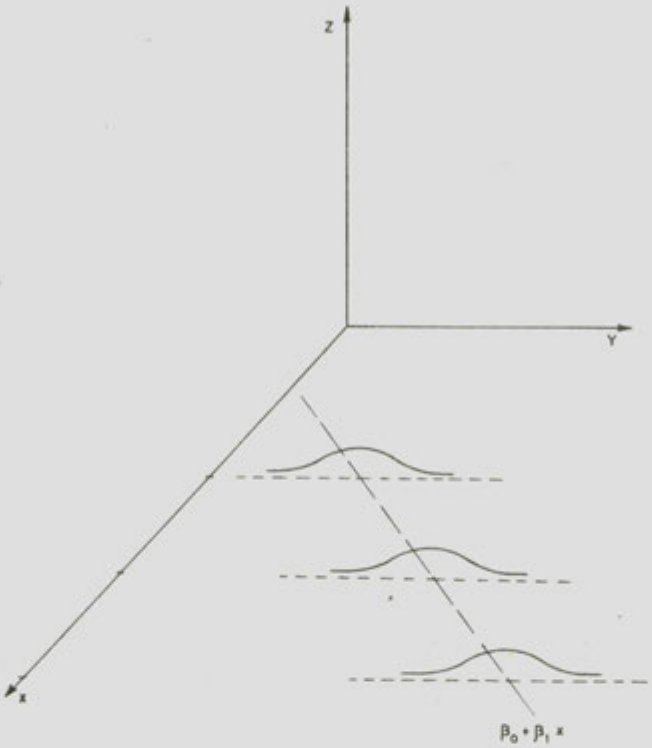


Figura 1.2.—Homoescedasticidad.

La primera intenta reflejar un grupo de datos reales, mientras que la segunda es una descripción conceptual del problema.

La figura 1.1. representa las parejas de datos (x, y) de un experimento cualquiera. En la figura 1.1.a se observa que la dispersión de los datos es mayor a valores altos de la x concentrándose más las observaciones con respecto a la recta cuando la x toma valores bajos. El caso opuesto queda reflejado en la figura 1.1.b. En la figura 1.2 se simboliza la distribución teórica de la variable aleatoria Y para cada valor de x. Como se indica, la media de la distribución va aumentando al hacerlo la x siguiendo la recta de regresión, pero la varianza, denotada por la dispersión de la curva normal, es constante al variar la x.

Cuando los términos del error correspondientes a observaciones diferentes sean dependientes (en sentido estadístico), se dice que el proceso está autocorrelacionado. Autocorrelación negativa indica que errores negativos en un período están asociados con errores positivos en el siguiente. Por otro lado, existe autocorrelación positiva cuando errores en un período están asociados con una pauta de comportamiento, siendo primero todos de un signo y después del contrario. Estas dos situaciones están reflejadas en la figura 1.3.

Como corolario a las hipótesis ii y iv, se está suponiendo implícitamente que el término del error no depende de las variables regresoras ya que,

$$E(x_i \varepsilon_i) = x_i E(\varepsilon_i) = 0$$

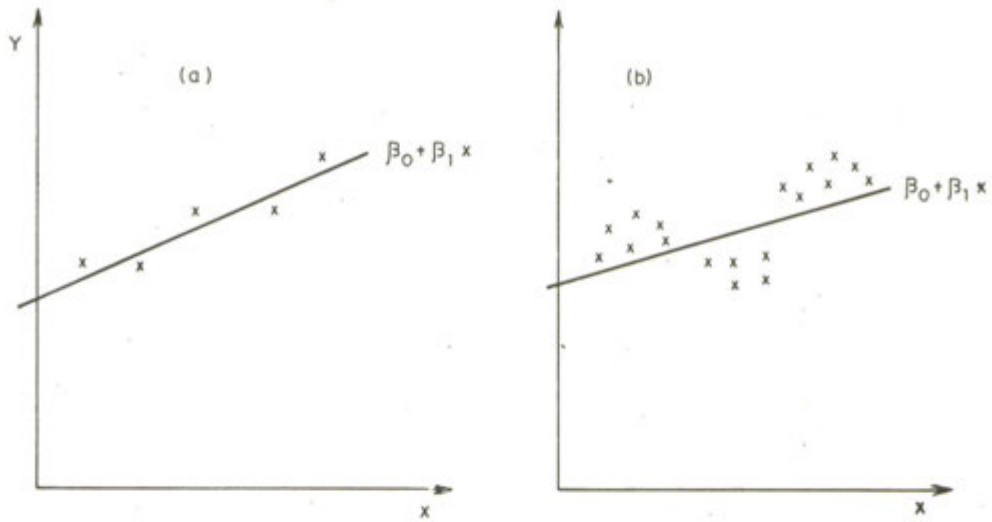


Figura 1.3.—Autocorrelación: *a*) negativa; *b*) positiva.

Esta propiedad es importante, pues afecta a la estimación de los parámetros.

Aunque no se considere como suposición o hipótesis de base, también es necesario tener en cuenta como importante el hecho de que, si se estudia más de una variable regresora, pudiera existir una relación lineal entre ellas. Este problema se llama colinearidad y se suele presentar cuando dos o más variables regresoras se encuentran altamente correlacionadas entre sí. Ahora bien, tal como más adelante se verá (capítulo 3), no existe una definición precisa para la colinearidad y es necesario analizar cuidadosamente la presencia de combinaciones lineales entre variables, pues, según el tipo de modelo que se esté ajustando, puede producir consecuencias totalmente diferentes. Sea, por ejemplo, la influencia de la granulometría de un suelo sobre una cierta variable Y. Naturalmente si las variables regresoras son los porcentajes de arcilla (x_1), limo (x_2) y arena (x_3), estas variables están relacionadas ya que $x_1+x_2+x_3=100$. Por tanto, si intentamos ajustar un modelo del tipo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

resulta que la variable regresora asociada a β_0 (que podríamos llamar x_0, en donde $x_0=1$) es precisamente la suma de las otras tres variables dividida por cien. En este caso existirá una colinearida extrema (ver definición en 3.1.1.), no pudiendo estimar los β_i por los métodos de regresión generalmente empleados. Sin embargo, si el modelo es

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

el problema desaparece, pudiéndose estimar la ecuación de regresión por mínimos cuadrados.

Esta salvedad tiene que ver con un concepto que se explicará más adelante en relación con el rango de la matriz **X** al presentar el enfoque matricial y, sobre todo, en el capítulo 3 dedicado específicamente al estudio de la colinearidad.

1.3.2. *Importancia de la violación de alguna hipótesis*

Como la mayoría de los efectos producidos por la violación de las hipótesis de base se refieren a estimación de los parámetros del modelo, postpondremos su presentación para más adelante, e incluso en algunos casos merecerá un capítulo aparte. Además, muchas veces no es posible efectuar las pruebas debido al tipo de datos de que se dispone, ya que algunas de ellas requieren la presencia de varias observaciones para cada x_i y otras, que los datos estén ordenados de acuerdo con algún criterio predeterminado. De todas formas, se puede señalar que por orden de importancia se pueden ordenar como sigue: independencia, homoescedasticidad y normalidad.

Con respecto a independencia, es necesario poner en duda la adecuación del método clásico de regresión en conexión con el estudio de las curvas de crecimiento. Si se trata de investigar el tipo de crecimiento que sigue un animal concreto, los pesos (variable dependiente) a unas edades determinadas del animal (variable regresora) se encontrarán correlacionadas debido principalmente a características intrínsecas del animal como composición genética, enfermedades, etc. Cuanto más próximas sean las edades, más alta será esta correlación. En este tipo de investigaciones debe de seguirse un procedimiento de cálculo distinto.

Para emplear los métodos aquí presentados es necesario disponer de varios animales describiendo el crecimiento de la estirpe o raza más que el de un animal determinado. La forma correcta de operar sería utilizar datos de edades procedentes de animales distintos para que, de esta forma, las medidas fueran independientes. De todos modos, si el número de animales no es demasiado pequeño, la influencia de las medidas correlacionadas puede considerarse despreciable. Esto proviene del hecho de que, si se dispone de «n» animales con «t» mediciones en cada uno de ellos, existirán $n\binom{t}{2}$ parejas de medidas dependientes de un total de $\binom{nt}{2}$ parejas, diluyéndose el efecto de las observaciones correlacionadas.

CAPITULO 2

REGRESION LINEAL: ESTIMACION

2.1. Regresión simple

Si basados en el conocimiento previo sobre el problema estudiado, se decide que la relación que liga a la variable Y con la variable x es de tipo lineal, el valor de Y para una específica x_i vendrá dado por la expresión, $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$ cuyo valor esperado es

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

Para una realización específica del experimento (es decir, para unos datos determinados) la expresión de este modelo será:

$$y_i = b_0 + b_1 x_i + e_i$$

en donde los valores con letras latinas serán las estimas de los parámetros de la expresión obtenidas a partir de los datos del experimento de que se trate.

2.1.1. *Estimación*

El procedimiento de estimación seguido es el llamado de mínimos cuadrados. Este método consiste en hacer mínimas las discrepancias elevadas al cuadrado y sumadas para todo i según la figura 2.1.

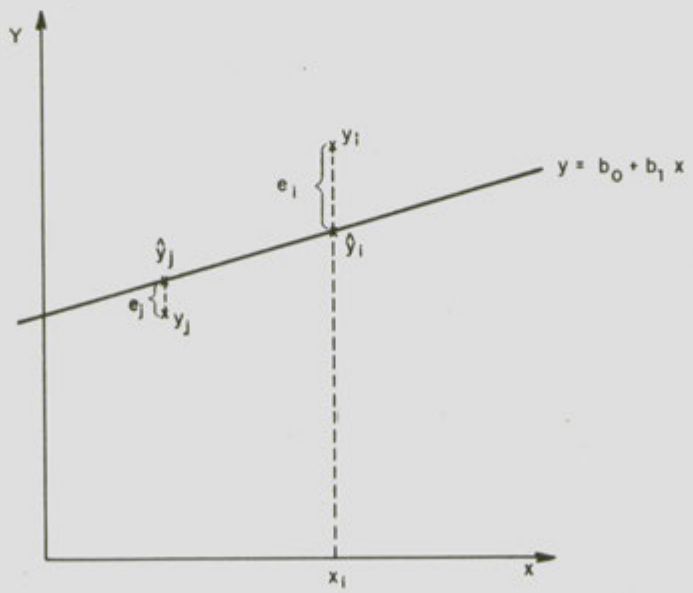


Figura 2.1.—Minimización de los residuos cuadráticos.

La razón de elevar al cuadrado las discrepancias es para evitar la anulación de las cantidades positivas y negativas al sumar. Para lograr esos mismos objetivos se han sugerido métodos alternativos. Uno de ellos consiste en sumar los valores absolutos de las discrepancias.

Este método ha sido usado ampliamente en estudios económicos. TAYLOR (1974), entre otros, lo recomienda como más apropiado para modelos predictivos econométricos. Incluso, recientemente, NARULA y WELLINGTON (1977) sugieren usar la suma mínima de errores relativos, $|e/y|$, para estimar los parámetros. Su método de estimación está formulado en términos de programación lineal.

Dado que se supone normalidad, el método de los mínimos cuadrados coincide con el de máxima verosimilitud (dar como estima el valor que hace más probable, a posteriori, la obtención de la muestra), lo que le da pleno valor estadístico.

Al aplicar la estimación minimocuadrática, lo que se pretende es, dado un modelo teórico $Y=\beta_0+\beta_1 x+\varepsilon$, tomando parejas de observaciones (y_i, x_i) en un experimento, tratar de estimar los parámetros β_0 y β_1 a partir de los datos. Se supone que b_0 y b_1 son las cantidades que estiman β_0 y β_1. Los β_0 y β_1 son constantes fijas desconocidas, mientras que b_0 y b_1 en un experimento concreto, son realizaciones que dependen del método de estimación empleado. En concreto, si se utiliza el método de los mínimos cuadrados, b_0 y b_1 deben minimizar la expresión siguiente:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Derivando parcialmente con respecto a b_0 y b_1 e igualando a cero, se obtienen dos ecuaciones con dos incógnitas que resueltas darán los valores estimados. Estas dos ecuaciones se conocen con el nombre de ecuaciones normales.

$$n\, b_0 + \sum_i x_i b_1 = \sum_i y_i$$

$$\sum_i x_i b_0 + \sum_i x_i^2 b_1 = \sum_i y_i x_i$$

Resolviendo el sistema se obtiene:

$$b_1 = \frac{\sum_i x_i y_i - \frac{(\sum_i y_i)(\sum_i x_i)}{n}}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}} = \frac{Sxy}{Sx^2}$$

$$b_0 = \frac{\sum_i y_i}{n} - b_1 \frac{\sum_i x_i}{n} = \bar{y} - b_1 \bar{x}$$

Dado que las expresiones del numerador y denominador de b_1 se emplearán con frecuencia en el texto, se representarán simbólicamente por Sxy y Sx^2. Este simbolismo tiene su lógica, pues equivale a Σxy y $E\Sigma^2$, pero cuando a las variables x e y se les ha restado su media respectiva. Es decir, podrían definirse como las sumas de productos y de cuadrados ajustados por las medias.

Estos resultados b_0 y b_1 son las estimaciones de los parámetros del modelo β_0 y β_1 y son cantidades obtenidas mediante operaciones algebraicas de los datos experimentales. Representan los valores que en la muestra han tomado unas variables aleatorias β_0 y β_1 que son los estimadores correspondientes a β_0 y β_1, respectivamente. Por tanto, «estimador» es una variable aleatoria y estimación o «estima» se entiende como el valor que toma el estimador en una muestra concreta.

2.1.2. *Pruebas de hipótesis e intervalos de confianza*

Los valores anteriores son estimaciones de los parámetros del modelo. Los estimadores que son variables aleatorias, tomarán valores diferentes al obtener muestras o realizaciones diferentes. Se debe, pues, estudiar cuál es la distribución de estos estimadores, con objeto de establecer pruebas de hipótesis sobre los parámetros del modelo.

En primer lugar, es necesario preguntarse si al variar la x se produce una variación concomitante en la Y (es decir, si $\beta_1 \neq 0$). Para ello, se descompone la variabilidad de los datos en dos componentes, una que mide la discrepancia entre el valor observado y otra que mide la discrepancia entre la recta estimada y una recta patrón, por ejemplo, $Y = \bar{y}$. Esta descomposición queda reflejada en la figura 2.2.

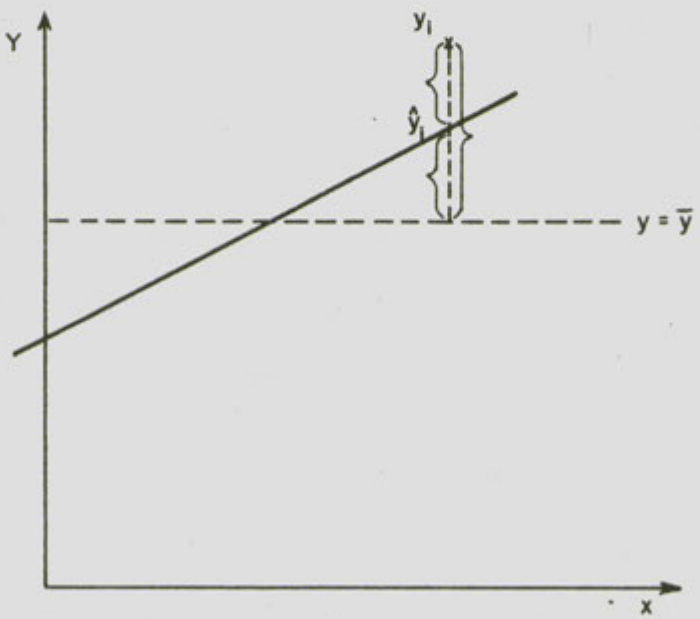


Figura 2.2.—Descomposición de (y_i-y) en (y_i-y_i) e (y_i-y).

Por tanto

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Elevando al cuadrado y sumando para todas las observaciones.

$$\sum_i (y_i - \bar{y})^2 = \sum_i \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2 =$$

$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Algebraicamente se puede demostrar que el doble producto es cero. En efecto,

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_i (y_i - b_0 - b_1 x_i)(b_0 + b_1 x_i - \bar{y}) =$$

$$= \sum_i (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i)(\bar{y} - b_1 \bar{x} + b_1 x_i - \bar{y}) =$$

$$= \sum_i \left[(y_i - \bar{y}) - b_1 (x_i - \bar{x}) \right] \left[b_1 (x_i - \bar{x}) \right] = b_1 \left[\sum_i (y_i - \bar{y})(x_i - \bar{x}) - \right.$$

$$\left. - b_1 \sum_i (x_i - \bar{x})^2 \right] = b_1 \left[b_1 \sum_i (x_i - \bar{x})^2 - b_1 \sum_i (x_i - \bar{x})^2 \right] = 0$$

Obteniéndose entonces la siguiente descomposición de la suma de cuadrados total:

$$Sy^2 = \sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

S. C. TOTAL = S. C. REGRESION + S. C. RESIDUO

Existen unas fórmulas alternativas más fáciles para el cálculo de la descomposición de la suma de cuadrados.

La expresión $\sum_i (\hat{y}_i - \bar{y})^2$ *se calcula como sigue:*

$$\sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (b_0 + b_1 x_i - \bar{y})^2 = \sum_i (\bar{y} - b_1 \bar{x} + b_1 x_i - \bar{y})^2 =$$

$$= \sum_i (b_1 (x_i - \bar{x}))^2 = b_1^2 \sum_i (x_i - \bar{x})^2 = b_1^2 Sx^2 = b_1 Sxy$$

Representando en forma tabular similar a la del análisis de varianza y llamando al residuo, «Desviación de la regresión», se tiene,

Fuentes de variación	*S. C.*	*g. l.*	*CM*
Debido a regresión	b_1Sxy	1	b_1Sxy
Desviación de la regresión $Sy^2 - b_1\ Sxy$		$n-2$	$(Sy^2 - b_1\ Sxy)/(n-2)$
Total	Sy^2	$n-1$	

Alternativamente, se puede hacer la descomposición de la suma de cuadrados de las observaciones brutas; es decir, sin corregirlas por el valor medio; en este caso sería:

$$y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Para obtener la descomposición de la suma de cuadrados se puede seguir un procedimiento análogo al anterior, es decir, elevar al cuadrado, sumar y demostrar que los dobles productos se anulan. Sin embargo ,aquí se utiliza un método más corto basado en la fórmula anterior:

$$Sy^2 = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2$$

Por tanto,

$$\sum_i y_i^2 - n\bar{y}^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

De donde

$$\sum y_i^2 = n\bar{y}^2 + \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

S. C. TOTAL S. C. MEDIA S. C. REG. S. C. RES.

La tabla correspondiente a esta descomposición es,

Fuentes de variación	*S. C.*	*g. l.*
Media	$n\ \bar{y}^2$	*1*
Regresión/media	b_1Sxy	*1*
Desviaciones	$Sy^2 - b_1Sxy$	$n-2$
Total	Σy_i^2	n

Para saber qué están estimando los CM de cada una de las líneas es necesario calcular previamente cuál es la varianza del estimador β_1.

Se sabe que:

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{Sx^2} = \frac{\sum_i y_i (x_i - \bar{x}) - \sum_i \bar{y}(x_i - \bar{x})}{Sx^2} =$$

$$= \frac{\sum_i y_i (x_i - \bar{x}) - \bar{y} \sum_i (x_i - \bar{x})}{Sx^2} = \frac{\sum_i y_i (x_i - \bar{x}) - \bar{y}(\sum x_i - n\bar{x})}{Sx^2} =$$

$$= \frac{\sum_i y_i (x_i - \bar{x}) - \bar{y}(\sum_i x_i - \sum_i \bar{x})}{Sx^2} = \sum_i \frac{x_i - \bar{x}}{Sx^2} y_i = \sum_i K_i y_i$$

donde

$$K_i = \frac{x_i - \bar{x}}{Sx^2}$$

Tomando varianzas y dado que x_i no es variable aleatoria

$$var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sum_i (x_i - \bar{x})^2}{[Sx^2]^2} \sigma^2_y = \frac{\sigma^2_y}{Sx^2} = \frac{\sigma^2}{Sx^2}$$

Resumiendo, pues: var $(\hat{\beta}_1) = \frac{\sigma^2}{Sx^2}$

Siendo $\sigma^2_y = \sigma^2$, la varianza constante del error experimental, cuya estima es el cuadrado medio de las desviaciones del modelo (o cuadrado medio del residuo).

A partir de la varianza de $\hat{\beta}_1$ se puede calcular el valor estimado por el cuadrado medio de la regresión de la forma siguiente:

$$\sigma^2_{\hat{\beta}_1} = E(\hat{\beta}_1 - E(\hat{\beta}_1))^2 = E(\hat{\beta}_1 - \beta_1)^2 = E(\hat{\beta}^2_1 + \beta^2_1 - 2\hat{\beta}_1 \beta_1) =$$

$$= E(\hat{\beta}^2_1) + \beta^2_1 - 2\beta_1 E(\hat{\beta}_1) = E(\hat{\beta}^2_1) - \beta^2_1$$

de donde $E(\hat{\beta}^2_1) = var(\hat{\beta}_1) + \beta^2_1$

Por otro lado, dado que el cuadrado medio de la regresión es

$$\hat{\beta}^2_1 Sx^2$$

$E(\hat{\beta}^2_1 Sx^2) = Sx^2 E(\hat{\beta}^2_1)$ *Sustituyendo el valor anterior*

$$E(CM\ regresión) = Sx^2 (var(\hat{\beta}_1) + \beta^2_1) = Sx^2 \left(\frac{\sigma^2}{Sx^2} + \beta^2_1\right)$$

E (CM Regresión) $= \sigma^2_1 + \beta^2_1 Sx^2$

Igualmente se puede calcular qué estima el cuadrado medio del residuo. Para ello se parte de la esperanza de su suma de cuadrados:

$$E(Sy^2 - \hat{\beta}_1 Sxy) = E(Sy^2) - E(\hat{\beta}_1 Sxy) = E(Sy^2) - (\sigma^2 + \beta^2_1 Sx^2)$$

Por otro lado,

$$E(Sy^2) = E[\sum_i (y_i-\bar{y})^2] = E[\sum_i (\beta_0+\beta_1 x_i+\varepsilon_i-\beta_0-\beta_1\bar{x}-\bar{\varepsilon})^2] =$$

$$= E\{\sum_i [\beta_1(x_i-\bar{x})+(\varepsilon_i-\bar{\varepsilon})]^2\} = E\{\sum_i [\beta_1^2(x_i-\bar{x})^2+(\varepsilon_i-\bar{\varepsilon})^2+2\beta_1(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})]\} =$$

$$= E[\beta_1^2 Sx^2+\sum_i (\varepsilon_i-\bar{\varepsilon})^2]$$

ya que el doble producto es nulo por el corolario de las suposiciones ii y iv en 1.3.1. De donde

$$E(Sy^2) = \beta_1^2 Sx^2+(n-1)\sigma^2$$

Por tanto,

$$E(Sy^2-\hat{\beta}_1 Sxy) = \beta_1^2 Sx^2+(n-1)\sigma^2-\sigma^2-\beta_1^2 Sx^2 = (n-2)\sigma^2$$

Resumiendo, pues:

E (CM Regresión)$=\sigma_1^2+\beta_1^2 Sx^2$

E (CM Desviación)$=\sigma^2$

Por tanto, si $\beta_1^2=0$, la fórmula anterior implica que el cuadrado medio de la regresión está estimando σ^2 y entonces el cociente entre el CM de la regresión y el CM de las desviaciones se distribuirá según una F centrada. Así, pues, se puede verificar la hipótesis de que $\beta_1=0$ frente a la alternativa de $\beta_1\neq 0$ mediante una prueba F. Si el valor calculado del cociente es mayor que el correspondiente a una tabla de distribución F con los grados de libertad y nivel de significación dados, se rechaza la hipótesis concluyendo que la regresora influye sobre la variable dependiente.

Como ejemplo práctico de cálculo supóngase que se dispone de los siguientes datos:

x	7	6	8	3	5	2	10	2	10	9
y	13	12	12	9	9	8	15	6	17	14

A partir de ellos se obtiene que:

$$\sum x = (7+6+ \ldots +9) = 62$$

$$\sum x^2 = (7^2+6^2+ \ldots +9^2) = 472$$

$$\sum xy = (7)(13)+(6)(12)+ \ldots +(9)(14) = 805$$

$$\sum y = (13+12+ \ldots +17) = 115$$

$$\sum y^2 = (13^2+12^2+ \ldots +17^2) = 1429$$

Por tanto,

$$Sxy = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = 805 - \frac{(62)(115)}{10} = 92$$

$$Sx^2 = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = 472 - \frac{(62)^2}{10} = 87,6$$

$$Sy^2 = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = 1429 - \frac{(115)^2}{10} = 106,5$$

$$b_1 = \frac{Sxy}{Sx^2} = \frac{92}{87,6} = 1,05$$

$$b_0 = \bar{y} - b_1 \bar{x} = 11,5 - (1,05)(6,2) = 4,99$$

Para probar la hipótesis $\beta_0 = \beta_1 = 0$ se construye la tabla del análisis de varianza.

Fuentes de variación	*S. C.*	*g. l.*	*C. M.*	*F. c.*
Debido a la media	$n\bar{y}^2 = 1322,5$	1	1322,5	1068,69
Debido a la regresora ...	$b_1 Sxy = 96,6$	1	96,6	78,06
Residuo	$Sy^2 - b_1 Sxy = 9,9$	8	1,24	
Total	$\Sigma y^2 = 1429$	10		

Como ambos valores de Fc son mayores que el correspondiente a una distribución F con 1 y 8 grados de libertad al 1 por 100, concluimos que β_0 no es cero y que tampoco β_1 lo es; es decir, la x aporta información adicional a la suministrada por la media en el modelo.

Por otro lado, en lugar de verificar una hipótesis con respecto a los parámetros puede interesar obtener un intervalo de confianza del $(1 - \alpha)$ porcentaje.

Dado que se ha supuesto que la variable Y se distribuye normalmente, β_0 y β_1 también tienen ese tipo de distribución. Así, pues, el intervalo de confianza vendrá dado por la fórmula general que se aplica a estimadores con distribución normal: [Estimador $\pm t_{\nu, \alpha}$ (desviación típica estimada del estimador)], en donde ν son los grados de libertad asociados a la t. (Recuérdese

que la fórmula del intervalo de confianza para una media es $\bar{x} \pm t_{\nu;\alpha} s/\sqrt{n}$, en donde $s/\sqrt{n}$ es la desviación típica estimada de x.) Por tanto, aplicando esta fórmula al caso de $\hat{\beta}_1 \pm t_{n-2,\alpha}\hat{\sigma}_{\beta_1}$. El valor de $\hat{\sigma}_{\beta_1}$ se calcula sustituyendo el σ^2 por CM de las desviaciones en la fórmula $\sigma^2_{\beta_1}=\sigma^2/Sx^2$. Naturalmente, la amplitud del intervalo dependerá de la varianza de la estima, siendo más pequeño, y por tanto mayor precisión, cuando menor sea ésta.

Estudiando la forma de var $(\hat{\beta}_1)$ se observa que disminuye (aumentando, por tanto, la precisión de la estimación) al disminuir σ^2 o al aumentar Sx^2. Para disminuir el valor de la varianza del error experimental ε es necesario seguir las recomendaciones generales que se indican para ello en cualquier diseño experimental. Para aumentar Sx^2 deben usarse valores de x lo más alejados posibles. Es decir, concentrar las observaciones en los dos extremos del rango de variación estudiado para la x (también llamado recorrido). Esta práctica, recomendada por muchos libros, es un arma de doble filo. Solamente es correcta cuando el modelo verdadero es precisamente lineal. Cuando esta suposición no es cierta, el resultado de seguir esta recomendación puede ser desastroso. Si el modelo propuesto es cuadrático, por ejemplo, al no tomar medidas repartidas a lo largo del rango de variación será imposible detectar la posible curvatura de la ecuación.

Si se quiere obtener un intervalo de confianza para la ordenada en el origen β_0, la fórmula es

$$\hat{\beta}_0 \pm t_{n-2,\alpha}\hat{\sigma}_{\hat{\beta}_0}$$

Por tanto, es necesario obtener previamente la varianza del estimador $\hat{\beta}_0$.

Dado que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ *tomando varianzas.*

$$var(\hat{\beta}_0) = var(\bar{y}) + \bar{x}^2 var(\hat{\beta}_1) - 2\bar{x}\, cov(\bar{y}, \hat{\beta}_1)$$

ya que x es una variable matemática y no estadística. Se puede demostrar que la cov $(\bar{y}, \hat{\beta}_1)$ *es cero. Para ello hay que tener en cuenta la siguiente propiedad:*

Si u_i *y* v_i *son constantes, y*

$$u = u_1 y_1 + u_2 y_2 + u_3 y_3 + \ldots + u_n y_n$$

$$v = v_1 y_1 + v_2 y_2 + v_3 y_3 + \ldots + v_n y_n$$

Si y_i *e* y_j *no están correlacionadas para* $i \neq j$, *y var* $(y_i)=\sigma^2$ *para todo i, entonces*

$$cov(u, v) = (u_1 v_1 + u_2 v_2 + u_3 v_3 + \ldots + u_n v_n)\sigma^2$$

en este caso $u=\bar{y}$, $v=\hat{\beta}_1$, *por lo que*

$$u_i = \frac{1}{n}, \quad v_i = \frac{x_i - \bar{x}}{Sx^2}, \quad \text{luego } cov(\bar{y}, \hat{\beta}_1) = \sum_i \frac{x_i - \bar{x}}{nSx^2}\,\sigma^2$$

$$= \sigma^2 \frac{\sum x_i - \sum x_i}{n S x^2} = 0$$

Por tanto,

$$var(\hat{\beta}_0) = var(\bar{y}) + \bar{x}^2 \, var(\hat{\beta}_1) =$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{Sx^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{Sx^2} \right)$$

El valor estimado de var ($\hat{\beta}_0$) por la muestra se obtiene sustituyendo en la fórmula anterior el valor σ^2 por el cuadrado medio del residuo; es decir,

$$\hat{\sigma}^2_{\hat{\beta}_0} = CM \text{ (desviaciones)} \left(\frac{1}{n} + \frac{\bar{x}^2}{Sx^2} \right)$$

2.1.3. *Predicción*

Muchas veces el investigador no sólo está interesado en saber qué tipo de relación liga a la variable x con la Y, sino que también pretende establecer predicciones con respecto al valor que tomará la variable Y al tomar la x un valor x_w. Naturalmente, el valor estimado se obtendrá sustituyendo x_w en el modelo predictivo:

$$Y_w = \beta_0 + \beta_1 x_w + \varepsilon_w$$

cuyo estimador es $\hat{\beta}_1 + \hat{\beta}_1 x_w + \hat{\varepsilon}_w$, quien para un experimento concreto se estima por $y_w = b_0 + b_1 x_w + 0$, ya que según 1.3.1. la esperanza de ε es cero.

Ahora bien, es necesario preguntarse la precisión con que se está efectuando la predicción (es decir, la magnitud de la varianza de ε_k). La bondad de la predicción dependerá de muchas circunstancias: bondad del modelo propuesto, varianza de los valores estimados, etc.

Existen diversos métodos para evaluar la bondad del modelo. Estos métodos no son de uso exclusivo en regresión simple, pues pueden aplicarse a la regresión múltiple, de la que se hablará más adelante. Incluso algunos son mejor utilizados en múltiple que en simple. Todos estos métodos se presentarán en el capítulo 6.

La precisión con que se efectúa la predicción de un valor de la variable dependiente para un valor determinado de la regresora viene dada por la varianza del valor estimado. Dependiendo de qué se quiere predecir, la fórmula de la varianza será diferente. Si lo que se pretende predecir es el valor *medio* de la variable dependiente para un valor x_w de la regresora, su varianza es:

$$var(\bar{\hat{y}}_w) = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_w)^2}{Sx^2} \right]$$

Por el contrario, el valor predicho de una observación *individual* sigue siendo $\hat{y}_w$, pero su varianza es diferente, ya que el valor observado varía alrededor de la media verdadera con varianza σ^2. Por tanto, como ambos efectos son independientes, su varianza será:

$$\text{var}(\hat{y}_w) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x}-x_w)^2}{Sx^2} \right)$$

Las estimas correspondientes a estas varianzas se obtendrán sustituyendo σ^2 por el cuadrado medio del residuo (o desviaciones de la regresión).

Como se desprende de las fórmulas, la varianza depende del cuadrado de la diferencia entre el valor de predicción y el valor medio. Por tanto, la precisión es máxima en el punto $\bar{x}$.

El intervalo de confianza del $(1-\alpha)$ por 100 para el valor medio predicho $\bar{\hat{y}}_w$ viene dado por la fórmula $\bar{\hat{y}}_w \pm t_{n-2,\,\alpha}\, \sigma_{\bar{\hat{y}}_w}$ que está representado gráficamente en la figura 2.3.

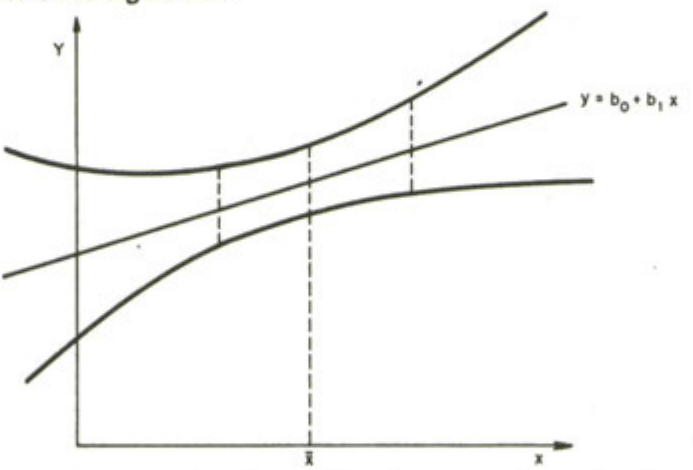


Figura 2.3.—Intervalo de confianza para un valor estimado y_w.

Esta gráfica da una clara idea de los peligros en la extrapolación de la predicción, puesto que el intervalo de confianza crece al alejarse del punto $\bar{x}$.

2.1.4. *Enfoque matricial*

Todas las fórmulas dadas anteriormente utilizando sumatorios son susceptibles de ser representadas bajo notación matricial. Utilizando el modelo de regresión lineal simple se estudiará la equivalencia entre ambas notaciones. Con ello, el pasar de regresión simple a regresión múltiple será simplemente una extensión de las fórmulas ya presentadas en este apartado.

Se definen: el vector aleatorio, *y;* la matriz de variables regresoras, *X;* el vector de parámetros a estimar, *b*, y el vector de errores, *e*.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad (n,1) \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix} \quad (n,2) \qquad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad (2,1) \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \cdot \\ \cdot \\ e_n \end{pmatrix} \quad (n,1)$$

La ecuación de regresión que liga a las observaciones $y_1 \ldots y_n$ con los valores $x_1 \ldots x_n$ puede expresarse para cada ocurrencia experimental (y_i, x_i) como

Este conjunto de igualdades se expresa en forma matricial como $y = Xb + e$.

$$\left.\begin{array}{lll} y_1 & = b_0 + b_1 x_1 + e_1 \\ y_2 & = b_0 + b_1 x_2 + e_2 \\ y_3 & = b_0 + b_1 x_3 + e_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ y_n & = b_0 + b_1 x_n + e_n \end{array}\right\}$$

Según las definiciones:

$$X'y = \underset{(2xn)}{\begin{bmatrix} 1 & 1 & 1 & 1 \ldots 1 \\ x_1 & x_2 & x_3 & x_4 \quad x_n \end{bmatrix}} \underset{(nx1)}{\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_n \end{bmatrix}} = \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{bmatrix}$$

que representa el primer término de las acuaciones normales.

Por otro lado,

$$X'X = \underset{(2xn)}{\begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & x_4 & & x_n \end{bmatrix}} \underset{(nx2)}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

que son los coeficientes de b_0 y b_1 en las ecuaciones normales. Por tanto, se pueden expresar en forma matricial como:

$$X'X\ b = X'y$$

Premultiplicando por $(X'X)^{-1}$ se obtiene el valor estimado de los parámetros

$$\hat{b} = (X'X)^{-1}\ X'y$$

Al tratar con regresión múltiple, puede darse el caso de que no existiera inversa y sería necesario recurrir al concepto de g-inversa, que no se expondrá aquí. Cuando exista una combinación lineal entre las regresoras, el determinante de X'X será cero, por lo que habrá de aplicar las técnicas correspondientes para obviar esta dificultad. Otro caso en el que se presenta este problema es el mencionado en 1.3.1. En él, la primera columna de la matriz X multiplicada por 100 es exactamente igual a la suma de las otras tres. Por tanto, la X'X tendrá una fila (y columna) combinación lineal de las demás, no siendo una matriz de rango pleno.

$$(X'X)^{-1} = \frac{1}{n\ \Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & \Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} = \begin{bmatrix} \dfrac{\Sigma x_i^2}{nSx^2} & \dfrac{\bar{x}_i}{Sx^2} \\ \dfrac{-\bar{x}}{Sx^2} & \dfrac{1}{Sx^2} \end{bmatrix}$$

Si se representa simbólicamente $(X'X)^{-1}$ por

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

y recordando los valores para var $(\hat{\beta}_0)$ y var $(\hat{\beta}_1)$, se observa que

$$\text{var}\ (\hat{\beta}_0) = C_{11}\sigma^2$$
$$\text{var}\ (\hat{\beta}_1) = C_{22}\sigma^2$$

ya que

$$\left[\frac{1}{n}+\frac{\bar{x}^2}{Sx^2}\right]=\frac{Sx^2+n\bar{x}^2}{nSx^2}=\frac{\sum_i x_i^2-n\bar{x}^2+n\bar{x}^2}{nSx^2}=\frac{\sum x_i^2}{nSx^2}=C_{11}$$

Basándose en la propiedad de que cov $(\bar{y}, \hat{\beta}_1)=0$ (demostrada en 2.1.2), se calculará la covarianza entre los estimadores: cov $(\hat{\beta}_0, \hat{\beta}_1)$=cov $[(\bar{y}-$ $-\beta_1\bar{x}), \hat{\beta}_1]$=cov $(\bar{y},$ $\hat{\beta}_1)-\bar{x}$ cov $(\hat{\beta}_1, \hat{\beta}_1)=$

$$=-x\sigma^2_{\hat{\beta}_1}=\frac{-x}{Sx^2}\sigma^2,$$ que precisamente es $C_{12}\sigma^2$.

Por tanto, se ha encontrado una expresión muy importante de aplicación en regresión múltiple

$$\text{Var}(\hat{b}) = V(\hat{b}) = (X'X)^{-1}\sigma^2$$

La cantidad $b'X'y$ es igual a

$$b'X'y=\begin{bmatrix} b_0 & b_1 \end{bmatrix}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}=\begin{bmatrix} b_0 & b_1 \end{bmatrix}\begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_i \end{bmatrix}=$$

$$=b_0\Sigma y_i + b_1\Sigma y_i x_i = (\bar{y}-b_1\bar{x})\, n\bar{y} + b_1\Sigma y_i x_i =$$

$$= n\bar{y}^2 - b_1 n\bar{x}\bar{y} + b_1\Sigma y_i x_i = n\bar{y}^2+b_1\left(\Sigma y_i x_i - \frac{\Sigma x_i \Sigma y_i}{n}\right) = n\bar{y}^2+b_1 Sxy$$ de

donde $b_1 Sxy=b'X'y-n\bar{y}^2$, que es precisamente la suma de cuadrados debida a la regresión.

Igualmente,

$$y'y=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}=\Sigma y_i^2$$

Dado que $Sy^2 = \Sigma y^2 - n\,\bar{y}^2 = y'y - n\bar{y}^2$

se puede construir la tabla del ANOVA de la regresión en forma matricial.

Fuente	*g.l.*	*S.C.*
Debido a regresión	1	$b'X'y - n\bar{y}^2$
Desviación de la regresión	n — 2	$y'y - b'X'y$
Total	n — 1	$y'y - n\bar{y}^2$

Así, pues el coeficiente de determinación será

$$R^2 = \frac{b'X'y - n\bar{y}^2}{y'y - n\bar{y}^2}$$

El valor $n\bar{y}^2$ puede representarse en forma matricial como $y'1(1'1)^{-1}1'y$, siendo 1 un vector columna formado por n unos.

Por otro lado, definiendo $x_w' = (1, x_w)$ siendo x_w un valor determinado de la variable regresora.

$$\hat{y}_w = b_0 + b_1 x_w = (1, x_w) \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = x_w' b = b' x_w$$

Esta expresión corresponde a la del valor medio estimado de la variable dependiente y su varianza puede fácilmente demostrarse (ver, por ejemplo, SEARLE, 1971) que matricialmente equivale a

$$\text{var}(\bar{y}) = x_w'(X'X)^{-1} x_w\, \sigma^2$$

2.2. Regresión múltiple

Se trata ahora de considerar el problema en que un grupo de variables regresoras $x_1, x_2, \ldots, x_k$ se estudian conjuntamente, investigando su influencia sobre una dependiente Y. Si la relación lineal se puede expresar de la forma siguiente:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

responde al modelo lineal conocido, en donde la matriz χ es ahora:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} \cdots & x_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{nk} \end{pmatrix}$$

Para obtener los valores que minimocuadráticamente estiman β_0, β_1, ... β_k se sigue un procedimiento totalmente similar al presentado en el enfoque matricial de la regresión simple. Todo lo dicho entonces, puede repetirse aquí, La única salvedad estriba en que en la tabla del análisis de la varianza, los grados de libertad de la línea correspondiente a «Debido a Regresión» son k en vez de 1 y los de «Desviaciones de la Regresión» son n-k-1.

Mediante la prueba oportuna se puede verificar que $\beta_j=0$ para todo j, (j=1, ..., k) frente a la alternativa de que al menos un β_j es diferente a cero. Igualmente, se puede verificar que un subconjunto $\beta_h=0$ (h=1, ..., p; $p \leq k$). Esta prueba será presentada más adelante en el capítulo correspondiente a la selección de un conjunto de variables.

Es necesario resaltar también el significado de los coeficientes de regresión β_1, ..., β_k en una regresión múltiple. El valor de β_j expresa el cambio producido en la variable dependiente Y al variar en una unidad la variable regresora x_j, *manteniendo el resto de las variables regresoras constantes.* Se trata pues de coeficientes de regresión parcial y así es como deberían llamarse siendo precisos en el lenguaje.

Existe otro punto importante dentro de la regresión múltiple que se presenta a continuación. Si las variables regresoras son independientes (su correlación empírica es cero), la influencia de una de ellas sobre la variable Y no depende de las demás. En este caso, el coeficiente de regresión indica la variación en la dependiente al variar en una unidad la regresora considerada, independientemente de lo que puedan hacer el resto de las regresoras, incluso aunque alguna de ellas fuera eliminada de la ecuación de regresión. Además, la suma de cuadrados se puede descomponer ortogonalmente de la forma siguiente:

$$SC(x_1, x_2, x_3 \ldots x_k) = SC(x_1) + SC(x_2) + SC(x_3) + \ \ldots \ + SC(x_k)$$

Según esta descomposición se puede demostrar que la suma de cuadrados de un subconjunto de variables, dado que otro grupo de regresoras se encuentran ya en la ecuación, es

$$SC(x_1, x_2, x_3 / x_4, x_5, \ldots, x_k) = SC(x_1, x_2, x_3, x_4, x_5, \ldots, x_k) - SC(x_4, x_5, \ldots, x_k) = SC(x_1, x_2, x_3)$$

Por lo tanto, la información que aportan x_1, x_2, x_3 sobre el comportamiento de la Y es independiente de la que aporten x_4 ... x_k. Esto quiere decir que el orden de entrada de las variables en la ecuación no influye ni en el valor de su coeficiente de regresión ni en la suma de cuadrados (ver capítulo de modelos inclusivos). Ahora bien, esta situación no se suele presentar generalmente y las variables regresoras se encuentran más o menos correlacionadas entre sí. En estos casos el coeficiente de regresión y la suma de cuadrados de una regresora dependerán del resto de regresoras que se encuentren en el modelo. (Esta situación quedará bien patente al presentar la selección de un subconjunto de variables y su tratamiento será objeto del capítulo dedicado a la colinearidad.) Considerando el caso sencillo de dos regresoras, se puede descomponer la suma de cuadrados de dos formas diferentes suponiendo, tal como se hizo en regresión simple, que la ordenada en el origen entre siempre la primera en el modelo. Estas descomposicions son:

$$SC(total) = SC(\beta_0) + SC(x_1/\beta_0) + SC(x_2/x_1, \beta_0)$$

$$SC(total) = SC(\beta_0) + SC(x_2/\beta_0) + SC(x_1/x_2, \beta_0)$$

Estas dos descomposiciones deben de interpretarse de forma diferente ya que, en general,

$$SC(x_1/\beta_0) \neq SC(x_1/x_2,\beta_0)$$

pues la influencia de x_1 sobre la Y, medida por su suma de cuadrados, dependerá del hecho de que la x_2 se encuentre o no ya incluida en el modelo. Estas dos descomposiciones se presentarán también al comentar el enfoque gométrico de la regresión múltiple. Cuando se quiera probar exclusivamente si un β_j correspondiente a una regresora x_j, es cero, la suma de cuadrados asociada a esa hipótesis es

$$SC(x_j/x_1,x_2\ldots x_{j-1},x_{j+1},\ldots,x_k)=SC(x_1\ldots x_k)-SC(x_1,x_2\ldots x_{j-1},x_{j+1},\ldots,x_k)$$

con un grado de libertad; el cuadrado medio que actuará como denominador en la prueba F es el de desviaciones del modelo completo.

2.3. Regresión polinomial

Cuando, dada una sola variable regresora x, su influencia sobre la variable Y puede describirse mediante una ecuación en grado k, el modelo es:

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \beta_k x^k$$

Si ésta es la verdadera ecuación que representa el tipo de asociación entre la x y la Y, es posible estimar β_0, β_1, ..., β_k siguiendo la técnica de la regresión múltiple sin más que hacer $x=x_1$; $x^2=x_2$; $\ldots x^k=x_k$. Por lo tanto, se puede aplicar toda la metodología allí descrita sin introducir ningún concepto nuevo. Debemos resaltar, sin embargo, que en este tipo de regresiones es casi seguro que se presentarán los problemas de colinearidad, dada la correlación existente entre la variable x y la serie de sus potencias. Es, pues, recomendable que se extremen las precauciones para evitar el peligro asociado a este problema.

Por otro lado, la interpretación de los coeficientes de regresión es menos clara que en regresión múltiple, puesto que no se puede hacer variar a x_2, por ejemplo, manteniendo el resto de las x constantes. Por tanto, solamente se pueden interpretar como parámetros característicos que definen una función polinomial.

La regresión polinomial tiene una enorme utilidad práctica cuando se aproxima una regresión no lineal por medio de un desarrollo en serie por descomposición de Taylor.

CAPITULO 3

COLINEARIDAD

3.1. Definición

Aunque clásicamente los autores emplean el término «multicolinearidad», en este texto se utilizará el de «colinearidad», pues se considera inútil y redundante el prefijo «multi».

Es necesario tener presente que la colinearidad en su acepción estadística no se corresponde con ninguna definición matemática concreta. En efecto, existen una infinidad de estados intermedios entre la colinearidad extrema (que es, de hecho, la colinearidad en sentido algebraico) y la ausencia total de colinearidad. Sólo se caracterizarán y definirán ambos términos opuestos proporcionando un criterio para medir el grado de colinearidad de un conjunto de datos.

3.1.1. *Colinearidad extrema*

Se presenta colinearidad extrema cuando existe una (o varias) relacion(es) lineal(es) entre las variables regresoras x_1, x_2, ... x^k (representadas simbólicamente por x_j; j=1 ... k). Es decir, si existen u_j tales que

$$u_0 + \sum_{j=1}^{k} u_j \, x_{ij} = 0$$

para todo i=1 ... n, siendo *n* el número de observaciones.

Este problema aparece principalmente en el caso de regresión múltiple, aunque en el caso de manipulación automática de ficheros, la utilización a ciegas de un subconjunto de individuos puede ocasionar la constancia del valor del regresor, implicando, por tanto, la existencia de colinearidad.

En este capítulo se excluirá el término constante del modelo β_0, suponiendo implícitamente que tanto la variable dependiente y como las regresoras x_j están centradas; es decir ,se les ha restado sus medias respectivas:

$$\sum_{i=1}^{n} y_i = 0$$

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \text{para todo } j=1...k$$

3.1.2. *Ausencia de colinearidad*

Desde el punto de vista estadístico, se dice que no existe colinearidad

cuando todas las variables regresoras son ortogonales entre sí. Es decir, si se verifica que

$$\sum_i x_{ij}\, x_{ij'} = 0 \quad \text{para todo } j \neq j'$$

Dado que las variables son centradas, esta condición es equivalente a la nulidad de los coeficientes de correlación simple entre las regresoras.

3.1.3. *Medida del grado de colinearidad*

El valor absoluto de la correlación empírica entre dos variables regresoras varía entre 0 (ausencia de colinearidad) y 1 (colinearidad extrema). Así, pues, éste es el índice que se emplea para el caso de dos regresoras. De una forma general, se caracteriza el grado de colinearidad de un conjunto de regresoras por los valores propios de su matriz de correlaciones. Si al menos uno de los valores propios es nulo, hay colinearidad extrema; si todos son iguales, no existe colinearidad. En el caso particular de dos variables regresoras con correlación ρ, los valores propios son $1+\rho$ y $1-\rho$.

3.1.3.1. Valores propios y vectores propios

Sea C una matriz (por ejemplo, de correlaciones o covarianzas entre k variables x_1, x_2 ... x_k) cuadrada, simétrica y semidefinida positiva. Por semidefinida positiva se entiende el hecho de que cualquiera que sea el vector u de dimensión «k» el escalar u'Cu es positivo o nulo. (Esta propiedad se puede demostrar teniendo en cuenta que esta cantidad es la varianza de una variable z obtenida como combinación lineal de las departida: $x=u_1x_1+u_2x_2+ \ldots +u_kx_k$ siendo u_j, $j=1 \ldots k$ los coeficientes de dicha combinación lineal).

Se denomina primer valor propio a la varianza máxima sobre el conjunto de las variables z que son combinación lineal de las de partida con la restricción de que

$$\sum_{j=1}^{k} u_j^2 = 1$$

(esta restricción es indispensable, pues de lo contrario, el máximo sería infinito). Al vector «u» correspondiente se le llama primer vector propio.

El segundo valor propio se define como la varianza máxima sobre el conjunto de las variables z que son ortogonales al primer vector propio (es decir, covarianza o correlación entre ellos nula), manteniendo la misma restricción. Al vector asociado se le llama segundo vector propio. De forma similar se definen el resto de los vectores y valores propios.

Calcular los valores y vectores propios de una matriz también se llama diagonalizarla ya que esto se puede interpretar como un cambio de base que la transforma en una matriz diagonal. Cuando un valor propio es cero, implica que la variable descrita por el vector correspondiente es de varianza nula, por tanto, constante; es decir, las variables originales x_1, x_2 ... x_k son colineales

$$\sum_{j=1}^{k} u_j\, x_{ij} = u_0 \quad \text{para todo } i=1 \ldots n$$

Recíprocamente, si existe tal relación lineal, hay un valor propio nulo.

3.1.3.2. Análisis de componentes principales

Relacionado con el concepto que se acaba de exponer, se encuentra el análisis de componentes principales que será aplicado posteriormente en 3.2.3. y en 3.2.3.2. Este análisis consiste en la diagonalización de la matriz de covarianzas (o de la de correlaciones para el análisis de componentes principales tipificado). Los ejes principales son las variables que, siendo combinación lineal de las originales, son de varianza máxima y se encuentran definidas por los primeros vectores propios (naturalmente el número máximo posible de ejes principales será el de variables originales).

3.1.4. *Observaciones*

a) La colinearidad está directamente relacionada con las correlaciones entre las variables regresoras.

b) Cuando hay más de dos regresoras, no se puede taxativamente afirmar con exactitud que un caso concreto sea más o menos colineal que otro. Supóngase que para tres variables regresoras, en un caso los valores propios son 2.1; 0.5 y 0.4 y en otro 1.6; 1.3 y 0.1. El primer caso tiene el mayor valor propio siendo los otros dos valores pequeños, mientras que el segundo caso tiene el más bajo pero en cambio los otros dos son relativamente grandes. Por esta razón, y para comparación ,algunos autores proponen considerar como medida indicativa la razón entre el primer valor propio y el último. Este cociente se hace infinito en caso de colinearidad extrema.

c) En lugar de calcular los valores propios de la matriz de correlación, se podrían también hallar los de la de covarianza obteniendo el mismo tipo de información: independencia (u ortogonalidad) o colinearidad extrema. Sin embargo, hay dos razones fundamentales en favor del primer método. Primera, el cálculo numérico es más seguro y fiable. Segunda, y sobre todo, el resultado no varía como consecuencia del cambio de escala efectuado sobre las variables regresoras, y como ésta es una propiedad interesante de la regresión, conviene respetarla.

3.2. Efectos de la colinearidad

3.2.1. *Inestabilidad de la estimación de los coeficientes de regresión*

Cuando se ajusta un modelo con variables regresoras intrínsecamente muy relacionadas (como por ejemplo la estimación del volumen de un árbol en función de su altura y el diámetro al cuadrado) puede ocurrir que se obtengan coeficientes de regresión de signo opuesto al que sería de esperar en buena lógica. Este hecho puede ser consecuencia de varianzas de los estimadores demasiado grandes.

Si se toman, por ejemplo, dos variables regresoras medidas en unidades tipificadas (es decir, desviadas de su media y divididas posteriormente por su desviacinó típica) de tal modo que su varianza empírica sea la unidad, la matriz de varianzas-covarianzas de los dos estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ es proporcional a la inversa de la matriz de correlación. Téngase en cuenta que, según 2.1.3.

$$\text{var}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \sigma^2 (X'X)^{-1} = \sigma^2 \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}^{-1} = \sigma^2 \frac{1}{1-\rho^2}\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}$$

$$= \sigma^2 \begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}$$

, siendo ρ la correlación entre ambas *variables regresoras.*

Por tanto, var $(\hat{\beta}_1)=\frac{\sigma^2}{1-\rho^2}$ que tiende a infinito cuando ρ tiende a la unidad; es decir, cuanto más se aproxime a la colinearidad extrema. Nótese que en el caso de dos variables tipificadas var$(\hat{\beta}_1)$=var $(\hat{\beta}_2)=\frac{\sigma^2}{1-\rho^2}$, pero esta igualdad ya no se cumple a partir de tres regresoras.

3.2.2. *Dificultad de interpretar la parte de «explicación» de cada regresora*

Siguiendo en el caso de dos variables regresoras x_1 y x_2, se puede expresar la parte de la variación de la variable dependiente Y explicada por x_1 [simbólicamente representada por $\%(x_1)$], o por x_2 [$\%(x_2)$], o por x_1, teniendo en cuenta que x_2 ya está en la ecuación [$\%(x_1/x_2)$]; simétricamente, por x_2 dado x_1 [$\%x_2/x_1)$] y por x_1 y x_2 conjuntamente [$\%(x_1, x_2)$].

Si todas las variables se encuentran tipificadas, entonces

$$\text{var} \begin{pmatrix} y \\ x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & \rho \\ r_2 & \rho & 1 \end{pmatrix}$$

siendo r_1 y r_2 las correlaciones entre «$y \longleftrightarrow x_1$» y «$y \longleftrightarrow x_2$», respectivamente, y ρ la existente entre ambas variables regresoras.

Según esto, se obtiene que

$$\%(x_1) = 100\ r_1^2$$

$$\%(x_2) = 100\ r_2^2$$

$$\%(x_1/x_2) = 100\ (r_1 - \rho r_2)^2/(1-\rho^2)$$

$$\%(x_2/x_1) = 100\ (r_2 - \rho r_1)^2/(1-\rho^2)$$

$$\%(x_1, x_2) = 100\ (r_1^2+r_2^2-2\rho r_1 r_2)/(1-\rho^2)$$

Todas estas expresiones se pueden demostrar fácilmente de forma geométrica (Capítulo 7, figura 7.16).

En particular ,se puede verificar que

$$\% (x_1, x_2) = \% (x_1) + \% (x_2/x_1) = \% (x_2) + \% (x_1/x_2)$$

y si $\rho=0$ implica que

$$\% (x_1) = \% (x_1/x_2)$$

En general (con las matizaciones que más adelante se expondrán), cuanto más diferentes de cero sea $|\rho|$, menos se cumplirá la igualdad anterior. Estos conceptos son totalmente similares a los presentados en la regresión múltiple a nivel de sumas de cuadrados.

Esta dualidad en la descomposición produce una confusión en la interpretación de los efectos respectivos de las variables regresoras. Tal confusión está ligada a los datos, pues fueron ellos quienes produjeron la colinearidad, y estadísticamente es imposible separarla.

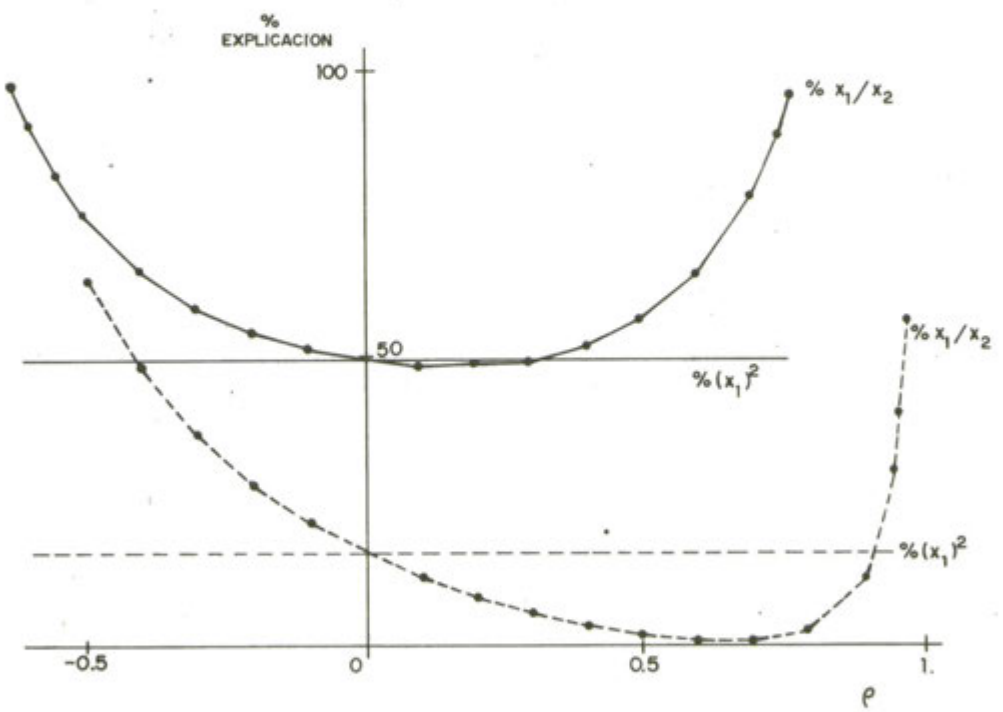


Figura 3.1.—Efecto de la colinearidad sobre el porcentaje de explicación de una de las regresoras. (Trazo continuo $r_1=0,7$, $r_2=0,1$; línea de puntos $r_1=0,4$, $r_2=0,6$.)

Los fórmulas precedentes y la figura 3.1. ponen de manifiesto que cuanto mayor sea el grado de la colinearidad, más difícil resulta establecer el aporte respectivo de cada uno de los regresores. En la figura se representa el efecto de la colinearidad ρ (abscisas) sobre el porcentaje de explicación de uno de los regresores ,ordenadas). La curva (a) es $\% (x_1)=r_1^2$ y la (b) es:

$$\%(x_1/x_2) = \frac{(r_1 - r_2)^2}{1-\rho^2} \times 100$$

En trazo continuo, se expresa el caso en que $r_1=0,7$; $r_2=0,1$ y en el trazo discontinuo cuando $r_1=0,4$; $r_2=0,6$. El rango de variación en el que se encuentra definido es (arc cos r_1 — arc cos r_2; arc cos r_1 + arc cos r_2), de tal manera que para los casos particulares de la figura 3.1 son [—0,64; 0,78] y [—0,49; 0,97], respectivamente. En la figua 3.1. se aprecia que, en general, conforme ρ se aleja del valor cero, menos coinciden ambos tipos de explicación de la variable Y por medio de la x_1. Ahora bien, es necesario reconocer que puede existir otros puntos en los que las curvas (a) y (b) coincidan, e incluso, como en la correspondiente a trazo continuo, que exista una zona en la que prácticamente sean iguales. Estas situaciones particulares dependen naturalmente de los valores r_1 y r_2. Así, pues, se observa que el tanto por ciento de explicación depende también en gran medida de las correlaciones r_1 y r_2 (comparación entre trazos continuo y discontinuo). Por tanto, para estudiar la influencia de la colinearidad, es necesario considerar además los valores de la variable dependiente. Por esta razón, WEBSTER, GUNST y MASON (1974) proponen examinar la colinearidad a partir de la diagonalización de la matriz de correlación de las regresoras orlada con la correlación de éstas con la dependiente. La crítica que se puede hacer a priori a esta proposición es la de no tener en cuenta el papel particular que desempeña la Y en relación con las variables regresoras.

3.2.3. *Influencia de la colinearidad cuando las regresoras no se conocen con exactitud*

Una de las hipótesis de base (ver 1.3.1.) es que las regresoras x_1, x_2 ... x_k son variables no estadísticas cuyos valores se conocen con certeza, como por ejemplo modificaciones de la temperatura, el ph, dosis dadas de un fertilizante, etc. Sin embargo, existen casos en que esta situación no se cumple. Como es difícil de formalizar la desviación del modelo con respecto a tal hipótesis, se ha procedido a simular esta situación para ver sus efectos con respecto a la colinearidad.

La simulación por medio de un ordenador consiste en generar un gran número de veces una situación experimental en la que algunos elementos son fijos y otros son aleatorios, variando según unas leyes determinadas (por ejemplo, distribuciones normales). Esto permite, en particular, estudiar estadísticamente cómo se comporta una fórmula de estimación determinada, una prueba concreta, etc. (dado el elevado número de repeticiones de la situación experimental básica en la que se opera). Permite, igualmente, considerar modelos diferentes a aquellos para los que se construyeron las pruebas y estimacions y juzgar así su robusted; es decir, su comportamiento frente a desviaciones del modelo.

En el caso que nos ocupa considérese el modelo de regresión con dos variables:

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

En donde ε son variables aleatorias normales, centradas e independientes (según las hipótesis de base clásicas).

Sin embargo, se incorpora a las variables fijas x_1 y x_2 una perturbación

aleatoria de distribución normal. Además se toman x_1 y x_2, de tal modo que tengan una correlación empírica $\rho=0$ o $\rho=0,7$; y los coeficientes de regresión β^2 y β_1 tales que $\beta_1 x_2+\beta_2 x_2$ tenga una correlación empírica con x_1+x_2 de cosθ, en donde θ varía de 0^o a 90^o de 10^o en 10^o. Este último parámetro se ha introducido porque, como $|\rho|$ es superior o igual a cero, x_1+x_2 representa el primer eje principal de los dos regresores (ver 3.1.3.2.) e interesa ver el efecto que tiene la Y sobre la medida de la colinearidad en relación con el hecho de que las variables regresoras no sean fijas. Para medir la discrepancia entre el modelo y la realidad, se ha utilizado la suma de los cuadrados de las desviaciones entre y_i e $\hat{y}_i$ (en donde $\hat{y}_i$ es el valor estimado de y_i por medio de la regresión). Los valores de esta suma de cuadrados (SCE) se presentan en la tabla 3.1. utilizando una muestra de tamaño 100 repitiendo la simulación 30 veces. Los diferentes parámetros empleados en la simulación se encuentran en la figura 3.2.

TABLA 3.1.

Suma de cuadrados del residuo (media de 30 simulaciones) en función de la colinearidad y de la localización de la variable dependiente

θ	$\rho=0$	$\rho=0,70$
0^o	5,8	4,2
10^o	5,5	4,6
20^o	5,7	5,1
30^o	5,5	6,2
40^o	5,2	7,1
50^o	5,0	8,2
60^o	5,3	9,6
70^o	5,9	11,6
80^o	5,6	11,4
90^o	6,8	11,9

De los resultados de la tabla 3.1. se pueden deducir las siguientes conclusiones:

a) Cuando las dos regresoras son independientes ($\rho=0$), sea cual sea la posición de Y (dada por el valor θ), la eficacia de la regresión no cambia, pues se obtienen valores del mismo orden de magnitud.

b) Por el contrario, en el caso de colinearidad ($\rho=0,7$), se comprueba que cuanto mayor sea el valor de θ, más desastroso será el efecto de la colinearidad si las regresoras no se conocen con exactitud.

c) Cuando los valores de θ estén próximos a cero (cuanto mayor sea la correlación de Y con el primer eje principal de x_1 y x_2) la presencia de colinearida des beneficiosa. Es decir, si las regresoras no se conocen con exactitud, es mejor que se enuentren correlacionadas si su eje principal está próximo a la variable dependiente.

Así pues, se concluye otra vez que el estudio de la colinearidad no se debe hacer sólo en base a las variables regresoras, sino que también es necesario considerar en el análisis la variable dependiente.

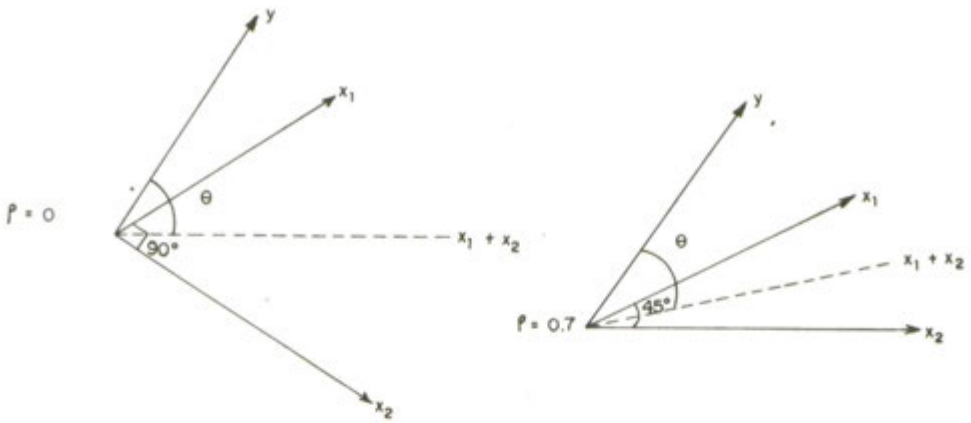


Figura 3.2.—Representación geométrica de los parámetros de la simulación de la

3.3. Estrategias a seguir

3.3.1. *Detección*

Es conveniente tomar conciencia de la posible existencia de colinearidad antes de actuar frente a ella. Si hay dos o más variables regresoras, ya se ha indicado al comienzo de este capítulo que el examen de la matriz de correlaciones no basta, aunque sí se ha detectado que un grupo de regresoras está correlacionado de manera importante, es muy probable que aparezcan problemas de colinearidad.

Si las pruebas de nulidad de los coeficientes de regresión conducen a la eliminación de variables que se consideran importantes, o si el signo del coeficiente de regresión es opuesto al que se espera, también aquí hay riesgo de encontrar una situación de colinearidad.

De hecho, solamente la diagonalización de la matriz de correlaciones y el examen de los últimos valores propios proporcionará una información precisa. Siendo k el número de regresoras, si designamos por $u_{(1)}$, $u_{(2)}$ ... $u_{(k)}$ los k valores propios en orden descendente, se verifica la relación

$$u_{[1]} + u_{[2]} + \ldots + u_{[k]} = k$$

Entonces si la razón $\frac{u_{[k]}}{k}$ es muy pequeña (inferior a 0.1 para cuatro variables o a 0.01 para una veintena) hay colinearidad. El grado de la colinearidad viene dado por el número de valores propios que se juzguen como muy pequeños. Naturalmente esta afirmación es subjetiva (el límite es arbitrario) y artifical (0.11 no indica colinearidad para cuatro variables y 0.09 sí), siendo consecuencia del hecho ya subrayado de que todos los casos intermedios son imaginables.

Existen otros indicios que pueden hacer sospechar la existencia de una situación de colinearidad. Uno de ellos consiste en observar que las varianzas de las estimas de los coeficientes de regresión tienen valores anormalmente altos, disminuyendo drásticamente al eliminar una (o varias) variables regresoras. Igualmente, el encontrar coeficentes de correlación múltiple o de determinación entre una variable regresora y el resto de las otras regresoras

muy elevados. Algunos programas calculan estos coeficientes automáticamente definiendo el valor $1-R^2$ como «tolerancia». De todos modos, puede haber colinearidad sin que estos síntomas sean evidentes.

3.3.2. *Tratamiento*

Supóngase que el examen de los valores propios indica la presencia de una fuerte colinearidad, ¿qué se debe hacer en este caso? La actitud a adoptar debe venir marcada por dos consideraciones:

a) El origen de la colinearidad.

b) La presencia de valores de las regresoras no conocidos con exactitud.

3.3.2.1. Origen de la colinearidad

Sea el ejemplo ya citado del investigador forestal que quiere ajustar el volumen de los árboles (V) en función del diámetro al cuadrado (S) y la altura (H). Si los árboles fueran perfectamente homotéticos, existiría una relación perfecta entre las dos variables regresoras de la forma

$$S = c\ H^2$$

que induciría, sobre el rango o dominio de variación de los valores de las regresoras, una correlación elevada y, por tanto, colinearidad (Figura 3.3). Es muy posible que, en un caso como éste se obtenga la regresión siguiente:

$$V = 70 + 10S - 0{,}1\ H$$

Podría sorprender encontrar un coeficiente negativo ($-0{,}10$) afectando a la altura del árbol, pues, evidentemente, cuanto más alto sea un árbol, mayor será su volumen. Se puede llegar, además, a que la eliminación de la variable dudosa (H) ocasione una pérdida apreciable de ajuste, lo que fuerza a mantener la ecuación con ambas variables en ella. ¿Qué significa, pues, el coeficiente negativo? El error de razonamiento cometido se debe al hecho de que se había interpretado el aumento de H aisladamente y sin tener en cuenta el valor de S. En la realidad esto no es así, ya que si se hace crecer la altura es necesario hacer crecer el diámetro, y así se verifica que

$$\Delta V = 10\,\Delta S - 0{,}10\,\Delta H$$

es positivo, tal como en buena lógica debe ser.

En un caso como el anterior en que la colinearidad viene originada por una relación cuasi-funcional entre las variables regresoras, conviene mantener a todos los regresores en la ecuación si se comprueba que la eliminación de uno de ellos ocasiona una disminución apreciable de la suma de cuadrados de la regresión.

La relación observada entre las regresoras del ejemplo precedente tiene una explicación evidente; sin embargo, puede producirse una situación parecida, aunque de interpretación más sutil, cuando la relación procede del muestreo empleado en la toma de los datos: Se desea predecir la renta de explotaciones agrícolas (R) en función de la potencia de sus tractores (P) y la superficie labrada (L). Una muestra en una región cerealista producirá una fuerte relación entre P y L y, por tanto, una elevada colinearidad. Si la ecuación de regresión calculada es

$$R = 32 + 10L - 3P$$

existe el mismo problema en la interpretación del signo que anteriormente. La diferencia estriba en el hecho de que la relación entre L y P no es de naturaleza intrínseca, sino debido al muestreo. Por ejemplo, si el estudio se hubiera hecho en una región con ganadería intensiva no existiría una rela-

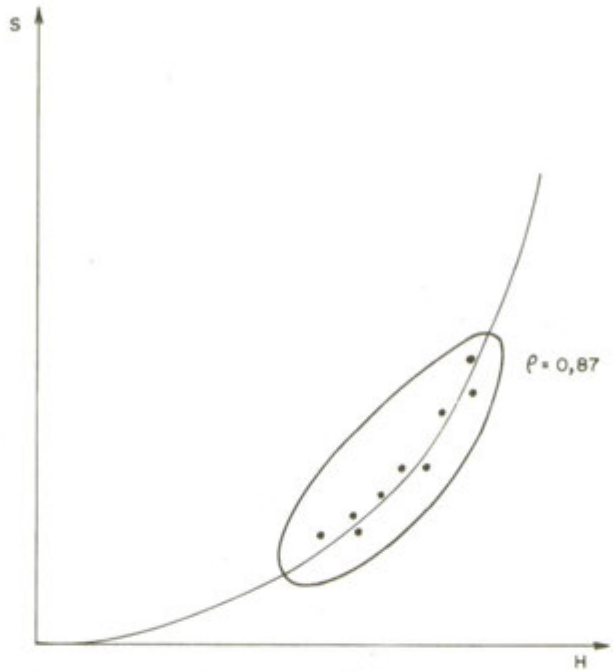


Figura 3.3.—Las relaciones cuasi-funcionales entre las regresoras pueden inducir a colinearidad.

ción tan fuerte entre L y P, puesto que en este caso la cantidad de tierra cultivada es pequeña y no estará ligada a la potencia de los tractores (fig. 3.4). Con este ejemplo simple se ha pretendido exponer un problema importante relacionado con la colinearidad en particular y con regresión en general: no se debe extrapolar la fórmula o modelo de regresión de una población a otra, sino aplicarla solamente a aquella de la cual se obtuvo el modelo. La presencia de la colinearidad puede significar que se está trabajando sobre una población particular y con objetivo deseado concreto. Más grave sería disponer de una muestra sesgada inadvertidamente o que la colinearidad fuera la consecuencia de un número de individuos demasiado reducido con relación al número de variables regresoras empleadas.

3.3.2.2. Exactitud de los valores de las regresoras.

Una vez concluidas las dos primeras etapas, es decir, la colinearidad ha sido detectada y aceptada la explicación de su origen, los pasos posteriores dependerán de la actitud del investigador frente a la colinearidad.

Si se acepta no considerar más que fórmula global (sin importar que los coeficientes de regresión sean inestables), renunciando a interpretar las influencias respectivas de las diversas variables regresoras, el problema ha terminado. Por el contrario, si se insiste en querer estudiar dichas influencias particulares conviene examinar, en el caso de que las regresoras no se conozcan con exactitud, si no hay riesgo de obtener un mal ajuste. Para ello se investigan los ejes principales del análisis de componentes principales normalizado de las regresoras (vectores propios de su matriz de correlaciones tal como se explicó en 3.1.3.1. y 3.1.3.2.). Se calcula la correlación entre la va-

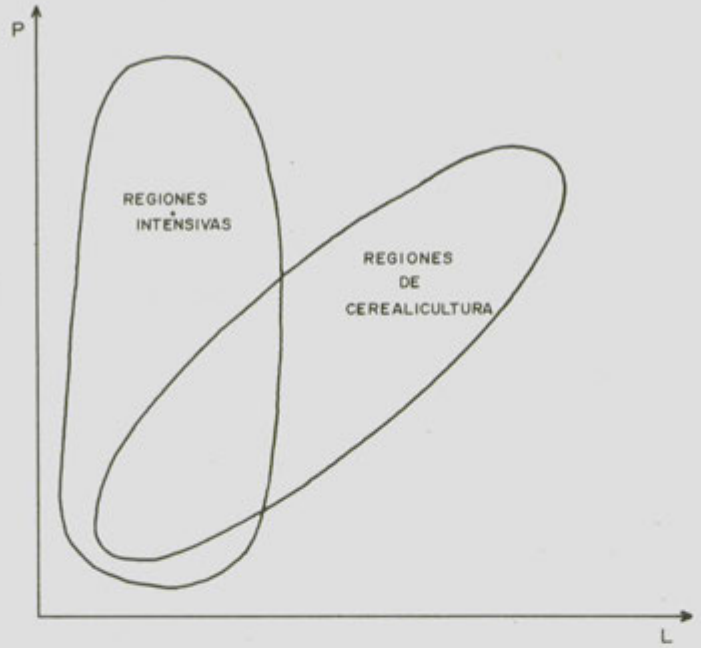


Figura 3.4.—Relación entre regresoras según la población muestreada.

riable dependiente y el primer eje principal; la correlación múltiple con los dos primeros ejes, los tres primeros, etc. Si se juzga que Y está suficientemente correlacionada (por ejemplo, superior a 0,7) con los ejes que absorben la mayor parte de la inercia (varianza), se puede admitir el usar la regresión habitual, porque en este caso, tal como se indicó en 3.2.2., la presencia de la colinearidad es beneficiosa. En el caso contrario no queda más remedio que eliminar variables (aquellas que sean combinación de otras, combinación identificada al estudiar el R^2 o la tolerancia) o emplear técnicas más sofisticadas, como la regresión «ridge» (Hoerl y Kennard, 1970 a, b).

La regresión «ridge» permite establecer una ecuación de regresión de mejor calidad (en un cierto sentido) que la obtenida por medio de la estimación mínimo cuadrática en caso de presencia de fuerte colinearidad.

Las estimaciones obtenidas por cuadrados mínimos eran

$$\hat{\beta} = (X'X)^{-1}X'\ y$$

Si hay colinearidad importante, la matriz (X'X) está mal condicionada, pues su determinante es casi nulo y su inversión puede acarrear problemas numéricos. La regresión «ridge» resuelve esta dificultad reemplazando (X'X) por (X'X+fI) para producir un mejor condicionamiento. En la fórmula, «f» es una constante e I es la matriz identidad. En este caso se obtienen los valores siguientes:

$$\beta^*(f) = (X'X+fI)^{-1}\ X'\ y$$

Naturalmente, las estimaciones halladas son sesgadas y dependen del valor de «f», y para f=0 coinciden con las minimocuadráticas, y si f=∞, β=0.*

El problema consiste en escoger el valor de «f» adecuado, haciendo un compromiso entre el sesgo y la varianza. La solución generalmente empleada es empírica y consiste en calcular β(f) para muchos valores de «f» y elegir aquel f_o que corresponda al comienzo de la zona de estabilización de las estimaciones tal como se aprecia en la figura 3.5.*

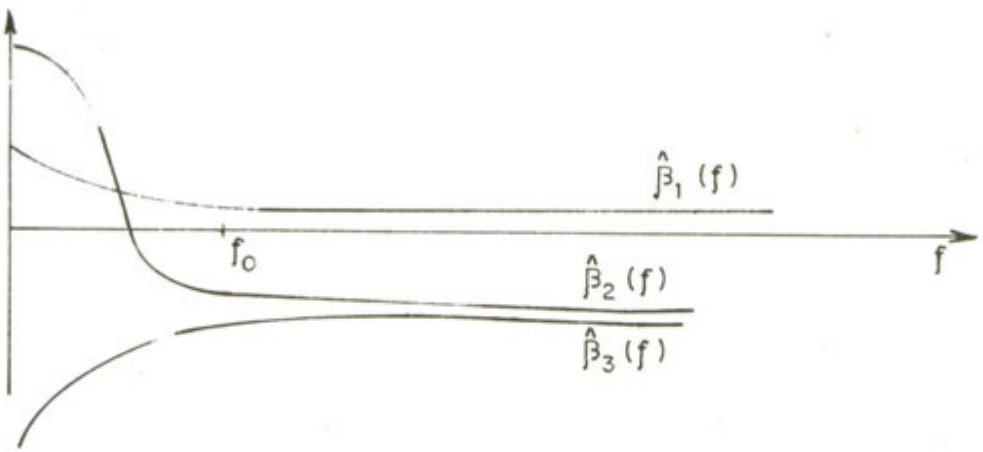


Figura 3.5.—«Ridge trace».

3.3.3. *Recapitulación*

En la figura 3.6 se presenta de manera simplificada el árbol de decisiones propuesto para tratar la presencia de colinearidad. En él se indican simplemente las preguntas y sus respuestas asociadas en los diferentes nudos de cada etapa. La manera de llegar a la respuesta se encuentra descrita a lo largo del presente capítulo.

Como conclusión debe decirse que esta estrategia es vaga y debe adaptarse a cada caso en función de los objetivos y características del estudio.

La investigación de la colinearidad es importante, en definitiva, para determinar algunas características de la población de donde se extrajo la muestra (Marquardt y Snee, 1975), y también, como indican Gunst y Mason (1977), para explicar el posible comportamiento errático de los procedimientos de selección de variables que serán tratados en el capítulo 4. Por ello, es muy conveniente antes de iniciar un estudio de selección de variables revisar la presencia de colinearidad.

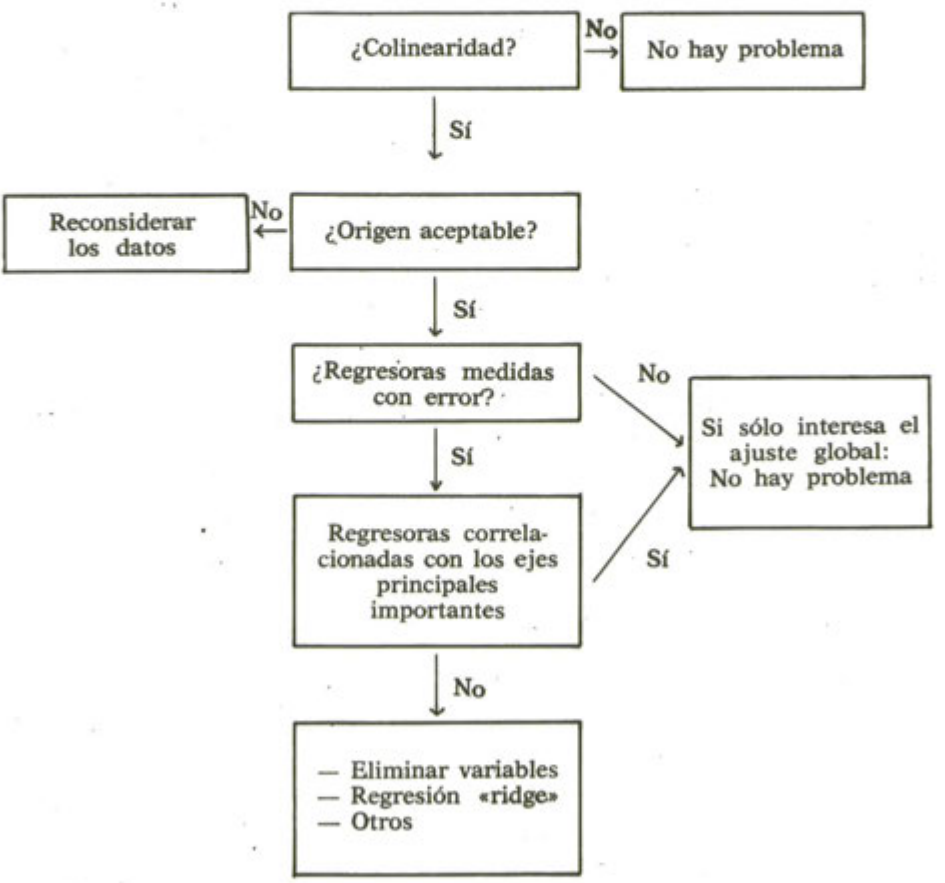


Figura 3.6.—Arbol de decisiones propuesto.

CAPITULO 4

SELECCION DE VARIABLES

4.1. Planteamiento del problema

Se desea establecer una ecuación de regresión lineal de la variable dependiente Y en función de las variables regresoras x_1, x_2, ..., x_k. Tal grupo constituye el conjunto completo de variables entre las que se han de elegir aquéllas que formarán parte de la ecuación buscada, pudiendo estar incluidas en dicho conjunto algunas funciones sencillas de las variables originales, tales como productos cruzados o potencias (este último aspecto puede ser muy interesante en ciertas aplicaciones agrarias; por ejemplo, tarifas y tablas de cubicación, respuestas de los cultivos a abonados).

En la selección de la ecuación buscada entran en juego diversos factores: Por una parte, para hacer la ecuación útil con una finalidad predictiva es necesario considerar un modelo lo más perfecto posible, incluyendo en él tantas variables regresoras con influencia en el fenómeno considerado como sea posible determinar. El disponer de un modelo muy completo puede tener como consecuencia la obtención de predicciones de la variable Y con una gran varianza debido, entre otras causas, a una posible colinearidad que suponga intervalos de confianza para la predicción demasiado amplios y, por tanto, con poco interés práctico. (Más adelante, al tratar el método gráfico de Mallows para comparar ecuaciones de ajuste, se pondrá de manifiesto la afirmación anterior.)

En consecuencia, las consideraciones anteriores sobre la fiabilidad de las predicciones pueden imponer ciertas restricciones en la elección de modelos demasiado complejos.

Por otro lado, si los costes que suponen la inclusión de determinadas variables en el modelo no se ven compensados por el interés de su aportación, en posteriores análisis que se realicen se puede tener en cuenta esta circunstancia y optar por incluir en el modelo solamente aquellas variables que se puedan considerar «rentables».

No existe un procedimiento estadístico único para resolver el problema apuntado anteriormente, como se verá más adelante. Por esto, el conocimiento que tenga el investigador acerca del fenómeno estudiado debe ser una ayuda necesaria y de gran valor en cualquiera de los métodos que se describen.

Todos los métodos que se citan son de uso corriente en la práctica estadística y su aplicación a un mismo problema no tiene por qué conducir necesariamente a la misma solución.

También es necesario hacer notar que las distintas pruebas que se realizan en un determinado método de selección de variables no son independientes estadísticamente y por tanto no tiene sentido hablar de un error α global para el conjunto de todas las pruebas o para el método en cuestión.

El proceso de selección de variables debe ser posterior a un estudio minucioso de la colinearidad, pues si ésta es importante puede conducir a resultados erróneos (ver capítulo 3).

Una amplia descripción de los métodos de selección se encuentra en Draper y Smith (1981), Hocking (1976) y Thompson (1978 a, b).

4.2. Medida de la calidad global de un modelo de regresión

Sea el modelo de regresión múltiple

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

anteriormente se vio que:

$$\underset{\text{Suma de cuadrados total}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \underset{\text{S. C. debida a la regresión}}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2} + \underset{\text{S. C. del residuo}}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Una interpretación geométrica de esta descomposición se encuentra en Cailliez y Pages (1971), Wonnacott y Wonnacott (1981) y en el capítulo 7.

Construyendo la tabla del análisis de la varianza de la regresión:

Fuentes	g.l.	S.C.	C.M.
Debido a regresión	k	$\hat{y}'\hat{y} - n\bar{y}^2$ (1)	$(\hat{y}'\hat{y} - n\bar{y}^2)/k$
Desviación de la regresión	n-k-1	$y'y - \hat{y}'\hat{y}$ (2)	$(y'y - \hat{y}'\hat{y})/n-k-1$
Total	n-1	$y'y - n\bar{y}^2$ (3)	$(y'y - n\bar{y}^2)/n-1$

Siendo $\hat{y}'\hat{y} = b'X'y$

En estas condiciones se puede afirmar que una ecuación de regresión será tanto mejor cuanto menor sea la suma de los residuos al cuadrado

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

o también cuanto mayor sea $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Una medida de la calidad global de un modelo de regresión puede ser la cantidad

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

que se define como «coeficiente de determinación» y cuyo valor varía entre 0 y 1 (ver 6.3 para una discusión más detallada de su utilidad).

La raíz cuadrada del coeficiente de determinación se denomina coeficiente de correlación múltiple «R» entre la variable dependiente Y, y el conjunto de variables regresoras x_j para $j=1, \ldots, k$.

La cantidad R también representa el coeficiente de correlación entre la variable dependiente Y y su estima y a través del modelo de regresión.

Otra forma de expresar la calidad de la regresión es a través del estadístico F_c, que se define como:

$$F_c = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / k}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n-k-1} = \frac{(\hat{y}'\hat{y} - n\bar{y}^2)/k}{(y'y - \hat{y}'\hat{y})/n-k-1} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$$

Este valor F_c permite, con ayuda de una tabla F de Snedecor, juzgar la significación de la regresión; es decir, apreciar si la estimación de Y por

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

es significativamente mejor que la estimación de Y por $y=b_0=\bar{y}$.

4.3. Medida de la significación de un subconjutno de variables regresoras

Sea el modelo de regresión múltiple

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Las estimas de Y obtenidas a través del modelo completo con k variables se denotan por $\hat{y}_k$.

Si se estima Y a través de un modelo más sencillo en el que solamente estén incluidas p variable de las k posibles (que, evidentemente, no tienen por qué ser las p primeras)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \qquad p < k$$

la estima de Y obtenida a través del modelo reducido con solo p variables se denotará por $\hat{y}_p$.

En estas condiciones la variabilidad individual puede descomponerse como sigue:

$$(y_i - \bar{y}) = (\hat{y}_{p_i} - \bar{y}) + (y_i - \hat{y}_{p_i}) = (\hat{y}_{p_i} - \bar{y}) + (\hat{y}_{k_i} - \hat{y}_{p_i}) + (y_i - \hat{y}_{k_i})$$

Elevando al cuadrado y sumando para todas las observaciones

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_{p_i}-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_{p_i})^2 =$$

$$= \sum_{i=1}^{n}(\hat{y}_{p_i}-\bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_{k_i}-\hat{y}_{p_i})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_{k_i})^2$$

puesto que todos los dobles productos son nulos.

El término $\sum_{i=1}^{n}(\hat{y}_{p_i}-\bar{y})^2$ mide la discrepancia existente entre los valores estimados a través del modelo con p variables y un modelo patrón $Y=\bar{y}$.

El término $\sum_{i=1}^{n}(\hat{y}_{k_i}-\hat{y}_{p_i})^2$ mide la discrepancia entre los valores estimados a través del modelo completo de k variables y el submodelo que contiene p variables.

El término $\sum_{i=1}^{n}(y_i-\hat{y}_{k_i})^2$ mide la discrepancia existente entre los valores observados y los valores estimados a través del modelo completo con k variables.

¿Cuál será el interés de incluir en el modelo con p variables aquellas k-p variables que en este momento no están presentes?

Se define el estadístico F* de la siguiente manera:

$$F^* = \frac{\sum_{i=1}^{n}(\hat{y}_{k_i}-\hat{y}_{p_i})^2/(k-p)}{\sum_{i=1}^{n}(y_i-\hat{y}_{k_i})^2/(n-k-1)} = \frac{(\hat{y}'_k\hat{y}_k - \hat{y}'_p\hat{y}_p)/(k-p)}{(y'y - \hat{y}'_k\hat{y}_k)/(n-k-1)}$$

Dicho estadístico F* se emplea para medir el «poder explicativo» de las k-p variables ausentes del modelo cuando las p restantes están presentes en él (concepto similar al expuesto en 2.2). Cuanto mayor sea el valor de F*, es decir, cuanto mayor sea el numerador frente al denominador, mayor será el interés de incluir dichas k-p variables en el modelo.

Teniendo en cuenta que

$$R_k^2 = \frac{\sum_{i=1}^{n}(\hat{y}_{k_i}-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad ; \quad R_p^2 = \frac{\sum_{i=1}^{n}(\hat{y}_{p_i}-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

se puede expresar el estadístico F* de otra forma:

$$F^* = \frac{n-k-1}{k-p} \quad \frac{R_k^2 - R_p^2}{1 - R_k^2}$$

Empleando la tabla F de Snedecor es posible juzgar si las k-p variables regresoras no incluidas en el modelo tienen todavía un poder explicativo significativo, teniendo en cuenta que hay p variables regresoras presentes en la ecuación de regresión.

Hay dos casos particulares extremos:

a) $p=k-1$, es decir, el número de variables que están presentes en el modelo es $k-1$, y solamente hay una variable, la x_k, que está ausente de él. En estas condiciones el estadístico F* está midiendo el poder explicativo de dicha variable x_k, teniendo presente que las restantes $k-1$ variables están incluidas en el modelo de regresión. (El contraste F*, en este caso, es equivalente a la t de Student.)

b) $p=0$, en este caso el submodelo no existe y equivale al caso del contraste del modelo completo incluyendo las k variables.

El estadístico F* será el que se emplee posteriormente al ir desarrollando los distintos métodos de selección de variables. Otras veces, el criterio de comparación entre modelos se basa en el cálculo del coeficiente de correlación parcial. Este enfoque será presentado en el ejemplo 7.5 y la relación entre ambos criterios se incluirá en la representación geométrica.

4.4. El método de todas las regresiones posibles

Este método de selección de variables consiste en calcular todas las ecuaciones de regresión que son posibles al combinar las distintas variables y posteriormente seleccionar la ecuación óptima.

El procedimiento sólo es fácilmente realizable cuando se tiene acceso a un computador de alta velocidad, y en consecuencia sólo ha empezado a emplearse desde que se dispone de computadores rápidos.

Sea el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

En estas condiciones cada variable regresora x_j puede estar o no incluida en la ecuación, y esto es cierto para cada x_j, variando $j=1, \ldots, k$.

El número de posibles ecuaciones de regresión es 2^k-1. Si $k=10$, $2^k-1=2^{10}-1=1023$. Esto da una idea de la magnitud que puede alcanzar el problema cuando k tiene valores relativamente grandes.

El método se ilustrará con el ejemplo siguiente:

Sea la variable dependiente Y, y las variables regresoras x_1, x_2, x_3; el número de ecuaciones posibles es de $2^3-1=7$, y son las siguientes:

(*Nota*: Aunque se representen los distintos modelos de regresión con las mismas letras β y ε no quiere decir que tales valores coincidan en las distintas regresiones.)

i) p=1; subconjuntos que contienen una variable:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$Y = \beta_0 + \beta_2 x_2 + \varepsilon$$

$$Y = \beta_0 + \beta_3 x_3 + \varepsilon$$

ii) p=2; subconjuntos que contienen dos variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$$

$$Y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

iii) p=3; subconjunto que contiene las tres variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_3 + \varepsilon$$

Para explicar el desarrollo de los distintos métodos de selección de variables, se han empleado los datos que muestra la tabla 4.1. Dicha tabla de datos se ha creado por simulación y en ella figuran 21 observaciones de una variable dependiente Y y cuatro variables regresoras x_1, x_2, x_3, x_4, cuyos estadísticos elementales y matrices de varianzas-covarianzas y correlación se incluyen allí.

TABLA 4.1

Datos y estadísticos elementales

Número de observación	x_1	x_2	x_3	x_4	y
1	5,6500	54,1100	16,9900	21,4600	98,2200
2	5,0000	48,5900	13,7000	31,6700	91,6800
3	1,5900	9,2000	16,7100	31,4600	58,0900
4	0,9200	40,1100	20,9600	35,7700	82,0200
5	4,0000	64,4300	16,4500	11,2800	100,0200
6	6,8600	71,1900	11,4700	7,7800	105,3700
7	1,4200	40,9500	18,9300	34,8000	82,9600
8	4,3900	54,1000	16,4200	21,7300	94,1900
9	8,8800	67,0900	15,6400	6,3300	110,6800
10	3,7200	59,3600	13,5000	21,0400	95,8100
11	17,8800	38,8700	5,9500	34,8100	103,9000
12	18,1200	69,7100	6,1000	4,0800	127,0400
13	10,4900	34,6100	10,4900	43,8100	91,8700
14	7,9100	33,7400	7,5300	48,6700	86,8900
15	3,5000	58,0700	12,7200	23,0500	94,5800
16	2,5200	60,0200	13,9700	20,8400	93,7200
17	0,4000	68,0400	22,6400	5,1100	97,8200
18	9,2900	38,8200	15,6600	34,3500	91,4100
19	1,9500	27,8700	12,0200	56,7200	70,9000
20	0,3800	36,5900	24,9200	35,0200	75,9900
21	8,2800	40,3500	16,6900	30,5200	93,6000

TABLA 4.1

Datos y estadísticos elementales (continuación)

Variable	*Media*	*Desviación típica*	$S/\bar{X}$	*Mínimo*	*Máximo*
x_1	5,86429	5,06635	0,86393	0,38000	18,12000
x_2	48,37238	16,20737	0,33505	9,20000	71,19000
x_3	14,73619	4,94863	0,33581	5,95000	24,92000
x_4	28,53333	17,35770	0,60833	4,08000	70,46000
y	92,70286	14,34299	0,15472	58,09000	127,04000

MATRIZ DE VARIANZAS-COVARIANZAS

	x_1	x_2	x_3	x_4	y
x_1	25,6679				
x_2	11,7116	262,6789			
x_3	−18,9990	−3,9962	24,4890		
x_4	−17,0201	−274,6969	−3,6102	301,2898	
y	48,3329	191,8062	−30,0066	−211,0289	205,7212

MATRIZ DE CORRELACIONES

	x_1	x_2	x_3	x_4	y
x_1	1,0000				
x_2	0,1426	1,0000			
x_3	−0,7578	−0,0498	1,0000		
x_4	−0,1935	−0,9764	−0,0420	1,0000	
y	0,6651	0,8251	−0,4228	−0,8476	1,0000

El primer paso del método consiste en calcular todas las ecuaciones de regresión posibles, que en este caso serán $2^4-1=15$.

En el paso siguiente se dividen las ecuaciones de regresión en cuatro grupos. El primero, con las ecuaciones que tienen una variable regresora; el segundo, con las ecuaciones que tienen dos variables, y así sucesivamente. Seguidamente se ordenan los grupos internamente en orden decreciente según el valor del estadístico F_c, tal como se muestra en las tablas 4.2 y 4.3. Esta ordenación es la misma que la resultante de emplear el valor de R o de R^2 debido a la relación que existe entre R, R^2 y F.

TABLA 4.2

Ajuste de diversos modelos de regresión

Variables regresoras	*R*	R^2	*Media de residuos al cuadrado*	F_c	*C.P.*
x_4	0,8476	0,7185	60,961	48,493	439,81
x_2	0,8251	0,6808	69,122	40,525	500,96
x_1	0,6651	0,4424	120,747	15,075	887,81
x_3	0,4228	0,1787	177,846	4,135	1315,68
x_1, x_2	0,9933	0,9867	3,034	668,948	6,54
x_1, x_4	0,9896	0,9793	4,722	426,632	18,52
x_3, x_4	0,9638	0,9290	16,234	117,720	100,25
x_2, x_3	0,9093	0,8268	39,586	42,969	283,10
x_2, x_4	0,8477	0,7186	64,315	22,986	441,58
x_1, x_3	0,6767	0,4579	123,909	7,603	864,64
x_1, x_2, x_3	0,9950	0,9901	2,392	567,709	3,04
x_1, x_2, x_4	0,9949	0,9898	2,460	551,733	3,50
x_1, x_3, x_4	0,9942	0,9885	2,784	486,891	5,76
x_2, x_3, x_4	0,9901	0,9802	4,786	280,890	19,09
x_1, x_2, x_3, x_4 ...	0,9951	0,9901	2,536	401,674	5,00

TABLA 4.3

Coeficientes de regresión estimados par diversos modelos de regresión

Variables regresoras	β_0	β_1	β_2	β_3	β_4
x_1	81,660	1,883			
x_2	57,382		0,730		
x_3	110,759			−1,225	
x_4	112,688				−0,700
x_1, x_2	51,516	1,582	0,660		
x_1, x_3	71,105	2,293		0,553	
x_1, x_4	101,670	1,474			−0,617
x_2, x_3	74,539		0,713	−1,109	
x_2, x_4	116,317		−0,049		−0,745
x_3, x_4	132,756			−1,331	−0,716
x_1, x_2, x_3	46,776	1,776	0,655	0,260	
x_1, x_2, x_4	69,141	1,540	0,431		−0,221
x_1, x_3, x_4	110,961	1,128		−0,445	−0,642
x_2, x_3, x_4	213,529		−1,024	−1,636	−1,654
x_1, x_2, x_3, x_4	38,194	1,863	0,742	0,354	0,086

El tercer paso consiste en seleccionar de cada grupo aquella ecuación que tiene el mayor valor de la F_c y estudiar su evolución de un grupo a otros.

Para este ejemplo las ecuaciones seleccionadas serán las siguientes:

Número de variables de la ecuación	*Ecuación de regresión*	R	R^2	F_c	P
1	$y=f(x_4)$	0,8476	0,7185	48,493	$0,17\times10^{-4}$
2	$y=f(x_1, x_3)$	0,9933	0,9867	668,948	$0,18\times10^{-7}$
3	$y=f(x_1, x_2, x_3)$	0,9950	0,9901	567,709	$0,26\times10^{-7}$
4	$y=f(x_1, x_2, x_3, x_4)$	0,9951	0,9901	401,674	$0,52\times10^{-7}$

En el cuarto paso se ha de tomar la decisión sobre qué ecuación escoger entre las cuatro seleccionadas, procurando conciliar el compromiso entre el número de variables presente en el modelo y las eventuales ganancias en el valor de F_c que pueden obtenerse al ir incorporando variables al modelo.

Aunque el modelo con dos variables proporciona el mayor valor para la F, no necesariamente significa que es el «mejor», puesto que no es válida la comparación de F basadas en grados de libertad diferentes. En vez de este valor F, un criterio mejor sería el área que deja a su derecha; es decir, el nivel de significación. En el caso presente, sin embargo, ambos criterios coinciden en señalar el modelo $f(x_1, x_2)$ como el mejor.

También puede estudiarse la evolución del valor que toma el coeficiente de determinación R^2 al ir pasando de un grupo a otro. Puede ocurrir que la adición de una variable más a un determinado conjunto de éstas no produzca un aumento significativo en el valor de R^2. En este caso, sería prudente retener el conjunto más sencillo. En el ejemplo, el paso del modelo $\hat{y}=f(x_1, x_2)$ al modelo $\hat{y}=f(x_1, x_3, x_3)$ proporciona un incremento en el valor de R^2 del 0,3 por 100, lo cual parece aconsejar el retener el modelo $\hat{y}=f(x_1, x_2)$.

A la vista de la tabla 4.3, la ecuación de regresión seleccionada es:

$$y = 51,516 + 1,582x_1 + 0,66x_2$$

El examen de todas las regresiones posibles no proporciona una respuesta definitiva al problema. Otras informaciones, tales como un conocimiento profundo del fenómeno estudiado que tenga presente el papel que desempeña cada una de las variables, deben tenerse muy en cuenta antes de tomar una decisión.

4.5. Eliminación descendente de las variables

El método de eliminación descendente es más eficiente desde el punto de vista del cálculo que el método anterior, pues no es preciso examinar todas las regresiones posibles sino solamente la «mejor» regresión, que contiene un cierto número de variables.

Los pasos básicos de este procedimiento son:

1. Cálculo de la ecuación de regresión que contiene a todas las variables.

En el ejemplo se calculará la ecuación de regresión

$$Y = \beta_0+\beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon$$

que según se ve en la tabla 4.2 tiene un valor de $R^2=0,9901$ y de $F_c=401,674$.

2. Cálculo del valor explicativo de cada variable, teniendo en cuenta que el resto de las variables están incluidas en el modelo.

Para ello se obtiene el valor del estadístico F* para la eliminación sucesiva de las variables x_1, x_2, x_3, x_4.

$$F^* = \frac{n-k-1}{k-p} \quad \frac{R_k^2 - R_p^2}{1 - R_k}$$

Realizando los cálculos a la vista de la tabla 4.2 se obtiene:

$$F^*(x_1) = \frac{21-4-1}{4-3} \quad \frac{0,9901 - 0,9802}{1 - 0,9901} = 16,00$$

$$F^*(x_2) = \frac{21-4-1}{4-3} \quad \frac{0,9901 - 0,9885}{1 - 0,9901} = 2,58$$

$$F^*(x_3) = \frac{21-4-1}{4-3} \quad \frac{0,9901 - 0,9898}{1 - 0,9901} = 0,48$$

$$F^*(x_4) = \frac{21-4-1}{4-3} \quad \frac{0,9901 - 0,9901}{1 - 0,9901} = 0,00$$

La variable que tenga el menor valor de F* es la que posee menor valor explicativo, teniendo en cuenta que las demás están presentes. En consecuencia, es la primera variable que se elimina del modelo. En el ejemplo se trata de la variable x_4. El modelo completo queda constituido ahora por la ecuación

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

3. Se calcula el valor del estadístico F* para las variables que aún permanecen en el modelo operando de manera análoga al paso anterior, si bien la base de comparación sigue siendo el modelo completo y no el que previamente se había elegido con solamente tres variables (x_1, x_2 y x_3). Esta forma de actuar es conveniente para que el error α global no se vea demasiado afectado.

Así, pues,

$$F^*(x_1) = \frac{21-4-1}{4-2} \quad \frac{0,9901 - 0,8268}{1 - 0.9901} = 131,96$$

$$F^*(x_2) = \frac{21-4-1}{4-2} \quad \frac{0,9901 - 0,4579}{1 - 0,9901} = 430,06$$

$$F^*(x_3) = \frac{21-4-1}{4-2} \quad \frac{0,9901 - 0,9867}{1 - 0,9901} = 2,59$$

Por tanto, la variable x_3 sería eliminada, con lo que el modelo quedaría

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Calculando de nuevo los valores de F*

$$F^*(x_1) = \frac{21-4-1}{4-1} \quad \frac{0,9901 - 0,6808}{1 - 0.9901} = 166,63$$

$$F^*(x_2) = \frac{21-4-1}{4-1} \quad \frac{0,9901 - 0,4424}{1 - 0,9901} = 295,06$$

En este paso se eliminaría la x_1, quedando el modelo reducido a

$$Y = \beta_0 + \beta_2 x_2 + \varepsilon$$

La ordenación de las variables de mayor a menor importancia es x_2, x_1, x_3 x_4. Esta ordenación no coincide con la obtenida a través del método de todas las regresiones posibles que situaba a la x_4 como la más importante. La razón es porque ambos métodos siguen caminos diferentes y por tanto no son comparables. El orden de importancia para este método considera que el resto de las variables están presentes en la ecuación; por tanto se trata de la importancia relativa de cada variable.

Si con un error α determinado se selecciona previamente un determinado nivel límite F_0 para el valor de F*, el proceso de eliminación de variables se detendrá cuando haya que eliminar una variable x_j, que es «significativa» en el sentido de F*, es decir, que $F^*(x_j \geq F_0$.

En el ejemplo: si antes de eliminar x_1 comparamos $F^*(x_1)=166,6$ con $F_0=F(3;16;0,05)=3,24$, la decisión será no eliminar x_1, deteniéndose el proceso. En este caso se considerará como mejor ecuación de regresión aquélla que precede a la eliminación de una variable significativa.

La mejor ecuación de regresión será, por tanto,

$$y=51,516+1,582x_1+0,66x_2$$

El esfuerzo de cálculo necesario para llegar a determinar la mejor ecuación de regresión por este método supone la determinación de $\frac{k(k+1)}{2}$ ecuaciones de regresión. Esto implica una reducción considerable en comparación con el método de todas las regresiones que implicaba el cálculo de 2^k-1.

4.6. Introducción ascendente de las variables

El método de introdución ascendente de las variables persigue un objetivo similar al anterior, pero trabajando en sentido inverso.

La técnica consiste en ir introduciendo una a una las variables regresoras en el modelo hasta que se obtenga una ecuación de regresión considerada como satisfactoria o hasta que se introduzcan todas las variables.

Los pasos a seguir son:

1. Determinación de la ecuación de regresión, que, considerando sólo una variable, proporciona el mejor ajuste. Siguiendo con el mismo ejemplo, hay que calcular las ecuaciones

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$Y = \beta_0 + \beta_2 x_2 + \varepsilon$$
$$Y = \beta_0 + \beta_3 x_3 + \varepsilon$$
$$Y = \beta_0 + \beta_4 x_4 + \varepsilon$$

Según la tabla 4.2, la ecuación que proporciona el mejor ajuste es $y=f(x_4)$ con un valor de $R^2=0,7185$ y $F_c=48,493$. El modelo es ahora

$$Y = \beta_0 + \beta_4 x_4 + \varepsilon$$

2. Incorporación al modelo anterior de aquella variable que proporcione el modelo con mejor ajuste. En el ejemplo se deben ensayar las regresiones

$Y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \varepsilon$; $F_c = 426,632$

$Y = \beta_0 + \beta_4 x_4 + \beta_2 x_2 + \varepsilon$; $F_c = 22,986$

$Y = \beta_0 + \beta_4 x_4 + \beta_3 x_3 + \varepsilon$; $F_c = 117,720$

De estos resultados se deduce que la siguiente variable a introducir es x_1.

El valor de F* para x_1 es:

$$F^*(x_1) = \frac{21-2-1}{2-1} \frac{0,9793 - 0,7185}{1 - 0,9793} = 226,78$$

El modelo es:

$$Y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \varepsilon$$

Reiterando el paso anterior, las ecuaciones a ensayar serán:

$Y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$; $F_c = 551,733$

$Y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$; $F_c = 486,891$

La siguiente variable a introducir es x_2.

El valor de F* para x_2 es

$$F^*(x_2) = \frac{21-3-1}{3-2} \frac{0,9898 - 0,9793}{1 - 0,9893} = 17,49$$

El modelo ahora es

$$Y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Si se introduce la variable que queda, x_3, se obtiene la ecuación

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$; $F_c = 401,674$

El valor de F* para x_3 es

$$F^*(x_3) = \frac{21-4-1}{4-3} \frac{0,9901 - 0,9898}{1 - 0,9901} = 0,4848$$

Finalmente, el modelo completo es

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

La ordenación de las variables de mayor a menor importancia es x_4, x_1, x_2, x_3, que no coincide con ninguna de las anteriores.

Conocido el valor del estadístico F* para las variables que se han ido introduciendo en el modelo, es posible contrastar la significación de dicha variable teniendo en cuenta el efecto de las variables presentes en el modelo.

El proceso de introducción de variables en el modelo puede detenerse cuando el valor de F* para la variable que se ha introducido en último lugar no sea significativo a un nivel que previamente se determine.

En el ejemplo, la variable x_3 pudiera no introducirse, pues fijado el nivel

$F(1;16;0,05)=4,49$, $F^*(x_3)=0,4848<1$.

En estas condiciones, la mejor ecuación de regresión sería

$$\hat{y} = 69,141 + 1,54x_1 + 0,431x_2 - 0,221x_4$$

El método de introducción ascendente de las variables supone un esfuerzo de cálculo inferior a los métodos anteriores, pues evita manejar en cada fase del proceso un número de variables superior a las necesarias debido a la mejora progresiva de la ecuación de regresión. Un inconveniente del método es que no se investiga el efecto que puede producir la introducción de una nueva variable en el modelo en el papel que desempeñan aquellas variables que se habían introducido en fases anteriores. Este inconveniente se subsana en el método de regresión progresiva («stepwise regresssion»). Otro inconveniente del método es que en el proceso no se compara con la mejor estima del error experimental σ^2 a través del modelo completo, sino con estimaciones obtenidas a partir de modelos reducidos que, por tanto, pueden ser erróneos.

4.7. Regresión «stepwise»

Aunque los dos métodos anteriores actúan «paso a paso» («stepwise»), este nombre se suele reservar al método que se presenta a continuación, propuesto originariamente por Efroymson (1960).

Este método es un perfeccionamiento del anterior. La mejora consiste en reconsiderar en cada fase de proceso la inclusión o exclusión de aquellas variables que se habían introducido previamente.

Una variable que fue el mejor candidato para ser incluida en el modelo en una fase anterior puede resultar superflua en una fase posterior debido a las relaciones existentes entre dicha variable y aquellas otras que se encuentran actualmente en el modelo.

Para investigar este supuesto se calcula en cada fase del proceso el valor del estadístico F* de todas las variables presentes en el modelo después de introducir una. Estos valores de F* se comparan con un valor seleccionado previamente F_0.

Este procedimiento proporciona una valoración sobre la contribución que aporta cada variable como si fuera la que se ha introducido en último lugar, independientemente de su orden de entrada en el modelo.

Toda variable que proporcione una contribución no significativa se elimina del modelo.

El proceso continúa hasta que se alcanza una fase en la que ninguna variable puede ser introducida y ninguna eliminada.

Siguiendo con el mismo ejemplo, los pasos son los siguientes:

1. Introducción en el modelo de la variable x_4, pues proporciona el mayor valor de $F_c=48{,}493$ dentro de los posibles modelos con una variable.

2. Introducción en el modelo de la variable x_1 como se había visto en el método anterior.

Cálculo de los estadísticos F* para las dos variables presentes en el modelo hasta el momento: x_4 y x_1.

$$F^*(x_4) = \frac{21-2-1}{2-1} \; \frac{0{,}9793 - 0{,}4424}{1 - 0{,}9793} = 466{,}86$$

$$F^*(x_1) = \frac{21-2-1}{2-1} \; \frac{0{,}9793 - 0{,}7185}{1 - 0{,}9793} = 226{,}78$$

Si se selecciona como valor $F_0=(1;\ 18;\ 0.05)=4{,}41$, las variables x_4 y x_1 permanecen en el modelo.

3. Introducción al modelo de la variable x_2, pues dado que x_4 y x_1 están presentes, la ecuación $y=f(x_1, x_2, x_4)$ es la que proporciona un mayor valor de $F_c=551{,}73$.

Cálculo de los estadísticos F* para las tres variables presentes en el modelo hasta el momento: x_4, x_1 y x_2.

$$F^*(x_4) = \frac{21-3-1}{3-2} \; \frac{0{,}9898 - 0{,}9867}{1 - 0{,}9898} = 5{,}16$$

$$F^*(x_1) = \frac{21-3-1}{3-2} \; \frac{0{,}9898 - 0{,}7186}{1 - 0{,}9898} = 451{,}99$$

$$F^*(x_2) = \frac{21-3-1}{3-2} \; \frac{0{,}9898 - 0{,}9793}{1 - 0{,}9898} = 17{,}49$$

Se utiliza como valor $F_0=F(1;17;0.05)=4{,}45$, ninguna variable puede ser eliminada.

4. Introducción en el modelo de la variable que falta por incluir x_3.

Cálculo del estadístico F* para la variable x_3.

$$F^*(x_3) = \frac{21-4-1}{4-3} \; \frac{0{,}9901 - 0{,}9898}{1 - 0{,}9901} = 0{,}48$$

Si se selecciona como valor $F=(1;16;0.05)=4{,}49$, resulta que la variable x_3 no debe ser incluida en el modelo.

La ecuación de regresión definitiva es:

$$y=69{,}141+1{,}54x_1+0{,}431x_2-0{,}221x_4$$

El procedimiento de regresión «stepwise» es el que normalmente se emplea al utilizar programas estándar, aunque debido a que es un método todavía menos riguroso estadísticamente que los anteriores, no es conveniente abusar de él.

Como en todos los métodos expuestos, se requiere un conocimiento del papel desempeñado por cada variable en el fenómeno estudiado, lo cual nos permitirá hacer un juicio objetivo y seleccionar un buen subconjunto inicial de variables.

4.8. Variantes de los métodos anteriores

Los tres últimos procedimientos expuestos no seleccionan necesariamente el mejor modelo, aunque normalmente sí llegan a un modelo aceptable. A efectos prácticos, se pueden considerar algunas variantes:

a) En algunos casos, debido al gran interés biológico que puedan presentar ciertas variables, puede forzarse la introducción de éstas en el modelo inicial.

b) La investigación de resultados intermedios y la mayor importancia biológica de una variable en comparación con otra, puede aconsejar su inclusión a pesar de que su significación estadística sea inferior.

Por ejemplo, si los valores del estadístico F* para las variables x_1 y x_2 fueran $F^*(x_1)=2{,}33$ y $F^*(x_2)=2{,}31$, pero la variable x_2 tiene una mayor significación en el problema estudiado, será más conveniente introducir, o no eliminar x_2 en vez de la x_1.

c) También se pueden idear procedimientos combinados para intentar mejorar la selección del modelo.

Algunos de éstos son los siguientes:

a) Realizar un procedimiento «stepwise» con unos niveles altos para la aceptación y el rechazo. Cuando el procedimiento de selección ha terminado, supongamos que con un modelo que incluye p variables, se calculan todas las ecuaciones de regresión que tienen p variables de las k posibles seleccionando la combinación que proporciona la mejor ecuación de regresión. Una vez seleccionada la ecuación, se inicia de nuevo un procedimiento «stepwise» a partir de tal ecuación. Este procedimiento recibe el nombre de SWAP en el paquete de programas de la serie BMDP elaborado por la Universidad de California en Los Angeles (ver Anejo 5 para una descripción de estos programas en relación con el tema de regresión).

b) Otro procedimiento consiste en realizar el método «stepwise» con unos niveles superiores de aceptación y rechazo (es decir, un F_0 menor, lo cual fuerza al programa a aceptar algunas variables más de las que se hubieran aceptado con unos niveles más estrictos).

Este método permite la investigación de variables adicionales a las seleccionadas en un procedimiento «stepwise» normal y puede conducir a la obtención de un modelo diferente.

4.9. El criterio Cp

4.9.1. *Obtención del estadístico C_p*

Cuando se considera un gran número de ecuaciones de regresión alternativas, es necesario elegir un criterio de bondad del ajuste para caracterizar cada ecuación.

Una de estas medidas es la del «error cuadrático medio total», descrita inicialmente por MALLOWS (1964) y subsiguientemente considerado por DANIEL y WOOD (1980) y MALLOWS (1973) entre otros.

Esta medida, denominada C_p, tiene en cuenta la suma de las desviaciones al cuadrado respecto al modelo completo más el cuadrado de los errores aleatorios en Y, para el conjunto de los n datos.

El estadístico C_p es una función sencilla de la suma de los residuos al cuadrado de cada ecuación de regresión.

El error cuadrático medio total (desviación respecto al modelo verdadero más error aleatorio) para un conjunto de n datos empleando una ecuación de ajuste con p términos es:

$$\sum_{i=1}^{n}\left[E(\hat{y}_{pi}) - \sum_{j=0}^{k} b_j x_{ij}\right]^2 + \sum_{i=1}^{n} E\left[\hat{y}_{pi} - E(\hat{y}_{pi})\right]^2$$

Desviación del modelo — Error aleatorio

siendo:

$\hat{y}_{pi} = \sum_{j=0}^{p} b_j x_{ij}$ — Estimación de y_i con el modelo de p variables incluyendo entre las p la variable x_0 que siempre vale 1.

$E(\hat{y}_{pi})$ — Valor esperado de $\hat{y}_i$ con el modelo de p variables.

$\sum_{j=0}^{k} b_j x_{ij}$ — Valor estimado de y_i con el modelo completo de k variables, incluyendo entre las k la variable x_0 que siempre vale 1.

En la figura 4.1 se realiza la visualización de ambos términos, en la cual, la cantidad (a) representa la desviación del modelo reducido ($y=b_0$) respecto al modelo completo o verdadero ($y=b_0+b_1x_1$) y la cantidad (b) representa el error aleatorio o desviación de $\hat{y}_{pi}$ respecto de $E(\hat{y}_{pi})$.

Representando simbólicamente a

$$\sum_{i=1}^{n}\left[E(\hat{y}_{pi}) - \sum_{j=0}^{k} b_j x_{ij}\right]^2$$

por SCD_p (suma de cuadrados de las desviaciones respecto al modelo), y al segundo término del error cuadrático medio total por $\sum_{i=1}^{n} var\,(y_{pi})$, *se define la variable Γ_p como el error cuadrático medio total dividido por σ^2, es decir:*

$$\Gamma_p = \frac{SCD_p}{\sigma^2} + \frac{\sum_{i=1}^{n} var(\hat{y}_{pi})}{\sigma^2}$$

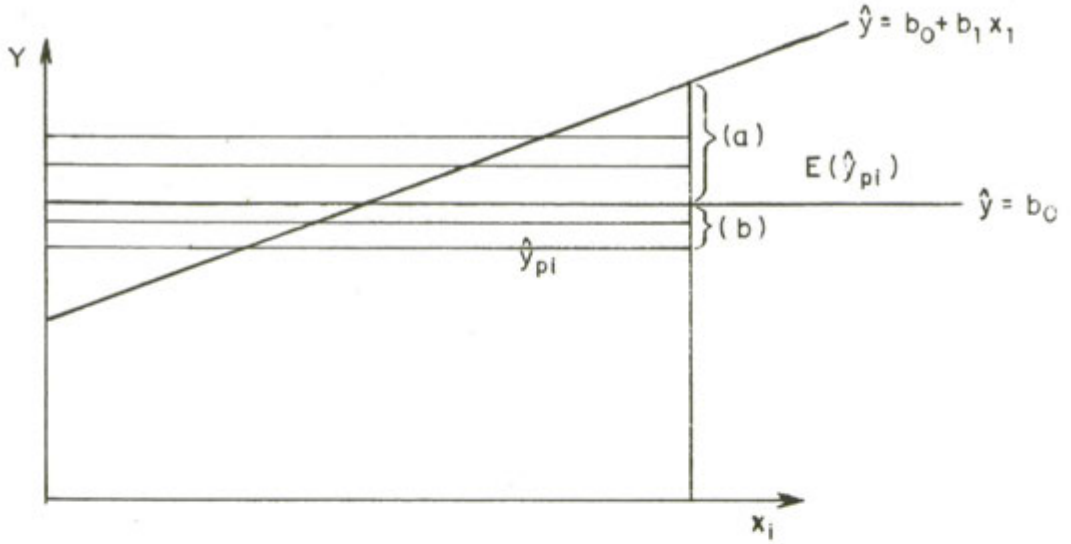


Figura 4.1.—Descomposición del error cuadrático total.

Se demuestra que (THOMPSON, 1978b) $\sum_{i=1}^{n} var(\hat{y}_{pi}) = p\sigma^2$

y que la esperanza de $\sum_{i=1}^{n} (y_i - \hat{y}_{pi})^2$ *es*

$$E(SCE_p) = SCD_p + (n-p)\sigma^2 \qquad [4.1]$$

siendo

$$SCE_p = \sum_{i=1}^{n} (y_i - \hat{y}_{pi})^2 \qquad [4.2]$$

Por lo tanto, $SCD_p = E(SCE_p) - (n-p)\sigma^2$*, sustituyendo estos valores en la fórmula de* Γ_p *queda*

$$\Gamma_p = \frac{E(SCE_p)}{\sigma^2} + 2p - n \qquad [4.3]$$

MALLOWS (1966) recomienda el empleo del siguiente estadístico para estimar Γ_p:

$$C_p = \frac{SCE_p}{\hat{\sigma}^2} + 2p - n \qquad [4.4]$$

siendo $\hat{\sigma}^2$ la estima de σ^2 a través del modelo completo.

Si la ecuación de regresión con p variables describe bien los datos $SCD_p = 0$; en este caso, SCE_p estimará $(n-p)\sigma^2$ y

$$C_p = \frac{(n-p)\hat{\sigma}^2}{\hat{\sigma}^2} + 2p - n = p \qquad [4.5]$$

La búsqueda del conjunto óptimo de variables equivale a identificar cuál es el conjunto que conduce al valor más pequeño de C_p y, a la vista de [4.5], los valores de C_p que están más próximos a p.

En la ecuación

$$C_p = \frac{SCE_p}{\sigma^2} + 2p - n$$

el término SCE_p (ver [4.2]) decrece cuando crece p mientras 2p crece cuantas más variables estén incluidas en la ecuación de regresión. Para calcular C_p debe usarse una estima insesgada de σ^2. En general, se usa el cuadrado medio del error de la regresión con las k-variables bajo la hipótesis de que este modelo es el verdadero. Debe notarse que $C_k = k$.

4.9.2. *Método gráfico de Mallows*

En la Tabla 4.2 en la columna «C_p» se reflejan los valores del estadístico C_p calculados para todas las regresiones según la fórmula [4.4], empleando los valores relativos a las tablas de análisis de varianza que figuran en la Tabla 4.4.

TABLA 4.4

Análisis de varianza de las regresiones

Variables regresoras	*Fuente*	*Suma de cuadrados*	*G.L.*	*Cuadrado medio*	*F*
x_1	Regresión	1820,227	1	1820,227	
	Residuo	2294,197	19	120,747	15,075
x_2	Regresión	2801,109	1	2801,109	
	Residuo	1313,316	19	69,122	40,524
x_3	Regresión	735,349	1	735,349	
	Residuo	3379,076	19	177,846	4,135
x_4	Regresión	2956,169	1	2956,169	
	Residuo	1158,255	19	60,961	48,493
x_1, x_2	Regresión	4059,804	2	2029,902	
	Residuo	54,620	18	3,034	668,948
x_1, x_3	Regresión	1884,064	2	942,032	
	Residuo	2230,362	18	123,909	7,603
x_1, x_4	Regresión	4029,422	2	2014,711	
	Residuo	85,003	18	4,722	426,632
x_2, x_3	Regresión	3401,884	2	1700,942	
	Residuo	712,541	18	39,586	42,969
x_2, x_4	Regresión	2956,752	2	1478,376	
	Residuo	1157,673	18	64,315	22,986
x_3, x_4	Regresión	3822,206	2	1911,103	
	Residuo	292,218	18	16,234	117,720
x_1, x_2, x_3	Regresión	4073,762	3	1357,921	
	Residuo	40,663	17	2,392	567,709

TABLA 4.4

Análisis de varianza de las regresiones (continuación)

Variables regresoras	*Fuente*	*Suma de cuadrados*	*G.L.*	*Cuadrado medio*	*F*
x_1, x_2, x_4	Regresión	4072,596	3	1357,532	551,733
	Residuo	41,828	17	2,460	
x_1, x_3, x_4	Regresión	4067,090	3	1355,697	486,891
	Residuo	47,335	17	2,784	
x_2, x_3, x_4	Regresión	4033,062	3	1344,354	280,890
	Residuo	81,363	17	4,786	
x_1, x_2, x_3, x_4	Regresión	4073,856	4	1018,464	401,674
	Residuo	40,569	16	2,536	

En la figura 4.2 se han representado en ordenadas los valores de C_p y en bcisas el número de variables incluidas en el modelo de regresión. (Solamente se han representado los valores correspondientes a C_p inferiores a 10).

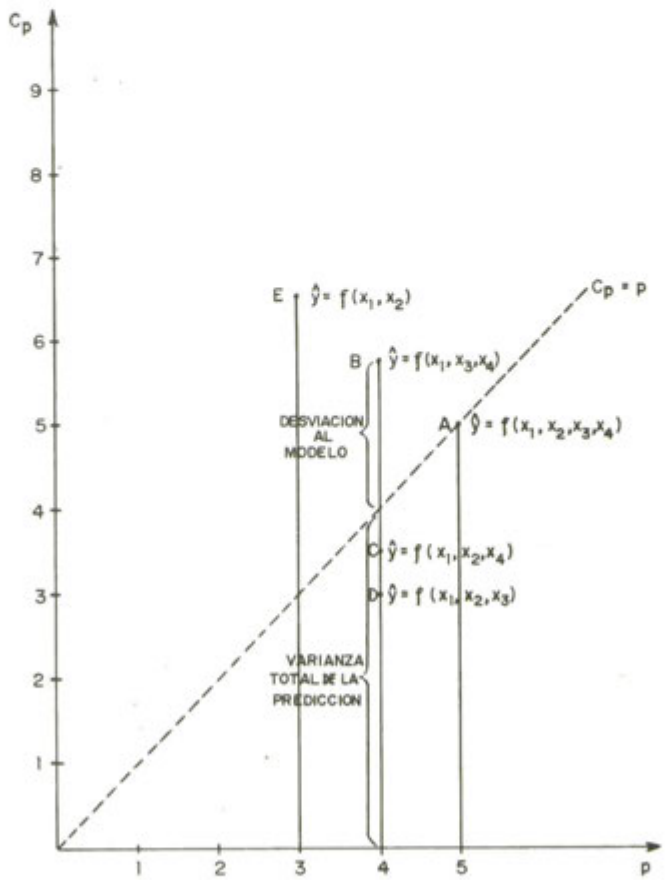


Figura 4.2.—Método gráfico de Mallows.

Las ecuaciones de regresión que tengan una desviación pequeña respecto al modelo verdadero tenderán a colocarse cerca de la recta $C_p = p$, por ejemplo el punto A; por el contrario, aquellas que tengan una desviación importante se colocarán por encima de dicha recta.

La introducción de variables en el modelo disminuye la desviación respecto al modelo verdadero, pero también aumenta la varianza total de la predicción para los n puntos

$$\sum_{i=1}^{n} \text{var}(\hat{y}_{pi})$$

tal y como se indicó en la introducción del capítulo. En la figura 4.2 puede observarse este efecto al pasar del punto E [$\hat{y} = f(x_1, x_3)$] al punto B [$\hat{y} = f(x_1, x_3, x_4)$].

El aumento producido en el término $\sum_{i=1}^{n} \text{var}(\hat{y}_{pi})$ tiene como consecuencia el aumento del promedio de la varianza de la predicción para cada punto.

Si interesa tener una dispersión más pequeña en la predicción, es necesario elegir como contrapartida un modelo que esté más desviado respecto al verdadero.

Si el criterio de selección consiste en minimizar el valor de C_p, se elegirá el punto D que representa a la recta de regresión $y = f(x_1, x_3, x_5)$ cuya ecuación es:

$$\hat{y} = 46,776 + 1,776x_1 + 0,655x_2 + 0,260x_3$$

CAPITULO 5

MODELOS INCLUSIVOS

5.1. Introducción

Cuando un investigador utiliza en su estudio algún cálculo de regresión, puede ocurrir que, debido a sus conocimientos a priori o a la forma de entender su problema considere como lógico un cierto orden de importancia entre las variables regresoras de que dispone. Dicho grado de interés puede tener justificaciones intrínsecas a los mecanismos del objeto de su estudio y también puede estar influido por problemas del coste de obtención de las variables; es decir, no se desea emplear una variable costosa sin estar seguro de que la información adicional que aporta con respecto a las demás, compensa suficientemente su precio. En algunos modelos se impone una jerarquía natural en las variables, ya que algunas de ellas solamente tiene razón de ser en función de la presencia de otras en el modelo.

En estos dos casos, es aconsejable utilizar lo que se podrían denominar «modelos inclusivos». Una razón estadística para preferir este enfoque a la búsqueda sistemática (presentada en el capítulo 4 de selección de variables) es el mejor control ejercido sobre el error α de cada una de las pruebas F efectuadas en el desarrollo de la tabla del análisis de la varianza.

5.2. Definición

Se dice que un modelo A está incluido en un modelo B cuando es un caso particular de éste. Esta situación se denotará por $A \subset B$. Por ejemplo:

Modelo A: $Y=\beta_0+\beta_1 x_1+\varepsilon$

Modelo B: $Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$

En este caso, basta hacer $\beta_2=0$ para pasar del modelo B al A.

5.3. Procedimiento para el análisis

Los pasos seguidos en el desarrollo de los modelos inclusivos son los siguientes:

a) Definir una serie de modelos incluidos a priori unos en otros. Si hay k modelos A_1, A_2, ... A_k, se puede definir, por ejemplo, la situación de inclusión $A_1 \subset A_2 \subset \ldots \subset A_k$.

b) Considerar siempre como suma de cuadrados del residuo la obtenida por medio del modelo más completo (A_k para el ejemplo anterior).

c) Probar la reducción en la suma de cuadrados de la regresión (o aumento de la del residuo) entre el modelo A_{k-1} y el modelo A_k. Esta prueba se puede feectuar mediante el estadístico F* (ver 4.3). Si la prueba es significativa, se debe mantener el modelo A_k. Por el contrario, si no lo es, rechazar el modelo A_k continuando con el proceso de la reducción.

d) Repetir el proceso con los modelos sucesivos descendiendo hasta encontrar una prueba significativa o hasta mantener el modelo más pequeño A_1 (ver el ejemplo de la heterogeneidad de pendientes tratado en 5.4.2 y la eliminación descendente de variables en 4.5).

5.4. Elección de la sucesión de inclusiones

En la mayoría de los casos la determinación de la serie de modelos inclusivos no resulta evidente. Si se admite que cuantos menos términos comporte el modelo más interesante será para la interpretación y el uso posterior, resulta lógico pensar que las variables consideradas como importantes deben eliminarse lo más tarde posible. En otras palabras, si las variables regresoras se ordenan según su orden decreciente de importancia a priori $x_1, x_2 \ldots x_k$, se denominarán los modelos siguientes:

$$A_1 : Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$A_2 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\vdots$$

$$A_k : Y = + x + x + \ldots + {}_k x_k +$$

El actuar de esta manera equivale a emplear un método descendente de selección de variables pero de manera totalmente dirigida o controlada.

En algunas situaciones particulares, las ordenaciones parciales sucesivas se encuentran impuestas por la índole del problema. En este contexto, se examinará el problema de las superfices de respuesta y, con algo más de detalle, el de las regresiones con pendientes variables en diferentes poblaciones.

5.4.1. *Superficie de respuesta*

Puede interesar obtener la respuesta de una variable en función de dos factores cuantitativos. Sea, por ejemplo, el rendimiento de un cultivo (Y) como consecuencia de un cierto abonado nitrogenado (N) y fosfórico (P). La denominación de superficie de respuesta se justifica por la representación gráfica que se puede hacer de estos modelos. En un eje se representan los niveles de nitrógeno, en otro los de fósforo y en el vertical el rendimiento (figura 5.1).

Entre los modelos a estudiar podríamos considerar, por ejemplo, los siguientes:

$$A_1 : Y = \beta_0 + \beta_1 N + \varepsilon$$

$$A_2 : Y = \beta_0 + \beta_1 N + \beta_2 P + \varepsilon$$

$$A_3 : Y = \beta_0 + \beta_1 N + \beta_{11} N^2 + \varepsilon$$

$$A_4 : Y = \beta_0 + \beta_1 N + \beta_{11} N^2 + \beta_2 P + \varepsilon$$

$$A_5 : Y = \beta_0 + \beta_1 N + \beta_{11} N^2 + \beta_2 P + \beta_{22} P^2 + \beta_{12} NP + \varepsilon$$

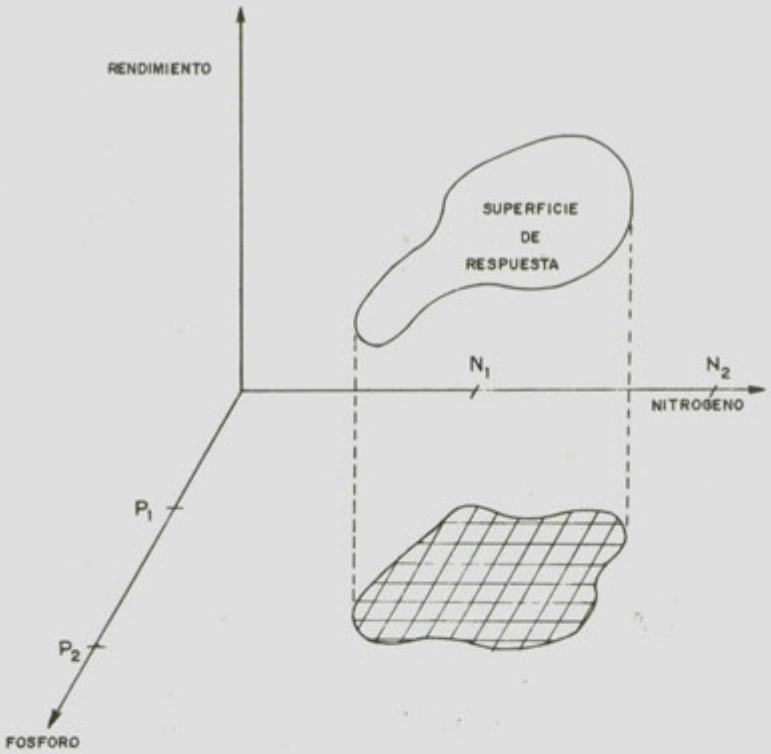


Figura 5.1.—Superficies de respuesta.

Las relaciones de inclusión para este ejemplo son:

$$A_1 \subset A_2 \subset A_3 \subset A_4 \subset A_5$$

En esta situación parece lógico emplear las pruebas siguientes:

a) A_4 contra A_5

b) A_3 contra A_4 (o A_3 contra A_4)

c) A_1 contra A_3 (o A_1 contra A_3).

El procedimiento de parada en estas pruebas encadenadas es, como anteriormente, obtener una significación que impida, por tanto, reducir más el modelo.

5.4.2. *Heterogeneidad de las pendientes*

5.4.2.1. Planteamiento del problema.

Puede ocurrir que los cálculos de la regresión se refieran a una muestra que posea una estructura determinada; es decir, que las observaciones se clasifiquen según ciertos criterios. Por ejemplo, se quiere predecir el número

de huevos puestos por un insecto en función de la temperatura a la que han sido sometidos durante el desarrollo, cuando se dispone de datos de diversas especies de un mismo género. En esta situación, resulta lógico considerar a los insectos de la misma especie más próximos entre sí que con respecto a los de otra especie. Ello puede traducirse a priori en un modelo de regresión en el que se postula que las rectas de regresión no tienen por qué ser iguales en ambas especies. La figura 5.2 representa gráficamente esta idea: La especie 1 es siempre más prolífica que las otras tres, si bien el aumento del número de huevos (pendiente de la recta) es inferior al de la especie 3. Por otro lado, la especie 3 produce menos huevos que la 4 a temperaturas bajas pero más a temperaturas altas, etc.

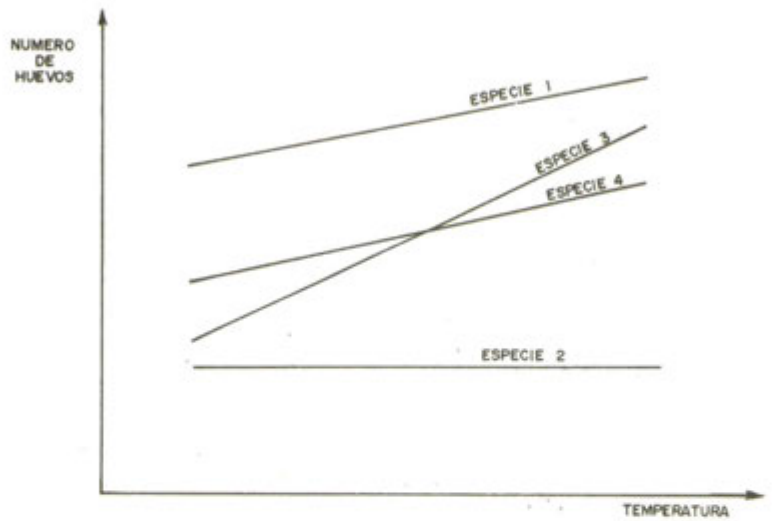


Figura 5.2.—Ejemplo de regresiones diferentes según la especie considerada.

Si H_{ij} es el número de huevos producidos por el insecto j de la especie i criado a la temperatura t_{ij}, el modelo general será:

$$A_3 : H_{ij} = \beta_{0i} + \beta_{1i}\, t_{ij} + \varepsilon_{ij}$$

Se pueden probar muchos modelos incluidos dentro de éste; sin embargo, para el problema que se trata, se consideran los dos siguientes:

$$A_2 : H_{ij} = \beta_{0i} + \beta_1\, t_{ij} + \varepsilon_{ij}$$

$$A_1 : H_{ij} = \beta_0 + \beta_1\, t_{ij} + \varepsilon_{ij}$$

El modelo A_1, corresponde a la identidad de las producciones de huevos en cada una de las especies (ni β_0 ni β_1 dependen de *i);* el A_2 postula únicamente que un mismo aumento de temperatura produce un mismo incremento en la producción sea cual sea la especie considerada (paralelismo de las pendientes: β_1 no depende de i).

Como estos modelos son inclusivos según el esquema $A_1 \subset A_2 \subset A_3$, se utilizará el procedimiento de estudio del análisis de varianza presentado en la tabla 5.1, y descrito genéricamente en 5.3.

TABLA 5.1

Tabla del Análisis de la Varianza para probar la heterogeneidad de las pendientes cuando existen varias poblaciones

Fuentes de variación	S.C.	g.l	Fc
$(H_0: \beta_{11}=\beta_{12}=\ldots=\beta_1)$ A_2 vs. A_3 Heterogeneidad de las pendientes	$S_1=SCEC-\sum_i SCE_i$	$G_1=I-1$	$F_1=\frac{S_1/G_1}{S_4/G_4}$
$(H_0: \beta_{01}=\beta_{02}=\ldots=\beta_0/$ Igualdad de las regresiones suponiendo las pendientes iguales	$S_2=SCET-SCEC$	$G_2=I-1$	$F_2=\frac{S_2/G_2}{S_4/G_4}$
Efecto de la regresora suponiendo las regresiones iguales	$S_3=Sy^2-SCET$	$G_3=1$	$F_3=\frac{S_3/G_3}{S_4/G_4}$
$(H_0: \beta_1=0/\beta_0, \beta_1)$ Residuo del modelo completo	$S_4=\sum_{i=1} SCE_i$	$G_4=N-2I$	

5.4.2.2. Cálculo de las sumas de cuadrados

Si se dispone de n_i observaciones para la especie *i* y de I especies, los cálculos de las sumas de cuadrados permiten obtener las pruebas de significación de algunos elementos de los modelos que pueden ser de interés para el investigador. Estos cálculos se basan en hallar las sumas de cuadrados del residuo de varias regresiones, cuyos valores se compararán posteriormente entre sí. De la misma manera que se trabaja con los residuos, se podría actuar con las sumas de cuadrados de las regresiones.

Los pasos son los siguientes:

a) Regresiones para cada una de las poblaciones: Se calcula una regresión para cada una de las *i* especies (i=1 ... I). La suma de cuadrados del residuo en cada una de ellas se denotará por SCE_i.

b) Regresión sobre el conjunto de las poblaciones: Sobre todas las especies, consideradas como un solo conjunto de datos, se calculan dos regresiones:

i. Para los datos iniciales se obtiene la suma de cuadrados del residuo SCET y la suma de cuadrados total Sy^2.
ii. Para los datos corregidos por la media de cada especie (es decir, para $H'_{ij}=\overline{H}_{ij}-_i$, en donde $\overline{H}_i$ es la puesta media de la especie *i)* se obtiene la suma de cuadrados del residuo SCEC.

Con todos estos datos se construye un análisis de varianza como el presentado en la tabla 5.1.

5.4.2.3. Ejemplo:

A continuación presentamos un ejemplo numérico aclaratorio de los pasos a seguir. Los datos son los siguientes:

Población 1		*Población 2*		*Población 3*	
x	*y*	*x*	*y*	*x*	*y*
24	27	26	26	20	29
18	26	29	28	26	29
30	29	28	23	14	24
15	23	16	21	30	32
21	28			24	28
27	29				

Los cálculos intermedios para llegar a la tabla del análisis de la varianza son:

a) Regresión por población:
Se calcula para cada población la suma de cuadrados del residuo. Para la población 1, por ejemplo, se procede de la forma siguiente:

$$Sx^2_{(1)} = 157,5 \qquad Sy^2_{(1)} = 26 \qquad Sxy_{(1)} = 57$$

$$b_{1(1)} = 0,3619 \qquad Sy^2_{(1)} - b_{1(1)} Sxy_{(1)} = SCE_1 = 5,3714$$

De manera similar se obtendría

$SCE_2 = 0,0924$

$SCE_3 = 5,3280$

i. Sobre los datos iniciales.

Suma de cuadrados del residuo: SCET = 46,7257

Suma de cuadrados total: $Sy^2 = 122,4$

ii. Sobre datos centrados $H_{ij} - \bar{H}_i$

Suma de cuadrados del residuo: SCEC = 12,6117.

Las sumas de cuadrados para la tabla del análisis serán:

$$S_1 = SCEC - \sum_i SCE_i = 12,6117 - (5,3714 + 0,0924 + 5,3280) = 1,8199$$

$$S_2 = SCET - SCEC = 46,7257 - 12,6117 = 34,1140$$

$$S_3 = Sy^2 - SCET = 122,4 - 46,7257 = 75,6743$$

$$S_4 = \sum_i SCE_i = 5,3714 + 0,0924 + 5,3280 = 10,7918$$

La tabla del análisis de Varianza se construirá de la forma siguiente:

Fuentes de variación	*S.C.*	*g.l.*	*C.M.*	*Fc*
Heterogeneidad de las pendientes	1,82	2	0,91	0,76
Igualdad de las regresiones suponiendo las pendientes iguales	34,11	2	17,06	14,22
Efecto de la regresora suponiendo las regresiones iguales	75,67	1	75,67	63,11
Residuo del modelo completo	10,79	9	1,20	

De los resultados de la tabla se deduce que no hay heterogeneidad de las pendientes aunque las regresiones son diferentes (debido a la diferencia del β_0).

CAPITULO 6

VALIDACION DEL MODELO

6.1. Planteamiento del problema

Para permitir las inferencias o generalizaciones de una muestra a una población, la Estadística dispone de los modelos. Sin embargo, es necesario reconocer que los modelos que se emplean, han sido escogidos por sus propiedades matemáticas interesantes y no siempre para concordar con una realidad biológica; además no es posible explotar todas las posibilidades en cuanto a su interpretación si el número de parámetros llega a ser demasiado grande. Pero es un tema todavía más delicado el que se va a tratar en este capítulo: poner en duda el modelo dentro de las fronteras en donde se refugian todos los razonamientos que conducen a proponer las estimaciones, pruebas de hipótesis, intervalos de confianza, etc., presentados en los capítulos anteriores.

Esta puesta en duda es de suma importancia pues el empleo de un modelo incorrecto puede inducir a conclusiones erróneas si se utiliza la regresión como simple herramienta de cálculo. Este capítulo se ha colocado casi al final de la publicación porque, en general, la validación del modelo con metodología estadística se efectúa a posteriori; es decir, una vez realizadas las fases anteriores de estimación de parámetros.

La validación del modelo se presentará en tres puntos. En primer lugar en relación con las hipótesis o suposiciones en las que se basa el modelo, en segundo lugar estudiando su capacidad o bondad predictiva y en tercer lugar a través del estudio de los residuos.

6.2. Verificación de las suposiciones del modelo

El cumplimiento de las suposiciones del modelo se puede llevar a cabo de dos maneras; una que se podría denominar conceptual, y otra mediante pruebas estadísticas.

La verificación conceptual consiste en tomar a las suposiciones como tales, es decir, algo que no se cuestiona. Se basa en el conocimiento que el investigador tiene de su problema concreto y de lo que está dispuesto a aceptar sin prueba en relación con el comportamiento de la variable aleatoria que está estudiando. Por ejemplo, si desea verificar la normalidad de la variable, el enfoque conceptual le llevaría a utilizar como argumento, el teorema central del límite. De una forma sencilla, se podría dar de él la interpretación siguiente: si una variable aleatoria continua Y puede considerarse como suma de las influencias aleatorias de un número grande de causas independientes cada una de ellas con varianza similar, la función de densidad de dicha variable será aproximadamente normal (más aproximadamente cuanto mayor sea el número de causas). Conviene tener presente este resultado para saber en qué circunstancias se puede aceptar la normalidad de la variable (o, con más generalidad, cualquiera de las suposiciones del modelo) y cuándo

es necesario recurrir a las pruebas estadísticas. Un inconveniente de emplear reiteradamente unos mismos datos para obtener diferentes conclusiones estadísticas (normalidad, existencia de β_1, etc.) es que la potencia de las pruebas disminuye.

En la literatura existen descritas varias pruebas para verificar estadísticamente las suposiciones del modelo. Aquí se presentarán algunas de las más utilizadas.

6.2.1. *Independencia*

La hipótesis de independencia de los errores se verifica por la prueba presentada por DURBIN y WATSON (1951). Para poder efectuarla es necesario que las observaciones se encuentren ordenadas de acuerdo con un criterio establecido, por ejemplo el tiempo de obtención de la observación, etc.

La prueba consiste en el cálculo del estadístico d

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Siendo e_i el valor del residuo obtenido con los datos de la muestra para la i-ésima observación el experimento, definido como la diferencia entre el valor observado para la variable Y y el valor predicho por el modelo para un valor dado x_i. A estos autores sólo les fue posible establecer los límites superior (d_s) e inferior (d_i) para los niveles de significación del estadístico «d». Estos valores críticos se encuentran en el Anejo 1.

La hipótesis nula que se verifica mediante estos límites es la independencia de errores, frente a la alternativa de autocorrelación positiva. La decisión se establece en el sentido siguiente:

1. Si $d < d_i$ se rechaza la hipótesis nula (se rechaza la independencia).
2. Si $d > d_s$ no se rechaza.
3. Si $d_i < d < d_s$ la prueba es ambigüa.

Si el valor del estadístico «d» es mayor que 2, se contrasta con la alternativa de autocorrelación negativa. Para ello se calcula $4 - d$ y este valor se compara con d_i y d_s como si se contrastara la autocorrelación positiva. Más recientemente, DURBIN (1970) ha diseñado otra prueba que resuelve la indeterminación del punto 3.

Es importante destacar que estas pruebas actúan sólo para el caso de modelo fijo.

A continuación se presenta un ejemplo para poner de manifiesto el método de cálculo a seguir. Sea la tabla de valores siguiente:

x	7	6	8	3	5	2	10	2	10	9
y	13	12	12	9	9	8	15	6	17	14

Empleando estimación mínimo cuadrática, tal como se explicó anteriormente (2.1.1.), se ha obtenido que la ecuación de regresión lineal es:

$$\hat{Y} = 4{,}99 + 1{,}05\ x$$

Por lo tanto, los valores predichos por la ecuación, simbolizados por $\hat{y}_i$ y sus correspondientes residuos son:

y_i	13,00	12,00	12,00	9,00	9,00	8,00	15,00	6,00	17,00	14,00
$\hat{y}_i$	12,34	11,29	13,39	8,14	10,24	7,09	15,49	7,09	15,49	14,44
e_i	0,66	0	-1,39	0,86	-1,24	0,91	-0,49	-1,09	1,51	-0,44

Suponiendo que los datos fueron tomados en el orden indicado y que esta ordenación tenía sentido real, se desea estudiar la independencia temporal de las observaciones utilizando la prueba descrita. Para ello, se calcula el estadístico d.

$$d = \frac{(0{,}71-0{,}66)^2+(-1{,}39-0{,}71)^2+\ldots+(-0{,}44-1{,}51)^2}{0{,}66^2+0{,}71^2+\ldots+1{,}51^2+(-0{,}44)^2} = \frac{31{,}39}{9{,}879} = 3{,}18$$

Como $d > 2$, se contrasta el valor $4-3{,}18=0{,}82$, por lo que al 1% la prueba resulta ambigua, rechazándose al 5% la independencia.

Cuando se rechaza la suposición, debe investigarse la causa de la dependencia. A veces, simplemente se trata de una especificación falsa del modelo y, por ejemplo, el modelo correcto es parabólico y no lineal (ver 6.4.5.2 y 8.2). En otros casos el origen de la dependencia no es tan aparente, siendo necesario averiguar qué tipo de estructura de correlación existe para incorporarla al modelo. Por lo específico y complicado del tema, no se trata aquí este segundo supuesto.

6.2.2. *Homoescedasticidad*

Para poder probar la homogeneidad de la varianza de la variable dependiente a valores fijos de la variable regresora, es necesario disponer de medidas repetidas de la variable dependiente para cada uno de los valores de la regresora. Con ello se tendrán estimas de la varianza.

Existen numerosas pruebas para verificar la homoescedasticidad. La más popular es la sugerida por BARTELETT (1937), aunque tiene el gran inconveniente de ser extraordinariamente sensible a ligeras desviaciones de la normalidad de la distribución. Una prueba muy simple y que no presenta ese inconveniente es la descrita por BURR y FOSTER (1972). Esta prueba resulta muy sencilla de cálculo y, además, aunque una varianza muestral valga cero no afecta a la prueba, como ocurre con otras que también son de cálculo muy simple (como la de Hartley, por ejemplo).

La prueba de Burr y Foster se calcula de la forma siguiente:

Para tamaños de muestra constantes «r», sea s_i^2 (i=1, ..., p) la varianza muestral de Y para cada x_i; el valor del estadístico q_c es

$$q_c = (s_1^4 + \ldots + s_p^4)/(s_1^2 + \ldots + s_p^2)^2$$

Para tamaños de muestra diferentes, el estadístico q_c se construye como:

$$q_c = \bar{\nu}(\nu_1 s_1^4 + \ldots + \nu_p s_p^4)/(\nu_1 s_1^2 + \ldots + \nu_p s_p^2)^2$$

siendo ν_i los grados de libertad en los que está basada cada varianza muestral ($\nu_i = r_i - 1$) y $\bar{\nu}$, la media armónica de los grados de libertad.

Si $q_c > q$, siendo q el valor de las tablas que se encuentran en el Anejo 2, se rechaza la hipótesis de homoescedasticidad.

A continuación se verá un ejemplo de cómo se calcula esta prueba:

Se dispone de la siguiente tabla de datos:

x	Y			
3	120;	123;	118;	125
4	128;	127;	132;	135
5	147;	132;	139;	142
6	192;	198;	210;	194
7	201;	213;	205;	220
8	300;	315	320;	310

Dado que existen medidas repetidas de la Y, es posible estudiar si la varianza de la Y para diferentes valores de la x es constante o no.

Empleando la prueba de Burr-Foster se obtendrían los siguientes resultados:

$$s_1^2 = [120^2 + \ldots + 125^2 - \frac{(120 + \ldots + 125)^2}{4}]/\ 3 = 9{,}67 \qquad s_1^4 = 93{,}44$$

$$s_2^2 = 13{,}67 \qquad s_2^4 = 186{,}78$$

$$s_3^2 = 39{,}33 \qquad s_3^4 = 1547{,}11$$

$$s_4^2 = 65{,}00 \qquad s_4^4 = 4225{,}00$$

$$s_5^2 = 71{,}58 \qquad s_5^4 = 5124{,}17$$

$$s_6^2 = 72{,}92 \qquad s_6^4 = 5316{,}84$$

$$\text{Total} = 272{,}17 \qquad 16493{,}25$$

El valor calculado para el estadístico q_c en la muestra presente es:

$$q_c = \frac{16493{,}25}{272{,}17} = 0{,}22$$

Según las tablas del Anejo 2 para $\nu=3$; $p=6$; y $\alpha=0{,}01$ q es 0,43 y para $\alpha=0{,}001$, q es 0,546. Como $0{,}22 < 0{,}43$, no se puede rechazar la hipótesis de homoescedasticidad al 1% de significación.

Para mayor descripción de la prueba, puede consultarse ANDERSON y MCLEAN (1974). Si la hipótesis se rechaza, puede intentarse efectuar transformaciones en la variable dependiente que homogeneicen las varianzas. Estos autores sugieren como regla práctica que:

1. Si se acepta la homogeneidad al 1%, no se transforman los datos.
2. Si se rechaza al 0,1%, transformar.
3. Si el resultado de la prueba está entre ambos márgenes es necesario conocer la distribución teórica de los datos. Si existe una razón fundada para la heteroescedasticidad, transformar; de lo contrario, no.

Si es necesario transformar, se sugieren algunas transformaciones de acuerdo con el tipo de diferencia que exista entre las varianzas. Muy a menudo, la heteroescedasticidad es consecuencia de la falta de normalidad de la variable. En una distribución normal, la media y la varianza empíricas son independientes. Si ocurre que las medias altas llevan asociadas varianzas altas, es indicación clara de falta de normalidad y/o patente heteroescedasticidad.

Cuando la media esté relacionada linealmente con la varianza, se sugiere la transformación $Y \rightarrow \sqrt{Y}$ que, además de homogeneizar las varianzas, automáticamente suele normalizar la distribución. En caso de que la media esté correlacionada con la desviación típica, puede utilizarse la transformación $y \rightarrow \log Y$. Esta transformación también se emplea en datos procedentes de problemas relacionados con curvas de crecimiento. Si la desviación típica está asociada al cuadrado de la media, la transformación más recomendada es $Y \rightarrow \sqrt{1/Y}$. Además, ANDERSON y MCLEAN (1974) indican que cuando la media está negativamente correlacionada con las varianzas debe transformarse según $Y \rightarrow \sqrt{B} - \sqrt{B-Y}$, siendo B el límite superior de los datos.

Cuando la variable Y sigue una distribución binomial, expresando los datos en porcentaje y de tal manera que el número de observaciones es pequeño y el porcentaje próximo al 0% o al 100% se sugiere utilizar la transformación $Y \rightarrow \text{arcsen} \sqrt{\text{porcent./100}}$. Cuando la distribución es de Poisson, debe emplearse la transformación de la raíz cuadrada anteriormente mencionada. Si en este caso existen muchos valores nulos ,se puede intentar $Y \rightarrow \sqrt{Y+1}$ o mejor $Y \rightarrow \sqrt{Y+1/2}$.

Naturalmente, después de efectuar la transformación es necesario comprobar que ésta ha sido efectiva. BOX y COX (1964) presentan una fórmula general para obtener transformaciones en distribuciones asimétricas, mientras que JOHN y DRAPER (1980) sugieren otra familia de transformaciones cuando la distribución es simétrica, pero con colas excesivamente largas.

6.2.3. *Normalidad*

Existen también numerosas pruebas de normalidad, sin embargo se recomienda entre todas la propuesta por SHAPIRO y WILK (1965). Esta prueba tiene la ventaja de no tener que incluir la media y la varianza de la distri-

bución como parte de la hipótesis, tal como ocurre en la de Kolmogorov-Smirnov o la chi-cuadrado. Estas y otras pruebas fueron comparadas, entre otros, por SHAPIRO et al. (1968) y PEARSON et al. (1977), demostrando que generalmente aquélla era superior en detectar la falta de normalidad al ser evaluada a través de varias alternativas (simétrica, asimétrica, colas largas o cortas, etc.) en muestras de tamaño 10 a 50.

Para utilizar la prueba de Shapiro y Wilk, es necesario obtener previamente los valo , residuales estimados con los datos, también llamados residuos, tal como se indicó en 6.2.1.

Esta prueba no fue originariamente descrita para ser utilizada en conexión con problemas de regresión sino para verificar la normalidad de la distribución de una serie de observaciones extraídas al azar de una distribución única. Por tanto, es necesario, reconocer que su uso es meramente indicativo a efectos prácticos y que se debe ajustar convenientemente la metodología de cálculo al problema que nos ocupa. Se sabe que estos valores residuales muestrales a diferencia de los teóricos del modelo, no son independientes (Ver apartado 6.4.2.1.); por lo tanto, se requiere que su número sea lo suficientemente grande como para paliar el efecto negativo de la falta de independencia (no aleatoriedad). Asimismo, el valor residual dependerá del número de variables regresoras del modelo y será necesario tenerlo en cuenta al efectuar la prueba. Si se dispone de más de una observación de Y por cada valor de x, podría efectuarse la prueba de normalidad tal como fue descrita por sus autores. Esta prueba habría que realizarla con los valores de las observaciones dentro de cada x y repetir sucesivamente la prueba para todas y cada una de las x. Ahora bien, como generalmente el número de repeticiones suele ser pequeño, la potencia de la prueba es muy baja, por lo que se declararía «normalidad» (aceptación de la hipótesis) en muchos casos de distribución no normal. Por tanto, se desaconseja esta práctica.

Si el número de valores residuales es «m» y «d» los grados de libertad de la línea correspondiente a las desviaciones del modelo en la tabla del análisis de la varianza de la regresión (ver 2.1.2) los pasos a seguir para el cálculo del estadístico Wc, en el que se basa la prueba, son:

1. Ordenar los residuos de forma creciente

$$e_{[1]} \leq e_{[2]} \leq e_{[3]} \leq \ldots\ldots \leq e_{[m]}$$

2. Calcular $S = \sum_{i} e_i^2$

3. Si m es par; $m=2k$, calcular

$$b = \sum_{i=1}^{k} a_{m,i} \left(e_{[m-i+1]} - e_{[i]}\right)$$

en donde los valores $a_{m,i}$ aparecen en el Anejo 3.

Si m es impar; $m=2k+1$, se omite del cálculo el valor de la mediana.

4. Calcular $W_c = b^2/S$
5. Comparar W_c con el valor W en las tablas del Anejo 4.

Valores $W < W$ indican falta de normalidad. El valor de W se busca empleando los grados de libertad «d» y el α elegido. Utilizando las observaciones del ejemplo para la prueba de la independencia, se obtendrían los siguientes valores:

1. Se ordenan los residuos

—1,39; —1,24; —1,09; —0,49; —0,44; 0,66; 0,71; 0,86; 0,91; 1,51.

2. Se calcula S

$$S = \sum_{i=1}^{10} e_i^2 = 9,879$$

3. Se construyen las diferencias siguientes multiplicadas por los coeficientes $a_{m,i}$

$$\begin{array}{ll} [1,51-(-1,39)] & (0,5739)=1,6643 \\ [0,91-(-1,24)] & (0,3291)=0,7076 \\ [0,86-(-1,09)] & (0,2141)=0,4175 \\ [0,71-(-0,49)] & (0,1224)=0,1469 \\ [0,66-(-0,44)] & (0,0399)=0,0439 \\ & \overline{\qquad\qquad} \\ & b=2,9802 \end{array}$$

4. $W_c = \dfrac{b^2}{S} = 0,899$

5. El valor de las tablas para $\alpha=0,01$ y $d=8$ es $W=0,749$. Como ocurre que $0,899 > 0,749$ se concluye que no hay evidencia en contra de la normalidad de la distribución.

Además de los métodos paramétricos, existen otras técnicas gráficas que, sin asociar afirmaciones probabilísticas a las pruebas de hipótesis, sí permiten obtener indicaciones de la presencia de normalidad en la distribución. Un gráfico muy utilizado consiste en representar los valores residuales en escala de probabilidad normal tipificada. Es decir, en el eje horizontal se ponen los residuos ordenados de menor a mayor $e_{(1)}$, $e_{(2)}$, $e_{(3)}$, ..., $e_{(m)}$; el eje vertical es el valor esperado de una distribución normal tipificada basado en el orden que ocupa cada observación. Si el subíndice «j» de los residuos representa el número de orden dentro de la ordenación anterior, la ordenada corresponde al valor normal esperado para el orden relativo a la observación. El valor normal esperado se estima por $\varnothing^{-1}[3j-1)/(3m+1)]$, siendo $\varnothing$ la función de distribución de una curva normal tipificada. Es decir, es el punto que deja a su izquierda un área igual a $(3j-1)/(3m+1)$. Si los datos proceden realmente de una distribución normal, la línea que forman los puntos será recta excepto por fluctuaciones aleatorias.

6.3. Bondad de un modelo predictivo

Un modelo será tanto mejor para predecir cuanto, a igualdad de otras condiciones, presente más estabilidad al tomar varias muestras. En principio se podría pensar que sería recomendable tomar dos muestras, una para estimar el modelo y otra para evaluar su capacidad predictiva. Normalmente, debido a las condicoines experimentales, no suele ser factible emplear este método. Si el número de observaciones que se han tomado es suficientemente amplio, se puede partir el conjunto de datos en dos subconjuntos, uno para estimación y otro para verificación. Recientemente Snee (1977) ha revisado

una serie de métodos para efectuar esta partición y ha presentado una nueva metodología para la valoración de modelos basada en el algoritmo «Duplex» de Kennard, que estaba sin publicar anteriormente. Se sugiere que esta partición no se haga a menos que el número de observaciones sea mayor que 2p+25, siendo «p» el número de variables regresoras del modelo (una en el caso de regresión simple). Este método suele recomendarse principalmente para regresión múltiple.

El algoritmo actúa sobre los valores de las variables regresoras de la forma siguiente: Los puntos, representados en el espacio de «p» dimensiones (una recta en regresión; plano en caso de dos regresoras, etc.) se tipifican para que todas las variables estén en las mismas unidades y posteriormente se ortonormalizan (en caso de regresión múltiple). Se calcula la distancia euclídea entre todas las parejas posibles de puntos. En el primer paso, se asigna al subconjunto de estimación aquella pareja de puntos que se encuentre más alejada. De los restantes, los que tengan la distancia mayor forman parte del subconjunto de valoración. Cada punto que ya haya sido asignado deja de ser tenido en cuenta en futuras asignaciones. En el tercer paso, el punto más alejado de los dos que ya forman el subconjunto de estimación, se añade a éste. La distancia de un punto a otros dos se calcula como la mínima de las dos distancias. Por tanto, el criterio de asignación es la distancia máxima de entre las mínimas. De forma similar se añade un punto al subconjunto de valoración. Este proceso se repite hasta que no queden puntos por asignar. Cuando existan puntos con la misma distancia a otros, la elección de puntos a asignar se hace al azar, entre los posibles candidatos.

Existen otros métodos que se basan en probar la bondad del modelo requiriendo el cálculo previo del ajuste. Lo que prueban, por tanto, es la bondad del ajuste de los datos experimentales al modelo propuesto. Uno de ellos es por medio del coeficiente de determinación R^2 que ya fue presentado en 4.2.

Este coeficiente expresa la parte de la variabilidad de la Y que ha sido explicada por el modelo. Se encuentra acotado entre los valores 0 y 1, y se define como

$$R^2 = \frac{SC\ (Regresión)}{SC\ (Total)} = 1 - \frac{SC\ (Residuo)}{SC\ (Total)}$$

Cuando la regresión es simple, este coeficiente equivale al de «correlación» elevado al cuadrado.

Ahora bien, existen ciertos problemas inherentes al uso del R^2. La dificultad más grave con el uso del coeficiente de determinación es que puede aumentarse ficticiamente su valor al aumentar el número de variables regresoras (en caso de regresión múltiple). Por tanto, una solución natural consiste en actuar en términos de varianzas en vez de sumas de cuadrados. Por ello, se define el coficiente de determinación corregido R_c como

$$R_C^2 = 1 - \frac{\hat{\sigma}^2}{\frac{Sy^2}{n-1}}$$

Cuando se dispone de varias observaciones de Y para unos valores dados x_i, se puede estudiar la falta de ajuste de los datos al modelo lineal. Si el modelo no es correcto, el cuadrado medio de las desviaciones no estará estimando el verdadero error experimental. Este habrá sido inflado, por el

sesgo debido a la incorrección del modelo. Mediante las medidas repetidas se puede obtener una estimación directa de la varianza del error experimental calculando una media ponderada de las variantes internas de cada grupo de datos. La descomposición para una observación particular y_{iu} es:

$$\underset{\text{Residuo}}{y_{iu} - \hat{y}_{iu}} = \underset{\substack{\text{Error}\\\text{verdadero}}}{(y_{iu} - \bar{y}_i)} + \underset{\substack{\text{Falta}\\\text{de ajuste}}}{(\bar{y}_i - \hat{y}_{iu})}$$

En donde:

y_{iu} es la observación correspondiente al u-ésimo valor repetido de x_i.

$\hat{y}_{iu}$ es la estimación por el modelo de la observación y_{iu}, que es igual para todo u.

$\bar{y}_i$ es la media de las observaciones repetidas para x_i.

La suma de cuadrados del error verdadero obtenido por las medidas repetidas se calcula de la forma siguiente:

$y_{11}, y_{12}, \ldots y_{1n_1}$ son n_1 medidas repetidas para x_1

$y_{21}, y_{22}, \ldots y_{2n_2}$ " n_2 " " " x_2

...

...

$y_{g_1}, y_{g_2} \ldots y_{gn_g}$ " n_g " " " x_g

La contribución por las observaciones en x_1 a la suma de cuadrados del error experimental verdadero es $\sum_{u=1}^{n1} (y_{1u} - \bar{y}_1)^2$.

Por lo tanto, para todas las observaciones, la suma de cuadrados del error verdadero o puro:

$$\text{SC (error puro)} = \sum_{i=1}^{g} \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2 \quad \text{con} \sum_{i=1} n_i - g$$

grados de libertad.

Por diferencia con la suma de cuadrados y grados de libertad del residuo se puede obtener las cantidades correspondientes a la línea de la falta de ajuste, según la tabla (siendo N el total de observaciones).

Fuentes de variación	*S. C.*	*g. l.*
Debido a Regresión	b_1 Sxy	1
Desviaciones	$Sy^2 - b_1$ Sxy	N − 2

Falta de ajuste	$Sy^2 - b_1 Sxy - \sum_{i}^{g} \sum_{u}^{n_i} (y_{iu} - \bar{y}_i)^2$	g-2
Error puro	$\sum_{i=1}^{g} \sum_{u}^{n_i} (y_{iu} - \bar{y}_i)^2$	N-g

Si el modelo es correcto, la división entre el cuadrado medio de la falta de ajuste y el del error puro se distribuye según una F centrada. Por tanto, se puede verificar la hipótesis de la bondad del modelo, que equivale a estudiar la linearidad del modelo.

Utilizando los datos del ejemplo de homogeneidad de varianzas:

X	Y_{iu}				$\bar{y}_i$	n_i	$\sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2$
3	120	123	118	125	121,50	4	29,00
4	128	127	132	135	130,50	4	41,00
5	147	132	139	142	140,00	4	118,00
6	192	198	210	194	198,50	4	195,00
7	201	213	205	220	209,75	4	214,75
8	300	315	320	310	311,25	4	218,75

SC (error puro) = 816,50

Con estos datos se ha efectuado el análisis de varianza de la regresión de acuerdo con la siguiente tabla:

Fuentes de variación	*S. C.*	*g. l.*	*C. M.*	*Fc*
Debido a Regresión	88572,86	1	88572,86	127,4**
Desviación de la Regresión	15287,64	22	694,89	
Total	103860,50	23		

Descomponiendo la suma de cuadrados de las desviaciones se obtiene que,

Fuentes de variación	*S. C.*	*g. l.*	*C. M.*	*Fc*
Debido a Regresión	88572,86	1	88572,86	127,4 **
Desviaciones de la Regresión	15287,64	22		
Falta de ajuste	14471,14	4	3617,78	79,75**
Error puro	816,50	18	45,36	

Por lo tanto, el modelo no ha sido correctamente especificado, ya que sale altamente significativa la falta de ajuste.

Si se obtuvo una F significativa para la hipótesis H_0: $\beta_1=0$ y una F no significativa para la falta de ajuste, esto no implica necesariamente que el modelo propuesto sea el adecuado, y sería recomendable profundizar en su estudio. Como indican DRAPER y SMITH (1981, pág. 93) a no ser que el recorrido de los valores predichos por la ecuación ajustada sea considerablemente mayor que el tamaño del error experimental, la predicción no tendrá generalmente ningún valor. Citan estos autores que, según sugerencia de WETZ (1964), para que una ecuación sea considerada como satisfactoria para predecir la F calculada en la prueba de $\beta_1=0$, debe ser mayor que cuatro vecs la F de las tablas para los grados de libertad y el nivel de significación correspondiente.

Actualmente SUICH y DERRINGER (1977, 1980) han presentado un criterio numérico más sólido, con objeto de que la predicción pueda considerarse como aceptable. Para evitar predicciones que no sean de ningún valor práctico, utilizan el criterio λ, parámetro de no centralidad de la distribución F, que cuantifica el rango de los valores predictivos relativo al tamaño de la desviación típica residual. Además propusieron una prueba del valor de λ que puede efectuarse comparando el estadístico usual F con diferentes valores críticos. Este criterio ha sido examinado comparativamente por HILL, HUDGE y FOMBY (1978, 1980).

6.4. Examen de los residuos

6.4.1. *Definición*

De una manera general, y no simplemente en el contexto de la regresión, el residuo se define como la diferencia entre el valor observado y el esperado, o estimado, dentro del marco del modelo que se utiliza. A cada observación elemental se le asocia, por tanto, un residuo (o valor residual) que, en valor absoluto, mide la mayor o menor adecuación del valor estimado al observado en el marco de la muestra tomada. Un residuo positivo implica que el valor observado es superior al estimado, mientras que un residuo negativo implica lo contrario.

En el caso de regresión simple, el residuo e_i asociado a la observación i-ésima fue definido como

$$e_i = y_i - b_0 - b_1 x_i = y_i - \hat{y}_i$$ cuyo estimador se denotará por

$$E_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = Y_i - \hat{Y}_i$$

Como es lógico, la suma de cuadrados de las desviaciones del modelo (SCE) mide globalmente la adecuación del modelo y está directamente asociada a los residuos, ya que

$$SCE = \sum_i e_i^2$$

6.4.2. *Propiedades estadísticas de los residuos*

Dado que se está interesado en el estudio de los residuos dentro de un contexto estadístico, es lógico conocer su distribución dado que son variables aleatorias. Precisamente debido a la presencia de comportamientos anómalos en su distribución, podrá ponerse en duda el modelo. Por tanto, es necesario

tener una idea clara de cuál es su comportamiento «normal»; para ello, se supone que todas las suposiciones de base se cumplen aunque, de hecho, la normalidad no sea necesaria nada más que para la distribución de los residuos.

6.4.2.1. Dependencia de los residuos

Cada ecuaciión normal (definida en 2.1) implica una restricción de tipo lineal sobre el conjunto de los residuos. En particular, la primera no es otra cosa que la condición de nulidad de su suma. Por consiguiente, el conocimiento de los $n-1$ primeros residuos, proporciona el valor del último. No son, pues, independientes (recuérdese lo indicado en 6.2.3 sobre la prueba de Shapiro y Wilk). Esto implica que, en la práctica, la información proporcionada por un residuo es parcialmente redundante con respecto a la del resto.

El número de «informaciones» independientes correspondientes al conjunto de residuos es $n-(k+1)$; es decir, el número de observaciones menos el número de parámetros del modelo (incluyendo β_0 y siendo k el número de variables regresoras). En particular, si solamente hay dos observaciones para ajustar una regresión simple, este valor es cero. Este cero está perfectamente justificado, puesto que por dos puntos siempre pasa una recta y los residuos son automáticamente nulos. Este número es importante y representa los grados de libertad del residuo.

Para que el examen individual de los residuos sea informativo es necesario que el cociente $[n-(k+1)]/n$ sea grande. El límite, como todo este tipo de límites, es difícil de marcar, pero de manera orientativa se puede decir que a partir de 0,9 es satisfactorio y que si es menor de 0,6, su interés es despreciable.

6.4.2.2. Independencia entre los residuos y los valores estimados

Esta es una propiedad general de los modelos lineales, pues la correlación entre E e Y es nula, ya que

En efecto, para regresión simple: $\sum_i E_i \hat{Y}_i = 0$.

$$\sum_i E_i \hat{y}_i = \sum_i E_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_i E_i \hat{\beta}_0 + \sum_i E_i \hat{\beta}_1 x_i =$$

$$\hat{\beta}_0 \sum_i E_i + \hat{\beta}_1 \sum_i E_i x_i = \hat{\beta}_0 0 + \hat{\beta}_1 0 = 0.$$

Así, pues, esta independencia, que lo es también en sentido estadístico, debe implicar que el reparto de los puntos P_i de coordenadas $(\hat{e}_i, \hat{y}_i)$ en una representación cartesiana, tenga una forma elíptica cuyos ejes principales sean paralelos a los ejes cartesianos.

6.4.2.3. Heteroescedasticdad de los residuos

Aunque las observaciones elementales y_i tengan, por suposición, la misma varianza, los residuos E_i (no confundir con el error aleatorio ε_i) no son homoescedásticos. Esta propiedad se entiende fácilmente si se consideran los dos puntos siguientes. Primero, la varianza de los residuos se deduce directamente de la de las estimas, pues $Y_i = \hat{Y}_i + E_i$, de donde:

$$\text{var}(Y_i) = \sigma^2 = \text{var}(\hat{Y}_i) + \text{var}(E_i)$$

ya que $\hat{Y}_i$ y E_i son independientes según la propiedad del apartado anterior. Como var ($\hat{Y}_i$) depende del punto x_i (ver 2.1.2), será variable y, por tanto, lo será var (E_i). Segundo, para algunos dominios del espacio de las variables regresoras la estimación es más precisa: aquellos en los que existen mayor número de observaciones o aquellos que están rodeados de muchas observaciones, beneficiándose de la información aportada por sus puntos vecinos (ver figura 6.1).

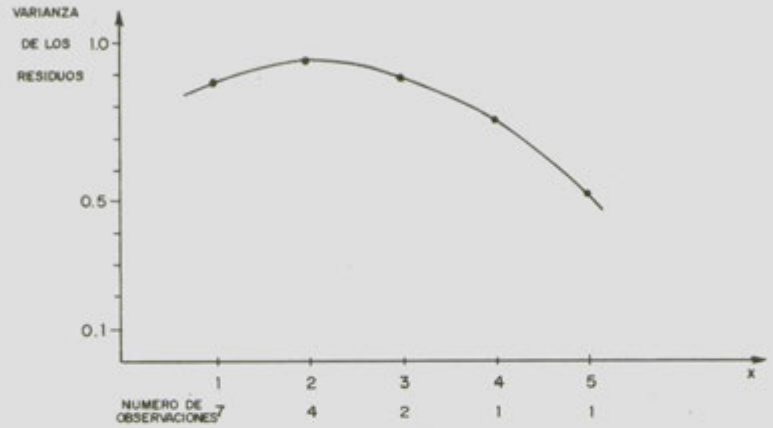


Figura 6.1.—Varianza de los residuos en función de una regresora para un modelo desequilibrado.

La varianza asociada a un residuo para regresión simple es

$$\text{var}(E_i) = \text{var}(Y_i) - \text{var}(\hat{Y}_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sx^2}\right]$$

Por tanto, será máxima en el baricentro de las variables regresoras. Los residuos con varianzas menores son los que mayor información aportan.

6.4.3. *Procedimiento de estudio*

La investigación de la distribución anormal de los residuos se hace esencialmente de manera gráfica. Se llevan sobre unos ejes cartesianos los residuos en ordenadas, y en abcisas una variable que se piense que pueda tener influencia sobre esa distribución. Cada punto corresponde a una observación elemental. La puesta en duda del modelo puede ser puntual (algunas observaciones se comportan de manera diferente a las otras) o global (la forma general de la nube de puntos es aberrante y permite suponer que el modelo no es adecuado).

En abcisas se puede representar:

a) Una de las variables regresoras con objeto de verificar que el ajuste propuesto es correcto. Por ejemplo, detectar heteroescedasticidad o autocorrelación (ver figuras 1.1, 1.2 y 1.3). Sin embargo, en el caso de regresión múltiple, el conjunto de gráficos (uno para cada variable regresora) está lejos de proporcionar toda la información necesaria para detectar los posibles fallos.

En el caso de dos variables regresoras se puede, eventualmente, proceder de la manera siguiente: representar en abcisas y ordenadas las dos regresoras y, a partir del punto correspondiente a cada observación, elevar (si el residuo es positivo) o bajar (en caso contrario) una flecha vertical de longitud igual al valor absoluto del residuo, tal como aparece en la figura 6.2. De todos modos, este procedimiento es complicado y poco informativo cuando el número de observaciones es grande.

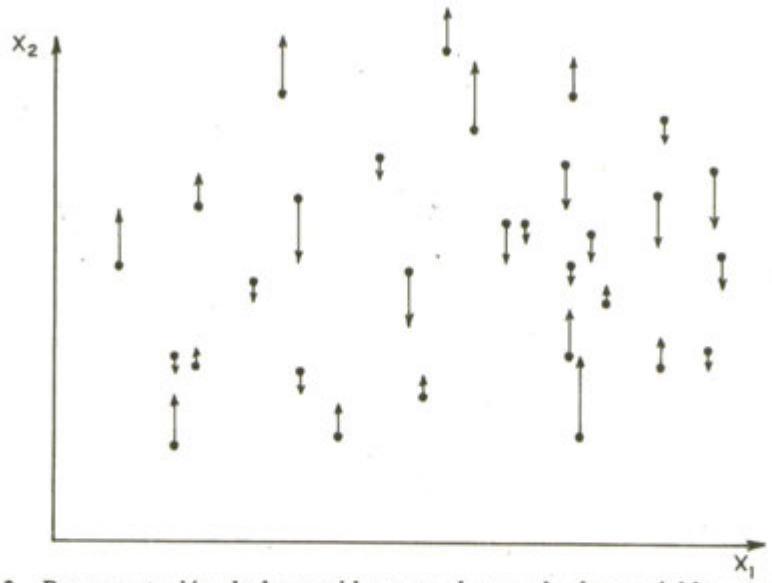


Figura 6.2.—Representación de los residuos en el caso de dos variables regresoras.

b) Una variable exterior al modelo (es decir, que no haya sido tenida en cuenta) pero que puede haber tenido influencia. Un caso particular que puede ser interesante estudiar, es cuando los individuos pertenecen a poblaciones diferentes. En este caso, se pueden representar los individuos identificándolos con un código correspondiente a cada una de las poblaciones. Un ejemplo de esta situación se presentará en el apartado 6.4.4.2 de «estructuras no manifiestas».

c) El valor estimado por el modelo. A causa de las propiedades de independencia entre los residuos y los valores estimados, esta representación es bastante sensible y permite identificar casos anómalos muy claramente. Es necesario seguir unas investigaciones posteriores para determinar cual es la causa del fallo detectado. Esta gráfica tiene la ventaja de ser única sea cual sea el número de variables regresoras. Es el método que se empleará sistemáticamente cuando no existan otras ideas a priori.

En el caso de regresión simple, los gráficos en función de los valores estimados o de la regresora son idénticos pues están relacionados linealmente por la ecuación $Y_i = b_0 + b_1 x_i$.

d) El orden de adquisición de los datos. Esta es una práctica a menudo recomendada porque el gráfico puede hacer patente los sesgos debidos a un modo de recogida de datos que se modifica a lo largo del tiempo sin advertirlo el investigador.

Hasta aquí se han presentado varios ejemplos que permiten visualizar

algunos tipos de situaciones posibles. Obviamente, no se puede dar una casuística total: en cada caso particular se debe adoptar un procedimiento particular y es el arte y la habilidad del estadístico quien debe fijarlo a partir de los elementos de que dispone.

Por una cuestión de simple pedagogía, los ejemplos que se describirán a continuación no están basados en datos reales sino en observaciones preparadas. Ello evita, por otro lado, disponer de un conjunto de datos demasiado elevado y tener que explicar el contexto de la experimentación incluyendo detalles importantes para el caso concreto, pero superfluos para el tema que se está tratando.

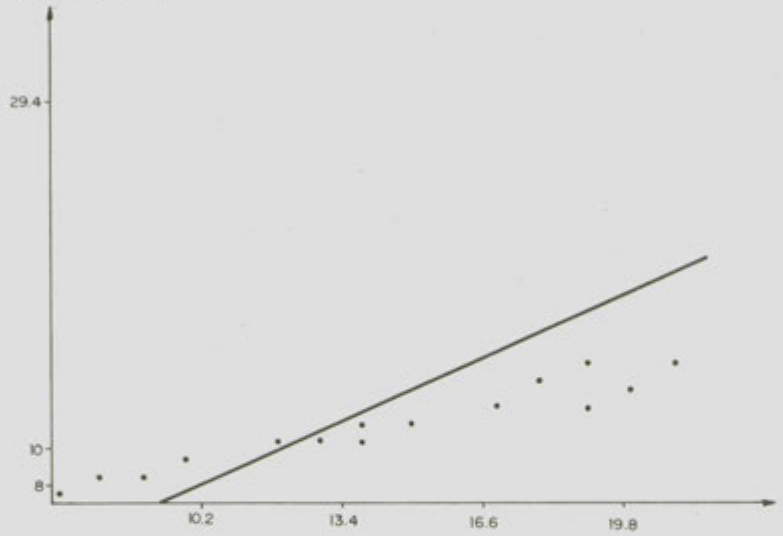


Figura 6.3.—Ejemplo de dato aberrante.

TABLA 6.1

Distribución discontinua de los residuos: Dato aberrante indicado por un asterisco.

Valores observados		*Valores predichos*	
V. dependiente	*V. regresora*	*V. dependiente*	*Residuo*
7,4	7	6,175	1,225
8,5	8	6,983	1,517
8,5	9	7,791	0,709
9,6	10	8,598	1,002
10,7	12	10,214	0,486
10,6	13	11,022	−0,422
10,5	14	11,830	−1,330
11,7	14	11,830	−0,130
11,7	15	12,638	−0,938
12,7	17	14,254	−1,554
13,9	18	15,062	−1,162
12,6	19	15,870	−3,270
15,0	19	15,870	−0,870
13,7	20	16,677	−2,977
14,9	21	17,485	−2,585
*29,4	23	19,101	10,299

6.4.4. *Distribución discontinua de los residuos*

6.4.4.1. Datos aberrantes

Detectar los valores muy aberrantes es el primer mérito del examen de los residuos. En la figura 6.3 y la tabla 6.1, el reparto de los valores de (x_i, y_i) muestra claramente que un punto se sale de la pauta general que siguen los demás. Este tipo de puntos, por lo general, no es más que el resultado de un error de transcripción de los datos. En el ejemplo, se ha tomado 29.4 en vez de 19.4. Sobre el gráfico de residuos (figura 6.4) esta situación se traduce en un residuo enorme (10.3) para el punto correspondiente, así como el comportamiento anormal de que los otros residuos son positivos primero y negativos después; comportamiento explicable, por otra parte, debido a que el dato aberrante forzaba a aumentar la pendiente de la regresión.

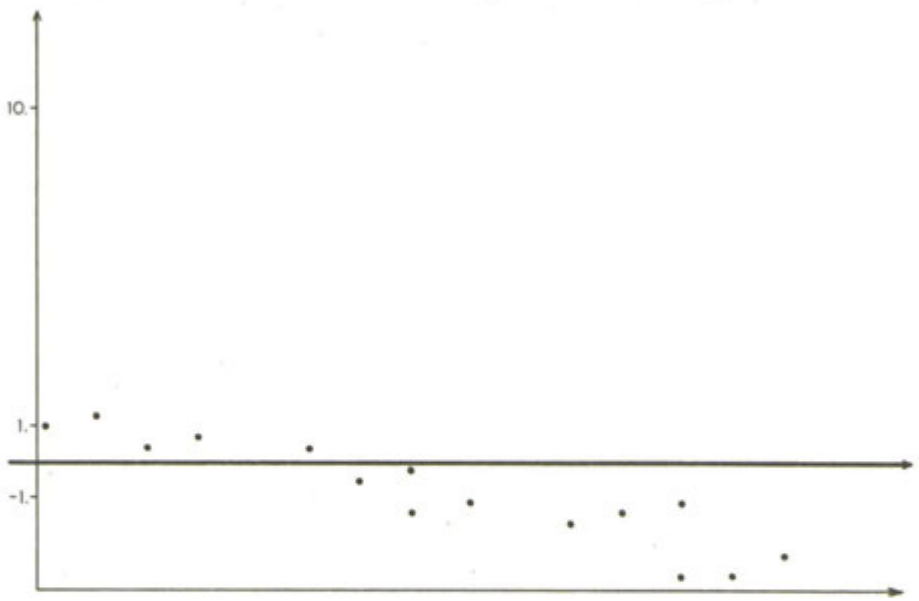


Figura 6.4.—Representación de los residuos con un dato aberrante.

El problema que se plantea en estos casos es determinar a partir de qué valor absoluto se puede considerar a los residuos como demasiado importantes, qué valores son realmente aberrantes y cuáles son indicativos de situaciones especiales y dignas de estudio por representar puntos singulares en la experimentación. No existe respuesta precisa a esta cuestión y a menudo la duda subsistirá; es un tema de estudio constante (ANDREWS y PREGIBON, 1978; COOK, 1977; ABRAHAM y BOX, 1978, entre otros) y es necesario también investigar la influencia que tienen estos datos aberrantes sobre la robustez de la regresión (ver, por ejemplo, BICKEL, 1978). En general, lo que se suele emplear para una comparación más fácil entre regresiones es utilizar los residuos tipificados, es decir, divididos por la estima de su desviación típica. Esta puede ser una precaución indispensable, pues ya se ha visto en 6.4.2.3 que su varianza difiere de acuerdo con la disposición de las variables regresoras. En la práctica, si esta disposición no es demasiado especial (por presencia de colinearidad, por ejemplo) y el número de parámetros de la regresión es pequeño en relación al de observaciones, para mayor sencillez,

se puede simplemente dividir por la estimación de la desviación típica de Y. Una vez tipificados, se considera que, aproximadamente, se comportan como una muestra de una normal tipificada. En consecuencia, todos aquellos que se encuentran fuera del intervalo [−2,5; +2,5] (que representa aproximadamente el 98,8 por 100 de la distribución) se consideran como aberrantes.

Una vez identificados los elementos aberrantes, conviene preguntarse si su eliminación no sesga la muestra; es decir, si son verdaderamente el resultado de un error en la toma o transcripción del dato. En otros casos, conviene investigar sobre el modelo las desviaciones sistemáticas (ver 6.4.4).

Otro ejemplo de datos que son aberrantes pero que corresponden a una causa concreta identificable se presenta en la figura 6.5. Se trata de la observación del grado de ataque de una enfermedad a lo largo de una línea de un campo que ha sido tratado uniformemente por un pesticida. Cada observación corresponde a la anotación sobre un metro lineal. Se ha efectuado la regresión del ataque en función del emplazamiento sobre la línea para estudiar si existía o no un gradiente.

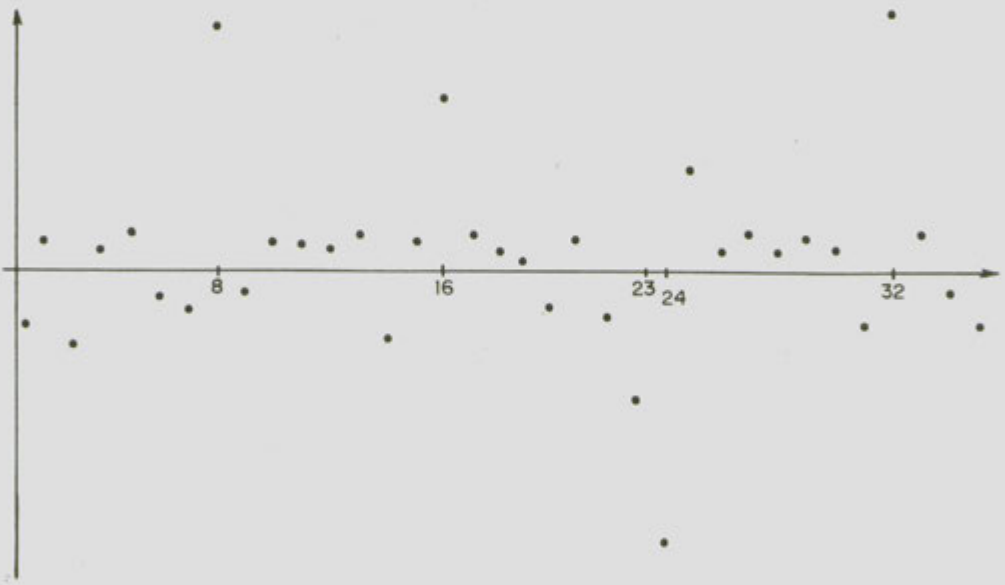


Figura 6.5.—Ejemplo de ataque de una enfermedad.

El examen del gráfico de residuos en función de los valores de la variable regresora, muestra la presencia de cuatro o cinco residuos demasiado grandes en los 8, 16, 23, 24 y 32 metros. Todos ellos con positivos excepto el 23 y 24. La explicación fue simple y se descubrió al hablar con el experimentador: el insecticida se esparció por tramos de 8 metros y los valores de ataque fuerte (residous altos) corresponden a lugares que no fueron tratados mientras que los residuos negativos son aquellos en donde hubo recubrimiento entre dos pasadas.

TABLA 6.2

Contorno de la cadera (y) y anchura de las espaldas (x) según el sexo. Estudio de los residuos (e).

Sexo	y_i	x_i	e_i
2	116	35,0	13,396
2	114	37,5	10,509
2	112	38,0	8,331
2	111	36,0	8,041
2	111	34,0	8,751
1	109	41,0	4,266
1	108	43,0	2,556
1	107	45,5	0,669
2	107	36,0	4,041
2	107	37,0	3,686
1	105	44,0	−0,799
1	105	37,0	1,509
2	104	34,5	1,573
2	103	36,5	−0,136
1	103	41,0	−1,734
1	100	40,5	−4,556
2	100	35,0	−2,604
2	100	34,0	−2,249
1	99	38,5	−4,846
2	99	32,5	−2,717
1	97	41,0	−7,734
2	96	34,0	−6,249
2	93	33,0	−8,894
1	91	38,0	−12,669
1	91	35,5	−11,781

6.4.4.2. Estructuras no manifiestas

En la tabla 6.2 se incluyen los datos del contorno de la cadera y la anchura de la espalda de un cierto número de individuos. Se intenta por regresión lineal explicar el contorno de la cadera (Y) por la anchura de la espalda (x). La ecuación obtenida es $y=90{,}18+0{,}35\ x$ con una suma de cuadrados de las desviaciones, SCE, de 1104,8. La regresión calculada no parece demasiado útil pues, si se ajusta la ecuación sin la variable, se obtiene $y=103{,}53$ con $SCE=1142{,}2$.. Sin embargo si se examina el gráfico de residuos en función del valor estimado y además se añade la información suplementaria del sexo al que pertenecen los individuos, está claro que la distribución de machos y hembras está lejos de ser al azar (figura 6.6). Conviene, por tanto, hacer intervenir el sexo calculando por ejemplo, una recta de regresión para cada sexo por lo que se obtiene

$$y_m = 90 + 2{,}7\ x_m$$

$$y_h = 49{,}6 + 1{,}3\ x_h$$

reduciendo la SCE a 597,5 (ver capítulo de modelos inclusivos para comparar ambas ecuaciones de regresión).

El objetivo de este sencillo ejemplo es poner de manifiesto que el olvido de un factor en el modelo puede conducir al rechazo de una relación entre las dos variables estudiadas. Más aún, si la influencia del factor sexo hubiera sido más marcada, este olvido podría haber producido una relación signi-

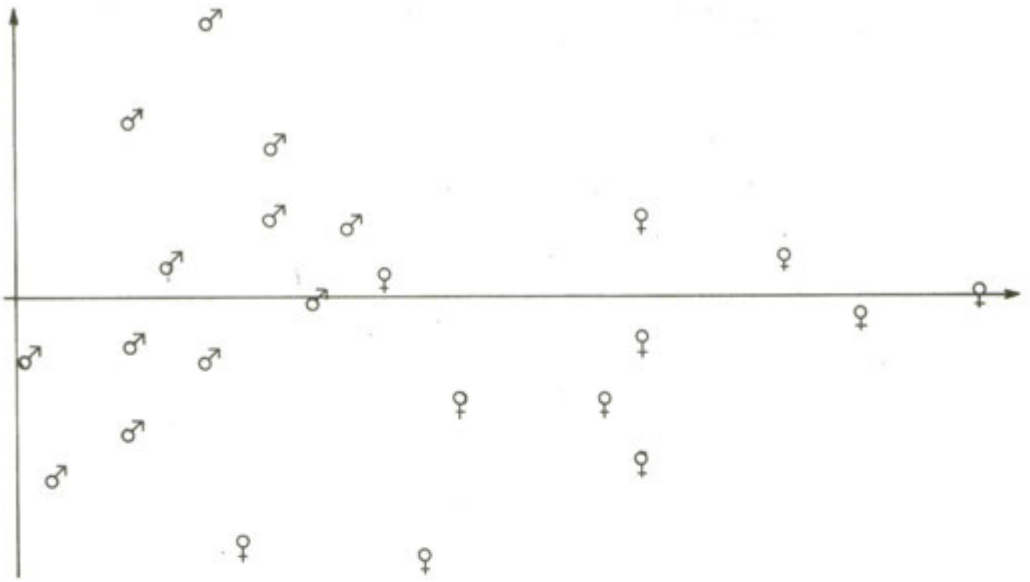


Figura 6.6.—Representación de los residuos del contorno de cadera y anchura de espalda por sexos.

ficativa pero de signo opuesto a la real, tal como muestra esquemáticamente la figura 6.7.

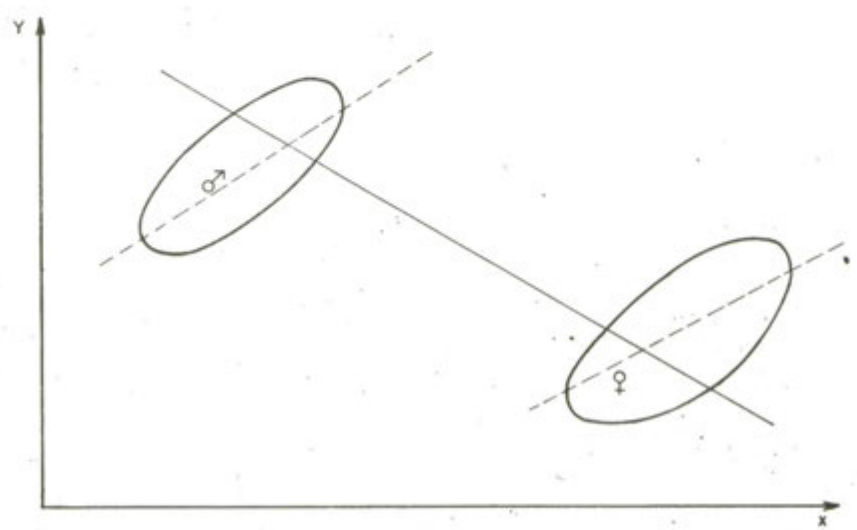


Figura 6.7.—Consecuencias de no tener en cuenta la estructura en los datos. (Trazo continuo: regresión dada por los datos; línea de puntos. regresión correcta.)

Existen diversas maneras de hacer intervenir la estructura (indicada por «s» en las ecuaciones siguientes) en un modelo de regresión:

La estructura sólo interviene a nivel de media de grupos:

$$(a) \quad Y = \beta_{0s} + \beta_1 x + \varepsilon$$

La estructura interviene exclusivamente en la manera de actuar la variable regresora

$$(b) \quad Y = \beta_0 + \beta_{1s} x + \varepsilon$$

La estructura actúa sobre ambos componentes

$$(c) \quad Y = \beta_{0s} + \beta_{1s} x + \varepsilon$$

Estas tres situaciones están esquematizadas en la figura 6.8 y muy relacionadas con la figura 5.2.

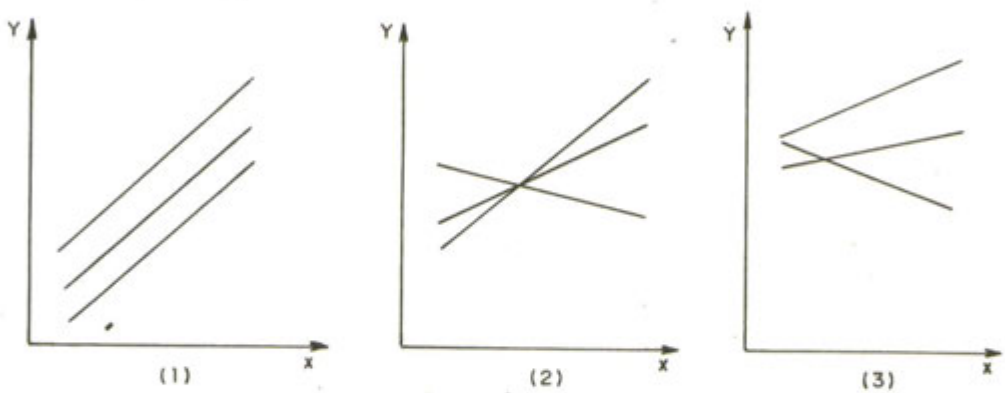


Figura 6.8.—Diferentes modelos de regresión según la estructura.

Es obio que la generalización de este tipo de modelos se puede extender a ajustes con varios regresores y a regresión polinomial.

6.4.5. *Distribución continua de los residuos*

6.4.5.1. Varianza no constante: Transformación de variables

Una de las hipótesis de base del modelo es la homogeneidad de las varianzas. Existen casos en los que claramente esta hipótesis no se cumple. Muy a menudo la varianza de las observaciones está ligada a la media (figura 1.1); por ejemplo, si se trabaja con porcentajes y son muy próximos a 0 ó 100 por 100, fuerzan a varianzas pequeñas. Otro caso se presenta al efectuar enumeraciones o conteos: cuanto mayor sea la frecuencia absoluta, mayor será la varianza. El examen de los residuos en función del valor estimado dará claramente una nube de puntos en forma de trompeta. En el ejemplo de la tabla 6.3, se estudia la población de un insecto sobre las tomateras en invernadero haciendo variar la temperatura del invernadero. La figura 6.9 indica claramente que a mayor nivel de población, mayor es la dispersión de las desviaciones del modelo (residuos).

TABLA 6.3

Población de insectos sobre tomateras en invernaderos (y) en función de la temperatura (x). Estudio de los residuos (e).

y_i	x_i	e_i
650	28,5	−175,045
980	28,5	186,627
800	27,5	38,299
620	26,5	−78,358
800	26,0	133,313
500	25,5	−135,015
750	24,5	178,328
610	24,5	38,328
390	24,0	−150,000
625	23,0	148,343
280	22,0	−133,313
380	22,0	−33,313
380	21,0	30,030
250	20,0	−36,627
260	19,5	5,045
140	19,0	−83,284
50	17,5	−78,269
125	17,0	28,403
70	16,0	36,746
50	15,0	80,089

Figura 6.9.—Ejemplo de la relación población de insectos-temperatura.

En este caso, tal como se recomienda en 6.2.2 una simple transformación logarítmica de los datos permite remediar el problema tal como aparece en la figura 6.10. El modelo finalmente ajustado es pues:

$$\text{Log } Y = \beta_0 + \beta_1 x + \varepsilon$$

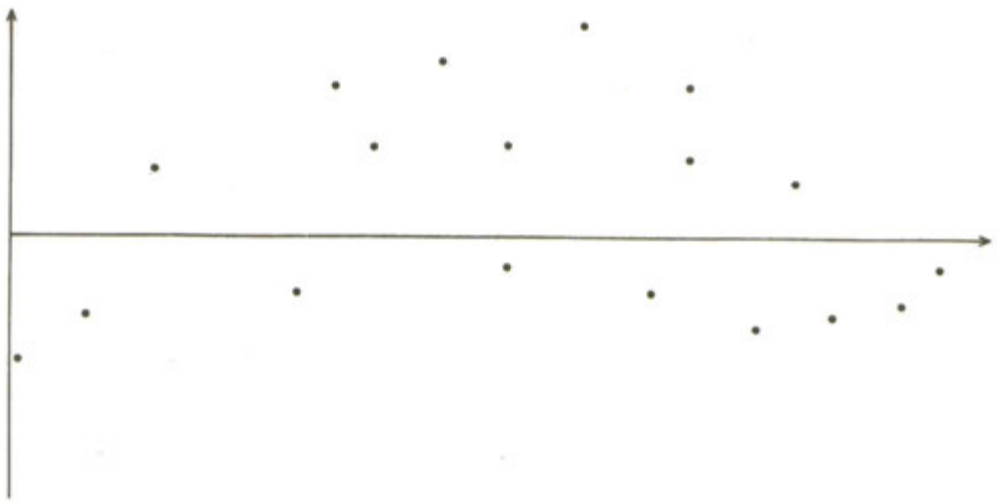


Figura 6.10.—Relación logaritmo de la población-temperatura.

Otras transformaciones no mencionadas anteriormente y susceptibles de ser utilizadas para los casos de dispersión sistemática son

$$Y = \beta_0 + \beta_1 \log x + \varepsilon$$

$$Y = \beta_0 + \beta_1 (1/x) + \varepsilon$$

$$\text{Log } Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

La forma de los ajustes a los que dan lugar se presentan en la figura 6.11.

6.4.5.2. Pruebas consecuetivas: Ajustes polinómicos

Si el modelo es correcto, los residuos son independientes de los valores estimados. Esto implica, entre otras cosas, que el signo de los residuos se distribuye en secuencias aleatorias. Existen tablas para practicar una prueba para la alternancia de signos (SIEGEL, 1956) y otros procedimientos no paramétricos (LEHMANN, 1975). Sin embargo, en los casos más obvios, no hay necesidad de recurrir a ellos. Sean los datos de la tabla 6.4. En ella se incluyen los rendimientos del trigo en función del abonado nitrogenado de 28 parcelas experimentales. Los residuos en función de los valores esperados se representan en la figura 6.12, claramente se observa que siguen una forma de parábola y que el número consecutivo de signos se reduce prácticamente a tres: negativos, positivos, negativos. Esto es una indicación ma-

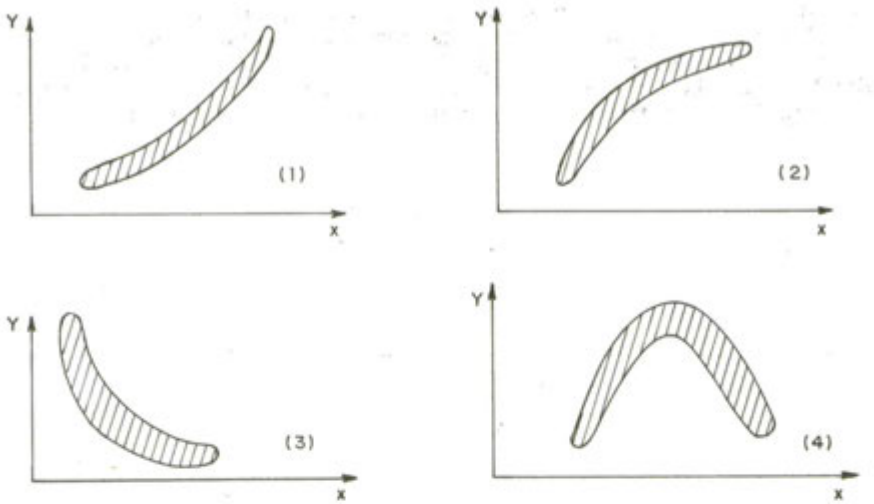


Figura 6.11.—Representación gráfica de diversos ajustes según transformaciones simples.

TABLA 6.4

Rendimiento del trigo (y) según el abonado nitrogenado (x). Residuos obtenidos en el ajuste lineal (e^l) y en el cuadrático (e^q).

y_i	x_i	e^l_i	e^q_i
39	120	−12,314	−1,986
45	120	−8,314	1,014
43	130	−10,238	−4,643
48	130	−5,238	0,357
53	140	−2,162	−0,512
58	140	2,838	4,488
56	150	−1,086	−2,591
61	150	3,914	2,409
63	160	3,990	0,120
65	160	5,990	2,120
66	170	5,066	−0,380
68	170	7,066	1,620
70	180	7,141	0,910
71	180	8,141	1,910
69	190	4,217	−2,011
71	190	6,217	−0,011
71	200	4,293	−1,143
73	200	6,293	0,857
71	210	2,369	−1,485
72	210	3,369	−0,485
70	220	−0,555	−2,038
72	220	1,445	−0,038
69	230	−3,479	−1,801
73	230	0,521	2,199
67	240	−7,403	−1,775
71	240	−3,403	2,225
66	250	−10,327	0,041
68	250	−8,327	2,041

nifiesta de un sesgo en el modelo, y la forma parabólica de los residuos induce claramente a ensayar la introducción del término x^2 en el modelo. Se pasa así de una ecuación $y=28{,}2+0{,}2x$ con una SCE de 1015, a otra $y=6100{,}5+1{,}7x-0{,}004x^2$ con una SCE de 107. Los residuos en este segundo caso, dibujados en la figura 6.13 toman una forma más ortodoxa.

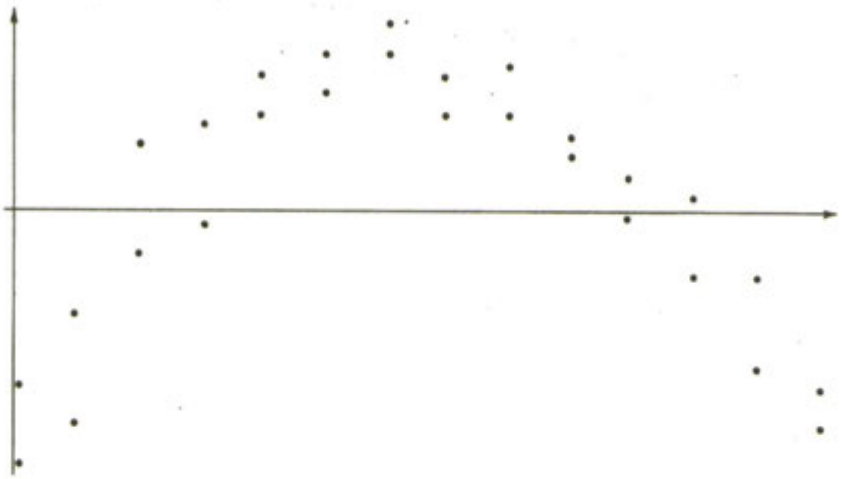


Figura 6.12.—Relación lineal entre el rendimiento y la cantidad de nitrógeno aportada (residuos).

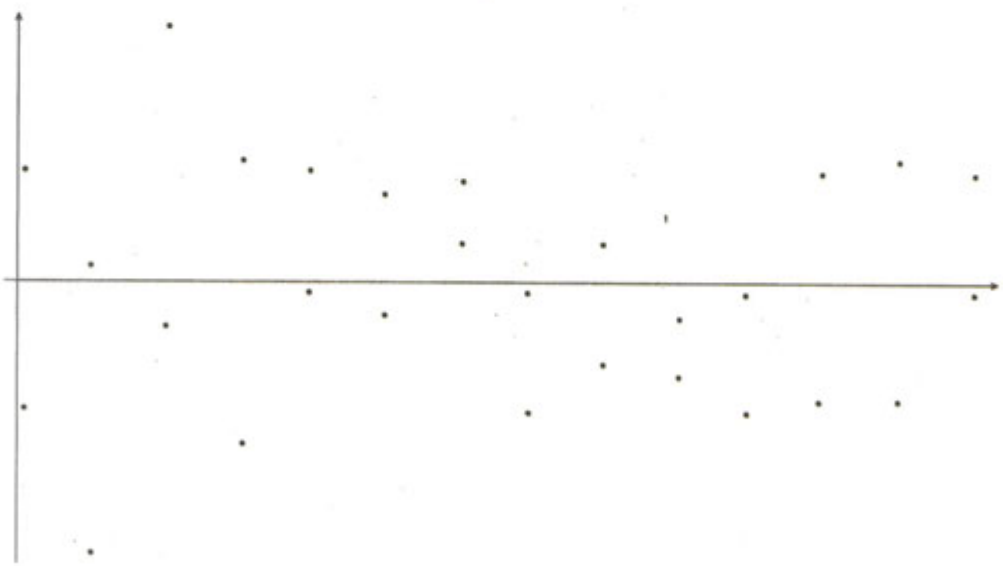


Figura 6.13.—Relación cuadrática entre el rendimiento y la cantidad de nitrógeno aportada (residuos).

Generalizando esta práctica, es a veces necesario realizar ajustes polinomiales de grado superior a dos (es decir, introducir igualmente x^3, x^4...). En el límite, se puede llegar hasta un grado g-1 siendo «g» el número de valores diferentes que toma la variable regresora. En el caso extremo, cuando se dispone de medidas repetidas, el análisis es idéntico (para grados de libertad, sumas de cuadrados, estimas y residuos) al análisis de varianza con un factor donde los niveles del factor son los diferentes valores que toma la variable regresora. Esta es una de las razones por las que las técnicas de regresión y análisis de varianza están muy próximas, pudiéndose englobar dentro de un marco más general constituido por la teoría de los modelos lineales.

CAPITULO 7

INTERPRETACION GEOMETRICA DE LA REGRESION

7.1. Introducción

La descomposición de un movimiento en el espacio en tres movimientos lineales representados por ejes perpendiculares, por un lado, y el sentido físico de ortogonalidad (o perpendicularidad) como independencia por otro (recuérdese que la fuerza producida en dirección ortogonal al movimiento puede considerarse como independiente de éste), son la clave de la representación de la geometría analítica. Además mediante la traducción Independencia-Ortogonalidad, permiten la asociación Estadística-Geometría con el consecuente acopio de conocimientos de ésta que son útiles en aquella.

Por ejemplo, la noción geométrica de dimensión es equiparable a la estadística de grados de libertad, ya que en ambos casos se trata únicamente de restricciones lineales, bien en un movimiento (Geometría), bien en la realización de una variable aleatoria (Estadística). De este modo, si se intenta reconstruir el vuelo de un pájaro a través de su proyección en el suelo, falta una dimensión o grado de libertad, de la misma manera que si se trata de reconstruir una muestra de dos elementos conociendo sólo su media.

Por otra parte, el concepto de producto escalar está inspirado en el efecto que produce una fuerza $\vec{F}$ libre en el plano sobre un cuerpo cuyo movimiento está restringido a una recta r como, por ejemplo, el movimiento de una barca que remonta un río empujada desde las orillas (figura 7.1) y resulta ser $\vec{F}\cos(\vec{F}, \vec{r})$. De una manera análoga, la explicación lineal de la variable aleatoria normal X sobre la variable aleatoria normal Y es $Y\,\mathrm{cor}(Y,X)$ en donde, como se aprecia, el coseno ha sido sustituido por la correlación.

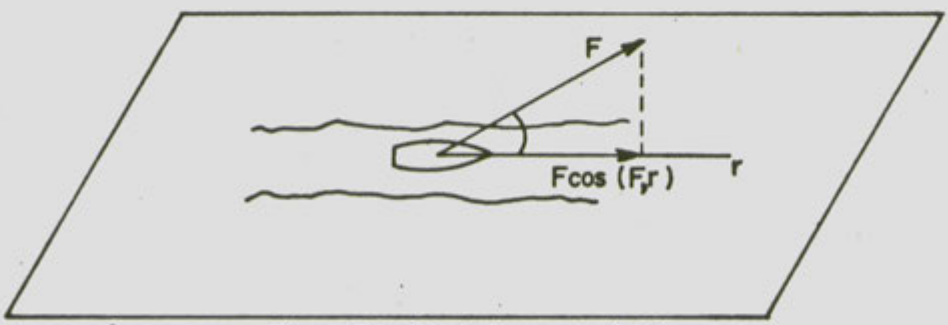


Figura 7.1.—Movimiento restringido.

7.2. Representación general de las variables aleatorias

Una variable aleatoria es por definición impredecible, pero debe cumplir cierta regularidad: las frecuencias relativas de cada posible resultado deben

converger y al valor límite se le llama probabilidad de ese resultado. Es importante destacar esta condición, ya que sin conergencia no puede hablarse de probabilidad y por tanto, no tiene sentido aplicar un modelo estadístico que se base en una estructuración de dichas probabilidades.

Cuando el número de posibles resultados es infinito real, es decir, existe una biyección con los números reales, y la variable aleatoria es continua, es necesario reconocer que las probabilidades deben darse sobre intervalos. Para entender mejor esta idea, considérese la variable tiempo. Si se dispone de un reloj con minutero y antes de mirar la hora se expresa la confianza de que sean, por ejemplo, las 9 h. 36' diciendo que existe probabilidad 2/3 de que esto ocurra, se está en realidad expresando la confianza respecto del intervalo que comienza en las 9 h. 35' 30" y termina en las 9 h. 36' 30" que es el que visualmente se identifica con las 9h.36'. La variable es, por naturaleza, continua ya que es posible, al menos en la imaginación, suponer que el intervalo podría hacerse tan pequeño como se quisiera (si el reloj tuviese segundero, el intervalo tendría amplitud de un segundo, etc.) y el concepto de probabilidad estará siempre aplicado a estos intervalos.

No obstante, resulta complicado expresar matemáticamente dichas probabilidades, ya que para ello se necesita una función de dos variables (extremo superior e inferior del intervalo). Además, sería interesante disponer de una función de punto que indicase de alguna forma la contribución de dicho punto a la probabilidad total; el modo de actuación es el mismo que se utiliza para la definición de derivada: se toman intervalos centrados en el punto de amplitud cada vez más pequeña (más precisión de medida) y se divide la probabilidad de cada intervalo por su amplitud, calculando el límite cuando ésta tiende a cero. Al resultado se le llama «densidad de probabilidad» en ese punto y permite, conociendo su valor en todos los puntos (función de densidad f), reconstruir la probabilidad en los intervalos mediante el proceso de integración (inverso al de derivada). Se tiene, en consecuencia, que la probabilidad en el intervalo A es:

$$\text{Prob}(A) = \int_A f(y)\,dy$$

cuya representación gráfica se presenta en la figura 7.2. Esta asimilación entre área (integral) y probabilidad es la misma que se realiza en la representación mediante histogramas, lo que da pie a otra obtención intuitiva de f (y) (ver fig. 7.3).

Generalizando del plano al espacio, puede considerarse ahora el proceso de tomar una muestra aleatoria independiente (m. a. i.) de tamaño tres (el

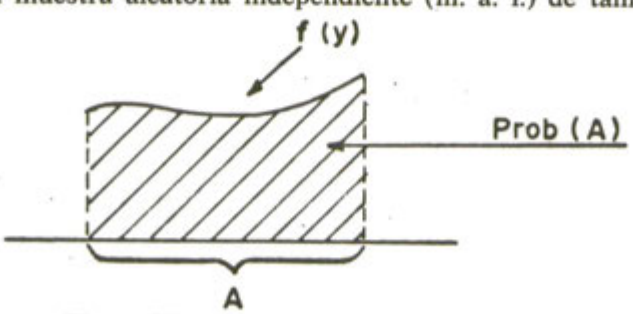


Figura 7.2.—Probabilidad de un intervalo.

Figura 7.3.—La función de densidad como límite de histogramas de amplitud decreciente.

razonamiento es ampliable a cualquier tamaño) y representar, de acuerdo con la idea ya expresada, la variable aleatoria y_i (extracción i-ésima) en el eje cartesiano i-ésimo (i=1, 2, 3). Un punto en el espacio $(y_1y_2y_3)$ será una realización concreta de la variable tridimensional $(Y_1Y_2Y_3)$, o bien una m. a. i. triple de la variable Y teniendo una densidad de probabilidad f $(y_1y_2y_3)$.

De esta forma, cada punto tiene asignado un valor que sirve para calcular las probabilidades de «zonas» del espacio a través de la integración de la función f; por ejemplo, para la zona Ω la probabilidad es

$$\text{Prob } (\Omega) = \int_{\Omega} f(y_1y_2y_3)\ dy_1dy_2dy_3$$

según se muestra en la figura 7.4.

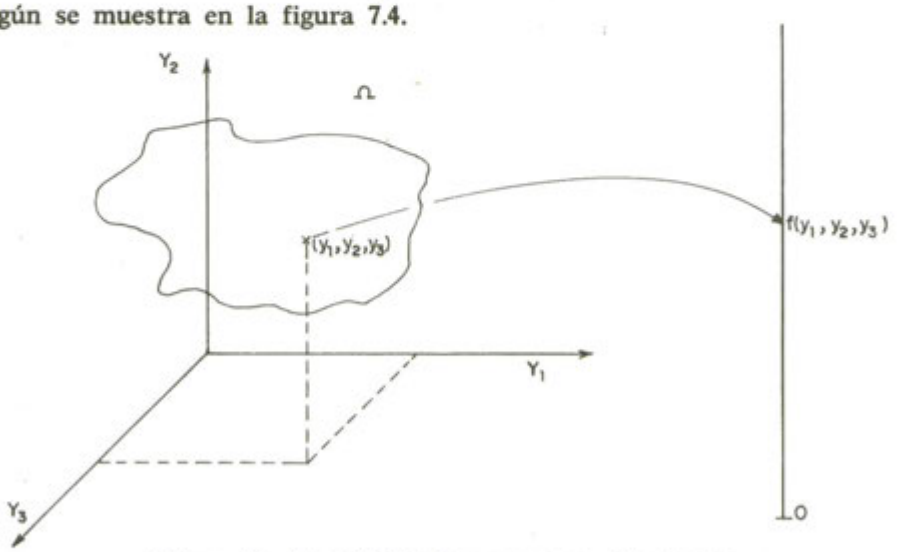


Figura 7.4.—Probabilidad en una zona del espacio.

Esta es la situación que se considera previa a la realización de la variable aleatoria y por razones prácticas conviene «resumir» la información de que se dispone. Para ello, el procedimiento usual consiste en el cálculo de puntos y zonas con especial significación; es decir, realizar una represen-

tación cartográfica a través de la elección de un punto central; por ejemplo, la esperanza, y curvas (o superficies, etc., según la dimensión del problema) de nivel de la función de densidad que correspondan a zonas de confianza conocida: 90 por 100, 95 por 100, 99 por 100 (es decir, que enmarquen conjuntos con probabilidades 0,9; 0,95; 0,99...). Realizada esta representación, que es un resumen bien de conocimientos teóricos sobre la variable en estudio o bien de experimentos previos, se obtiene mediante una muestra la realización concreta de la variable aleatoria, formulándose entonces preguntas del tipo siguiente:

¿Es coherente esta realización con la estructura probabilística que se había construido?

Supuesto que dicha estructura flotante no estaba del todo determinada, ¿cómo quedaría después de la información aportada por la realización de la variable aleatoria?

Estas dos preguntas son básicas en Estadística y en un aspecto parcial de la segunda se concretará más adelante (ver 7.4.1.).

7.3. Representación de la variable aleatoria normal: Propiedades

Frecuentemente se utiliza en Estadística Aplicada la suposición de normalidad sobre alguna de las variables en estudio. Ya se mencionó en el capítulo anterior en qué circunstancias puede aceptarse tal suposición; en este apartado se centrará el estudio en el por qué de la sencillez de los resultados obtenidos cuando se supone normalidad.

Principalmente este aspecto de sencillez se explica a través de la «forma» de la función de densidad normal. Así, para una normal de media cero y varianza unidad la función de densidad es

$$f(z) = (2\pi)^{-1/2} \exp(-z^2/2)$$

Esta «forma» da lugar a que la independencia muestral (producto de densidades de los elementos de la muestra) produzca un exponente que es una suma de cuadrados relacionada con el módulo del vector, que representa la muestra; es decir,

$$f(\vec{z}) = \prod_{i=1}^{n} f(z_i) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp(-z_i^2/2) = (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^{n} z_i^2/2\right) =$$
$$= (2\pi)^{-n/2} \exp(-||z||^2/2)$$

Esta expresión implica que las curvas de nivel correspondientes a variables normales independientes y tipificadas corresponden a los lugares geométricos de los puntos cuyas normas de sus vectores asociados sean iguales; es decir, se trata de n-esferas (esferas de n dimensiones) centradas en el origen tal como se indica en la figura 7.5.

Supóngase a continuación una situación más general a través del cambio de variables $y_i = Z_i\sigma + \mu$ para todo $i = 1, \ldots, n$, lo que significa considerar va-

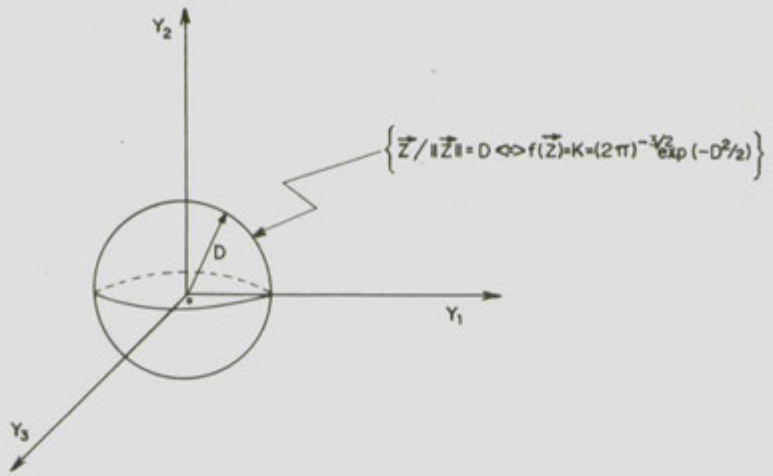


Figura 7.5.—Representación de las curvas del nivel correspondiente a distribuciones normales tipificadas.

riables aleatorias normales de media diferente μ_i y varianza igual (homoescedásticas) σ^2. En este caso, la función de densidad será:

$$f(y_i) = \sigma^{-1}(2\pi)^{-1/2}\exp\left[-(y_i-\mu_i)^2/2\sigma^2\right]$$

que proporciona vectorialmente una función de verosimilitud que se puede representar por

$$f(\vec{y}) = \sigma^{-n}(2\pi)^{-n/2}\exp(-||\vec{y}-\vec{\mu}||^2/2\sigma^2)$$

Con lo cual, las n-esferas están centradas en μ y el radio es diferente al caso anterior (fig. 7.6). Esta representación corresponde al modelo de regresión en el que se cumplen las suposiciones de base y en donde $\mu_i = x_i b$. Por tanto, es la que se empleará a lo largo de este capítulo.

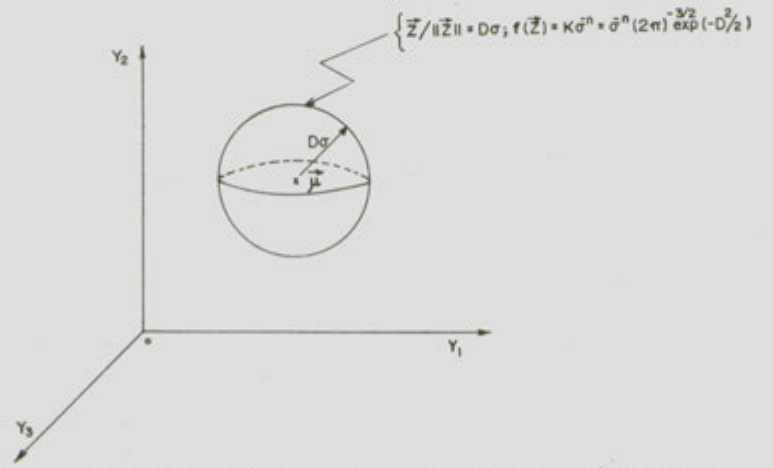


Figura 7.6.—Representación de las curvas de nivel para distribuciones normales homoescedásticas.

7.4. Regresión simple

7.4.1. *Representación de las hipótesis del modelo*

Para representar geométricamente las hipótesis del modelo se han indicado ya todos los elementos necesarios: En un modelo de regresión simple las hipótesis puede resumirse en que Y_i son variables aleatorias independientes, tales que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \mu_i + Z_i \sigma$$

en donde $\mu_i = \beta_0 + \beta_1 x_i$ y Z_i es una variable aleatoria normal con media cero y varianza unidad. Por tanto, se llega al caso analizado al final del apartado 7.3 con la única peculiaridad de que $\mu = \beta_0 1 + \beta_1 x$ debe pertenecer al plano L_x engendrado por 1 y x teniendo en él unas coordenadas β_0 y β_1 desconocidas. Estas coordenadas son, junto con el valor de σ, lo que se está interesado en estimar (fig. 7.7).

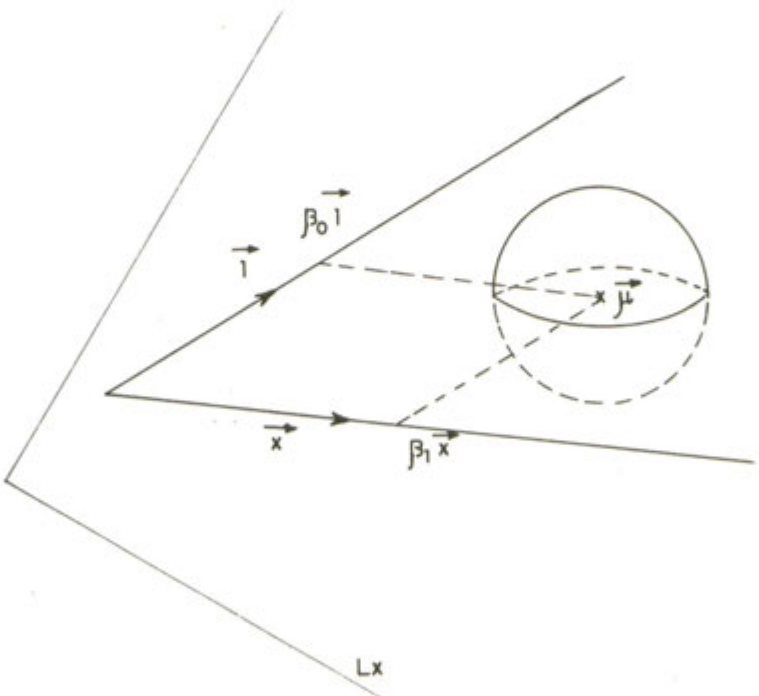


Figura 7.7.—Representación de las hipótesis del modelo de regresión.

Es necesario recalcar que lo que se está efectuando no es sino un caso particular de lo indicado al final del apartado 7.2 en relación con las preguntas que cabía formularse al obtener una realización concreta del experimento: Se tiene información previa sobre la estructura de la variable aleatoria, dada por curvas de nivel n-esféricas con radio desconocido (que depende de σ) centradas en un punto, también desconocido, «móvil» sobre un plano (que quedará fijado al conocer β_0 y β_1); por tanto, se utilizará la información proporcionada por la muestra para estimar los parámetros desconocidos σ,

β_0, β_1, quedando así perfectamente determinada la variable aleatoria. Este concepto conduce, pues, al proceso de estimación.

7.4.2. *Visualización del proceso de estimación*

Se estudiará primeramente la estimación de β_0 y β_1, empleando posteriormente otro mét › para σ.

El procedimiento a seguir, denominado máxima verosimilitud, se basa en un razonamiento intuitivo sencillo, pues se trata únicamente de tomar como valor del parámetro aquél que atribuya mayor densidad de probabilidad a la muestra de que se disponga; es decir, el que hace «más probable» a priori aquello que ha ocurrido a posteriori.

En la estimación de $\mu=\beta_0+\beta_1 x$ se concreta en obtener

$$\max\ \sigma^{-n}(2\pi)^{-n/2}\exp\left[-(\vec{y}-\vec{\mu})^2/2\sigma^2\right]$$

$$\vec{\mu} = \beta_0+\beta_1\vec{x}$$

lo que equivale a $\min\left|\left|\vec{y} - \vec{\mu}\right|\right|^2$

$$\vec{\mu} = \beta_0+\beta_1\vec{x}$$

Este valor mínimo se obtiene en el punto $\vec{m}$, proyección ortogonal de y sobre el plano engendrado por $\vec{1}$ y $\vec{x}$ tal como se muestra en la figura 7.8. Por tanto, en este caso, un proceso de estimación con significado exclusivamente estadístico como el de máxima verosimilitud se ha convertido en un mecanismo geométrico de proyección a través de un paso intermedio de minimización de la norma del error $\vec{\varepsilon}=\vec{y}-\vec{\mu}$ conocido como el método de estimación de mínimos cuadrados. Es decir, el punto central que era «móvil» sobre un plano ha sido fijado en la proyección de la realización de la variable aleatoria sobre dicho plano. Es necesario destacar que $\vec{m}$ es la representación vectorial de $\hat{y}=b_0+b_1x$.

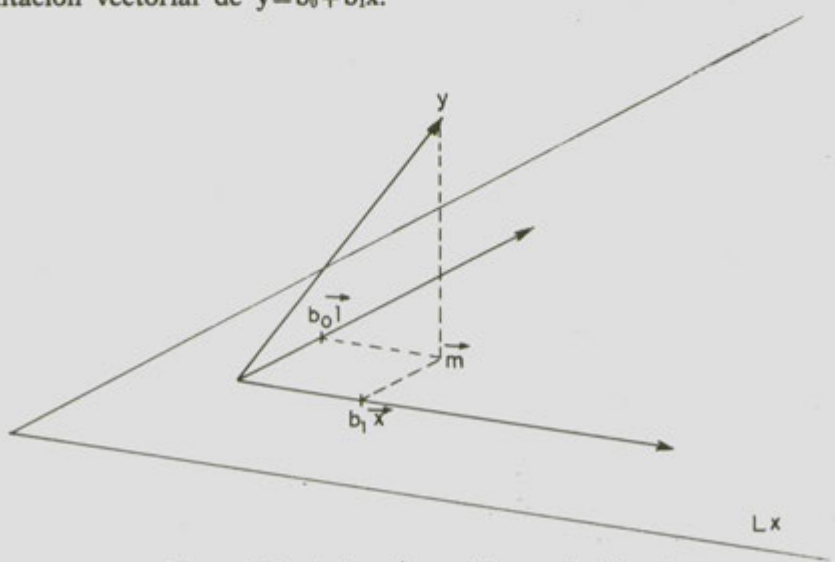


Figura 7.8.—Estimación mínica cuadrática de y.

Para estimar σ, considérense las n-esferas centradas en μ (valor auténtico, aunque desconocido para la media) y tómese, imaginariamente, aquélla que contiene al vector $\vec{y}$ de observaciones (ver fig. 7.9), y que será la formada por los extremos de los vectores de norma igual a la de $\vec{\varepsilon}$, ya que $\vec{y}=\vec{\mu}+\vec{\varepsilon}$. A continuación se utilizará el método usual para obtener estimadores insesgados, que consiste en igualar la realización a su esperanza. En este caso se calculará el radio cuadrado medio de n-esferas y se igualará al de la observación muestral.

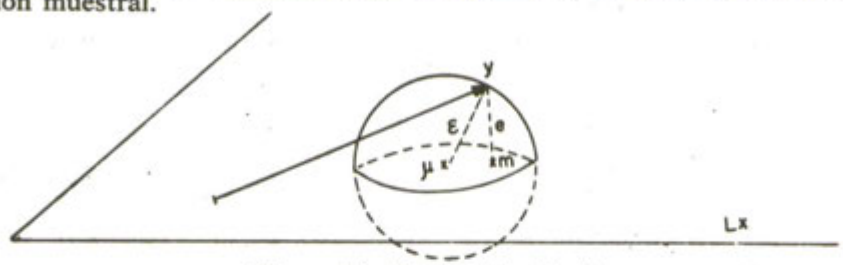


Figura 7.9.—Estimación de σ^2.

El problema, a priori, no parece complicado, ya que como $\|\vec{\varepsilon}\|^2/\sigma^2$ sigue una distribución X^2_n, se tiene que $E(\|\vec{\varepsilon}\|^2)=\sigma^2 n$, pero la dificultad surge al tener en cuenta que no se dispone del valor de $\vec{\varepsilon}=\vec{y}-\vec{\mu}$, ya que $\vec{\mu}$ es desconocida, por lo que hay que ceñirse a $\vec{e}=\vec{y}-\vec{m}$.

Es fácil, sin embargo, utilizando el Teorema de Cochran, demostrar que $E(\|\vec{e}\|^2)=\sigma^2(n-2)$, donde la reducción en grados de libertad se produce por la restricción en el «movimiento» de e (recuérdese lo expresado en la introducción), que es una proyección de ε en el subespacio perpendicular al plano L_x (en la fig. 7.9 con $n=3$ sólo hay un grado de libertad para e reflejado en el movimiento vertical).

Se tendrá, por consiguiente, igualando los radios cuadráticos observado y medio como antes se dijo: $\|\vec{e}\|^2=\sigma^2(n-2)$, de donde $\sigma^2=\|\vec{e}\|^2/(n-2)$, con lo que se obtiene una estima de σ, cuya varianza es $2\sigma^4/(n-2)$.

7.4.3. *Transmisión de la aleatoriedad desde Y a los estimadores*

Al depender el resultado de la estimación del valor particular de la realización, es claro que los estimadores serán variables aleatorias cuya función de densidad dependerá de la Y y del método concreto de estimación que se haya empleado. La aleatoriedad de Y viene representada por las n-esferas centradas en μ, y a partir de ella se llega a la del estimador $\hat{\beta}=(\hat{\beta}_0, \hat{\beta}_1)$, según se explica a continuación.

El procedimiento, fácilmente generalizable a dimensiones mayores que 3, consiste en realizar el mismo proceso de proyección utilizado anteriormente, obteniéndose circunferencias centradas en m y contenidas en el plano L_x (esto es posible por tratarse de una distribución normal, lo que implica que sus marginales son también normales). Proyectando ahora dichas circunferencias sobre las rectas engendradas por $\vec{1}$ y $\vec{x}$, se obtienen intervalos de confianza centrados en $b_0\vec{1}$ y $b_1\vec{x}$, respectivamente (fig. 7.10).

Por último, dividiendo por $\|\vec{1}\|$ y $\|\vec{x}\|$ (representado simbólicamente por $\vec{b}_0$ y $\vec{b}_1$, respectivamente), se tienen intervalos centrados en $\vec{b}_0$ y $\vec{b}_1$ a partir

de los cuales se pueden formar circunferencias de centro $\vec{b}=(\vec{b}_0+\vec{b}_1)=(b_0, b_1)$ que expresarán la aleatoriedad normal bivariable sobre el vector $\vec{b}$ de estimaciones (fig. 7.11).

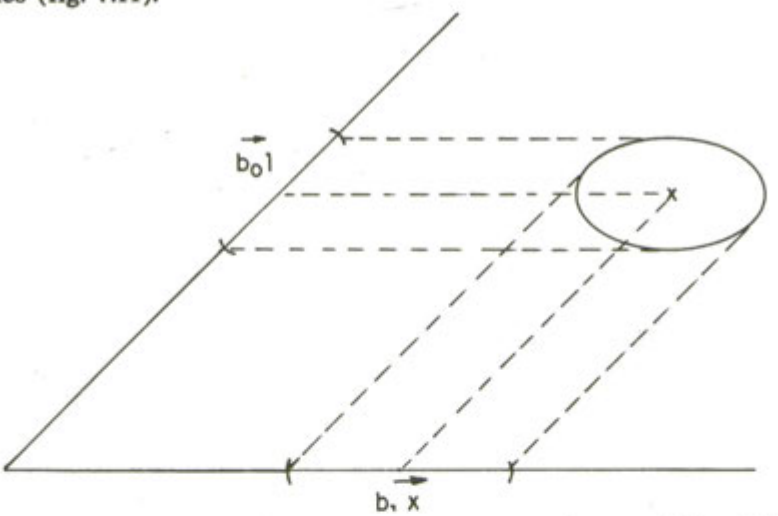


Figura 7.10.—Intervalos de confianza para los parámetros del modelo.

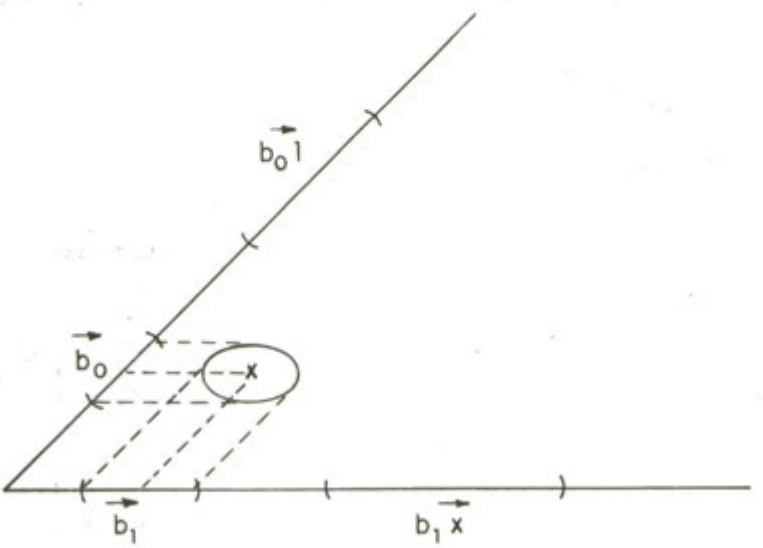


Figura 7.11.—Aleatoriedad normal bivariante sobre el vector de estimaciones.

7.4.4. *Descomposición de la suma de cuadrados en regresión simple*

Una práctica usual en regresión es la descomposición de la suma de cuadrados de la variable dependiente. Esta suma de cuadrados viene representada por $\|\vec{y}\|^2$, es decir, el módulo del vector de observaciones. Desde el punto de vista geométrico consiste en la aplicación del teorema de Pitágoras,

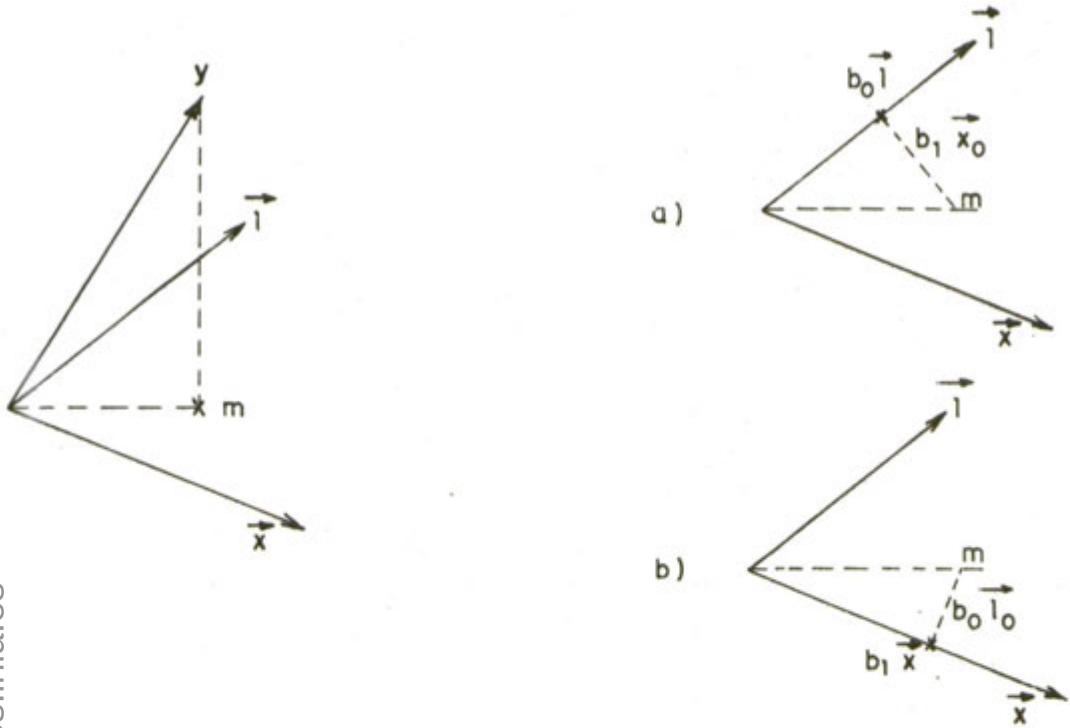


Figura 7.12.—Descomposición de la suma de cuadrados.

por tanto es necesario que los triángulos en cuestión sean rectángulos, para lo cual las proyecciones deben ser ortogonales. En las figuras 7.12 a y b se explica gráficamente las proyecciones que es necesario efectuar para llegar a la descomposición de $||\vec{y}||^2$. Según ellas,

$$||\vec{y}||^2 = ||\vec{e}||^2 + ||\vec{m}||^2 = \begin{cases} (1)\ ||\vec{e}||^2 + ||b_0\vec{1}||^2 + ||b_1\vec{x}_0||^2 \\ \quad \text{SC error + SC media + SC x/media} \\ (2)\ ||\vec{e}||^2 + ||b_1\vec{x}||^2 + ||b_0\vec{1}_0|| \\ \quad \text{SC error + SC x + SC media/x} \end{cases}$$

por tanto, se obtienen dos tablas del análisis de varianza diferentes que sólo coincidirán cuando $\vec{1}$ y $\vec{x}$ sean perpendiculares (ortogonales). La primera descomposición es la más usual y la empleada a lo largo del texto.

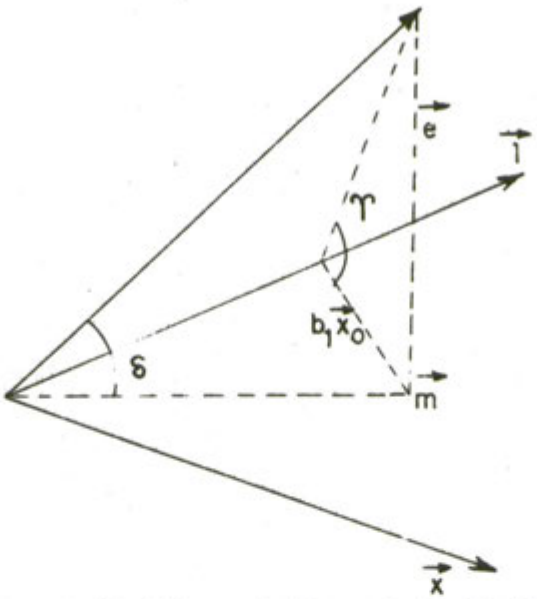


Figura 7.13.—Representación de la prueba F para la significación del modelo.

Los estadísticos para las pruebas F de la descomposición (1), según la figura 7.13, son:

$$\frac{n-2}{2}\ \frac{||\vec{m}||^2}{||\vec{e}||^2} = \frac{n-2}{2}\ \text{cotg}^2\ \delta$$

para la significación global de la regresión, y

$$(n-2)\quad \frac{||b_1\vec{x}_0||^2}{||\vec{e}||^2} = (n\cdot 2)\text{cotg}^2\gamma$$

para la significación de la variable regresora x, estando la media ($\vec{1}$) presente ya en el modelo.

Es decir, la regresión será significativa al α por 100 si δ es lo suficientemente pequeño como para que

$$\delta < \text{arc cotg}\ \sqrt{(F_{\alpha;2,n-2})\ \frac{2}{n-2}}$$

lo que es equivalente a

$$\cos^2\delta > \frac{2F_{\alpha;2,n-2}}{n-2+2F_{\alpha;2,n-2}}$$

que puede interpretarse como una cota para el cuadrado del coeficiente de correlación entre y e $\hat{y}$ (recuérdese lo indicado en la introducción).

Análogamente, la influencia de x adicional a la de la media es significativa al α por 100 si γ es lo suficientemente pequeño como para que:

$$\gamma < \text{arc cotg } \sqrt{(F_{\alpha;1,n-2})\ \frac{1}{n-2}} = \text{arc cotg}\left(t_{\alpha;n-2}\ \frac{1}{\sqrt{n-2}}\right)$$

que puede también expresarse como:

$$\cos^2\gamma > \frac{t^2_{\alpha;n-2}}{n-2+t^2_{\alpha;n-2}}$$

y representa una cota para el cuadrado del coeficiente de correlación entre y y $b_1\vec{x}_0$ cuando $\vec{1}$ está ya en el modelo (es decir, entre las proyecciones de y y $b_1\vec{x}$ sobre el plano perpendicular a $\vec{1}$).

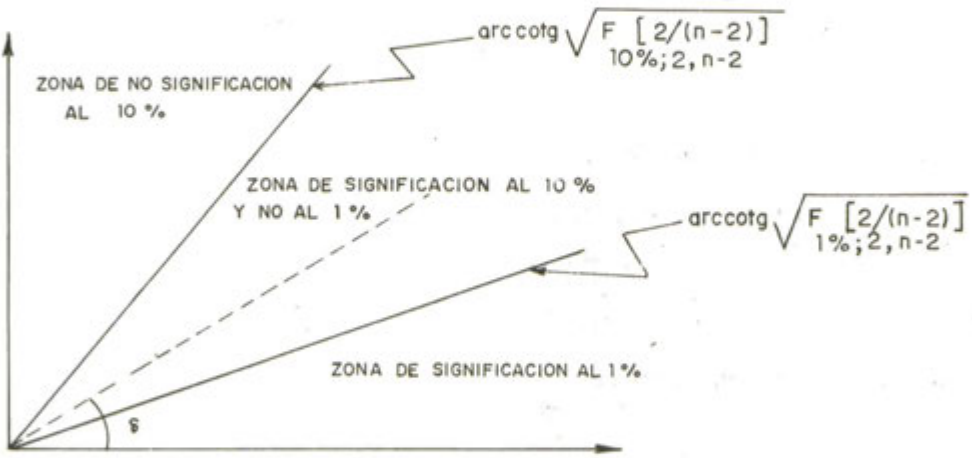


Figura 7.14.—Interpretación de la significación del modelo.

La interpretación de las fórmulas anteriores en sencilla recurriendo a la figura 7.14. Cuanto más pequeña sea δ, más se aproxima $\vec{y}$ al plano Lx y; por tanto, menor longitud tiene, es decir, más explicable es y por $\vec{1}$ y $\vec{x}$.

Por otra parte, como se puede observar en la figura 7.15, si manteniendo la longitud de $\vec{y}$ constante se le hace girar sobre el eje engendrado por 1 para que $\vec{Z}=\vec{y}'-\vec{y}$ sea constante, cuanto mayor sea γ, más cerca se encuentra y de la recta engendrada por $\vec{1}$ y, en consecuencia, menor es la coordenada que queda para x_0.

Por tanto, la situación es similar a la anterior. Cuando γ sea menor que un valor regulado a través de α y las tablas t, se considerará que la influencia de x, cuando $\vec{1}$ (la media) «está» en el modelo, es significativa, una vez declarada significativa la regresión total.

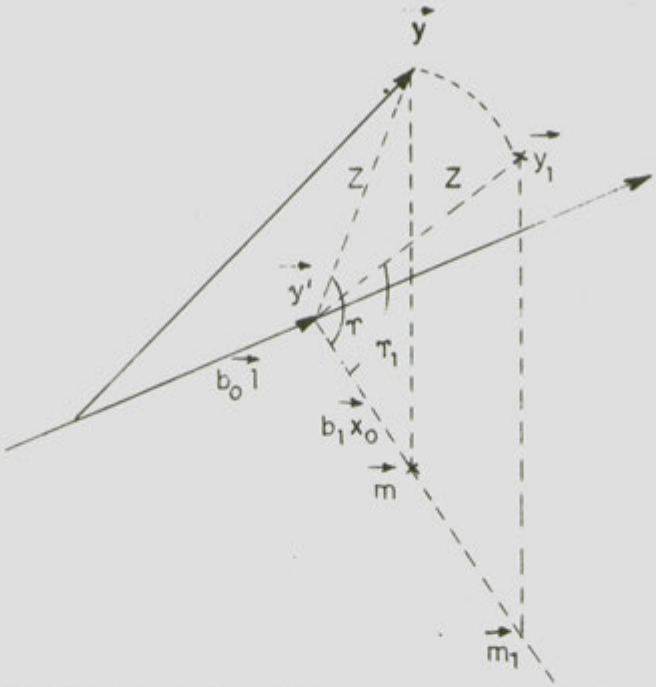


Figura 7.15.—Significación de la variable regresora según el valor del ángulo γ.

7.5. Regresión múltiple: Observaciones

Es interesante resaltar aquí dos cuestiones que, aunque pueden ser analizadas en regresión simple, cobran en la múltiple toda su importancia.

La primera es relativa al orden en que las variables son introducidas en el modelo cuando realizamos un Análisis de Varianza (capítulo 4).

En regresión simple habría dos posibilidades: primero $\vec{1}$ (media) y después $\vec{x}$, o al contrario, y normalmente se elegía la primera por razones bastantes claras.

En una regresión múltiple con p variables habrá p! formas de introducirlas, y, salvo que se trate de un modelo inclusivo, no se tienen indicaciones precisas sobre el orden adecuado.

En este punto la decisión corresponde al investigador, previo análisis de la naturaleza de las variables y utilizando el contraste F «sobre» los coeficientes de correlación parcial entre Y y la variable candidata x_j, teniendo en cuenta las ya presentes en el modelo.

Obsérvese que el criterio de la F para la valoración de la aportación de una nueva variable (acotando el coeficiente de correlación adecuado) no es equivalente a la consideración directa del porcentaje de explicación de la nueva variable dadas las demás (ver 3.2.2), ya que éste es el cociente de la aportación en S. C. de esta variable partido por la S. C. total ($\|y\|^2$) y al contrastar el coeficiente de correlación parcial se utiliza como denominador $\|y'y\|^2$ (eliminada la influencia de las variables ya presente en el modelo) (ver fig. 7.16).

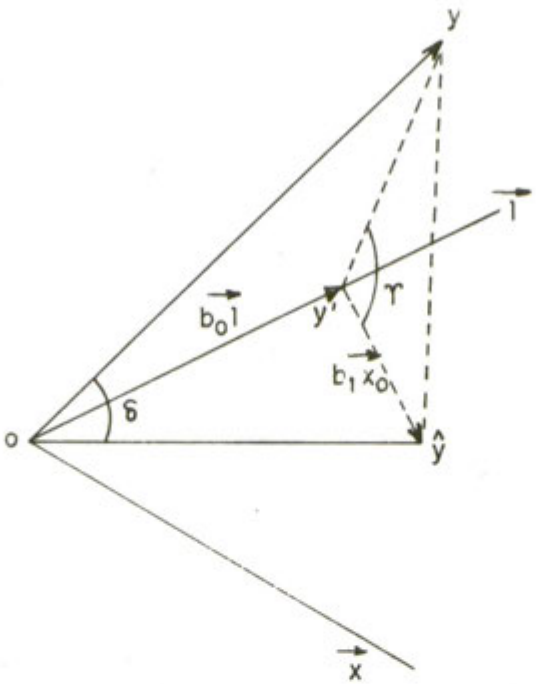


Figura 7.16.—Comparación entre el coeficiente de correlación parcial y el tanto por ciento de explicación.

Sin embargo, el porcentaje de explicación es un interesante valor de referencia para medir la aportación de cada variable una vez que el proceso de selección se considere terminado.

La segunda cuestión consiste en el análisis de las consecuencias de la colinealidad (variables regresoras con relación casi lineal).

Utilizando la figura 7.10 puede fácilmente analizarse las consecuencias en la regresión simple y generalizadas a la múltiple, que es donde tienen más interés.

Al cerrarse el ángulo entre $\vec{1}$ y $\vec{x}$ (mayor colinealidad) la proyección de las circunferencias que representan la varianza de y sobre $\vec{1}$ y $\vec{x}$ dan intervalos de confianza más amplios alrededor de $b_0\vec{1}$ y $b_1\vec{x}$, con lo que las predicciones son más imprecisas.

En caso de ortogonalidad entre $\vec{1}$ y $\vec{x}$ (colinealidad nula) los efectos de ambos vectores son independientes (recordar la introducción) y los estimadores tienen intervalos de confianza mínimos.

Si sustituimos la pareja $(\vec{1}, \vec{x})$ por las $(\vec{x}_i, \vec{x}_j)$ que podemos obtener en regresión múltiple y aplicamos el mismo razonamiento, podremos comprender la influencia negativa que la colinealidad tiene en la precisión de las predicciones.

CAPITULO 8

EJEMPLOS INTERPRETATIVOS

8.1. Introducción

Para completar esta publicación se incluyen a continuación una serie de ejemplos prácticos con intención de ayudar a comprender y resumir todo lo expuesto en los capítulos anteriores.

Los datos utilizados en la mayoría de los ejemplos son reales e intentan presentar un problema concreto.

Lógicamente, dada la naturaleza de realidad de las observaciones, los resultados no saldrán tan perfectos o críticos como en un ejemplo con datos simulados. En esta parte, más que presentar el cálculo en forma metodológica, pues ya fue incluido en las respectivas secciones, lo que se pretende es hacer hincapié en la parte de discusión e interpretación de resultados.

Solamente en un ejemplo se ha verificado la validez de las suposiciones del modelo empleado y se ha efectuado un estudio de los residuos. Naturalmente, en todos los demás ejemplos, y en general en cualquier estudio, debe prestarse gran atención a la validación del modelo, tal como se indicaba al comienzo del capítulo 6. Para evitar ser reiterativo no se ha incluido tal validación en el resto de los ejemplos.

En los ejemplos no se justifica el nivel de resignificación α empleado; sin embargo, en cualquier utilización práctica de la gresión (como en muchas otras técnicas de estadística inferencial) debe tenerse muy en cuenta el nivel de significación con que se efectúan las pruebas estadísticas. Este nivel de significación debe ser decidido por el investigador en base a su conocimiento del problema, experimentos anteriores, etc. Algunas consideraciones pueden ayudarle a decidir qué nivel α debe utilizar. Por un lado, el problema puede sugerir que se acepte la hipótesis nula. En este caso debería emplear un α pequeño; sin embargo, si pruebas experimentales y conocimientos previos parecen indicar que la hipótesis nula debe ser rechazada, el valor α debe ser grande (el 10 por 100, por ejemplo). Naturalmente ,estas decisiones deben tomarse antes de obtener ningún dato del experimento. Por otro lado, puede basar su decisión en el error que tenga más interés en controlar; si el peligroso es el error tipo I (declarar significación cuando no debía), el α debe ser pequeño; en caso contrario, si quiere controlar el error tipo II (no declarar significación cuando debía), el α debe ser grande. Con esto queda subrayada la importancia de la decisión del nivel de significación, que no debe hacerse de forma rutinaria, sino teniendo en cuenta las consideraciones anteriores y estableciendo algún tipo de compromiso cuando las circunstancias del problema no indiquen claramente qué es lo que se debe de hacer.

8.2. Ejemplo 1: Estimación en regresión simple

En este ejemplo se intenta mostrar el camino a seguir ante un problema general de ajuste por regresión lineal simple: suposiciones de base, estima-

ción, pruebas de hipótesis, estudio de residuos para validar el modelo, intervalos de confianza y predicción. Los datos se refieren al crecimiento de terneros alimentados con un pienso determinado. La variable regresora corresponde a los días de evaluación del crecimiento.

TABLA 8.1

Datos de crecimiento de cinco terneros en función de la edad

Edad	*Crecimiento de los terneros*				
	1	2	3	4	5
33	350	351	343	352	338
62	392	395	381	383	368
92	413	423	417	415	418
111	445	442	435	450	431
151	460	472	477	454	451
172	474	490	481	470	468
201	489	486	506	482	487
220	503	503	516	507	508

Se han tomado datos de cinco terneros de una misma raza para estudiar el crecimiento medio racial intentando amortiguar la dependencia existente entre las observaciones a distintos tiempos, tal como se indicó en 1.3.3. Los datos se presentan en la tabla 8.1 y en la figura 8.1.

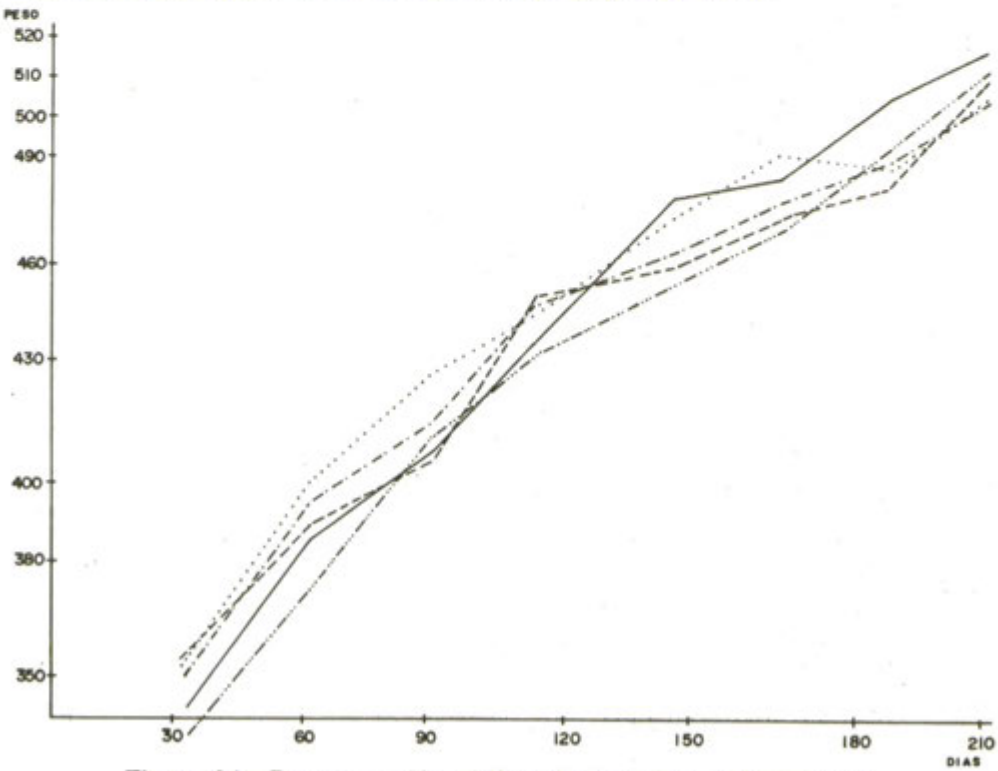


Figura 8.1.—Representación gráfica de los datos de ejemplo 1.

Dado que, aunque las medias presentan una ordenación temporal, el total de las observaciones no tiene una razón de ordenación especificada (la segunda observación del primer tiempo no tiene claramente definida una ordenación con respecto a la tercera de ese mismo tiempo, por ejemplo), no se efectuará una prueba de independencia, limitándose a un estudio posterior de los residuos. Si se utilizaran las medias en vez de los datos originales sería posible efectuar la prueba; sin embargo, por el escaso número de puntos carecería de interés, puesto que la potencia de la prueba sería demasiado pequeña.

Por tanto, se investigará en primer lugar la *homoescedasticidad*, tal como quedó descrita en 6.2.2. Aplicando, pues, la metodología de la prueba de Burr y Foster a los datos de la tabla 8.1, se tiene:

$$q_c = \frac{\sum_i s_i^4}{(\sum_i s_i^2)^2} = \frac{34453,2}{213573,4} = 0,16 \text{ con } p=8,\ \nu=4$$

Buscando en las tablas del anejo 2 el valor del estadístico q para el nivel de significación a=0,01, se obtiene q=0,276. Dado que $q_c < q$, no se encuentra evidencia contra la homoescedasticidad.

Para verificar la *normalidad*, según la prueba de Shapiro y Wilk, es necesario previamente estimar la recta de regresión.

Con las fórmulas pertinentes de 2.1.1 se halla:

$$\hat{y} = 334,78 + 0,81x$$

TABLA 8.2

Valores residuales para el ajuste lineal de los datos de la tabla 8.1

	Ternero				
Edad	1	2	3	4	5
33	−11,60	−10,60	−18,60	−9,60	−23,60
62	6,83	9,83	−4,17	−2,17	−17,17
92	3,44	13,44	7,44	5,44	8,44
111	20,00	17,00	10,00	25,00	6,00
151	2,48	14,48	19,48	−3,52	−6,52
172	−0,59	15,41	6,41	−4,59	−6,59
201	−9,16	−12,16	7,84	−16,16	−11,16
220	−10,60	−10,60	2,40	−6,60	−5,60

A partir de esta ecuación se pueden calcular los valores residuales e_i que se incluyen en la tabla 8.2. Con los valores residuales se siguen los pasos descritos en 6.2.3 obteniendo

$$\sum e_i^2 = 5478,92; \quad b^2 = 5330,27; \quad W_c = \frac{b^2}{\sum e_i^2} = 0,97$$

El valor del estadístico W de las tablas del anejo 4 para $\alpha=0,01$ y $d=38$ es 0,916. Como $W_c > W$ no se puede rechazar la normalidad.

Una vez comprobadas las hipótesis de base que se pueden verificar en este estudio es necesario preguntarse si hay influencia de la variable regresora.

La tabla del análisis de la varianza de la regresión es:

Fuentes de variación	*S.C.*	*g.l.*	*C.M.*	*Fc*
Debido a regresión	103020,17	1	103020,17	714,51
Desviaciones	5478,93	38	144,82	
TOTAL	108499,10	39		

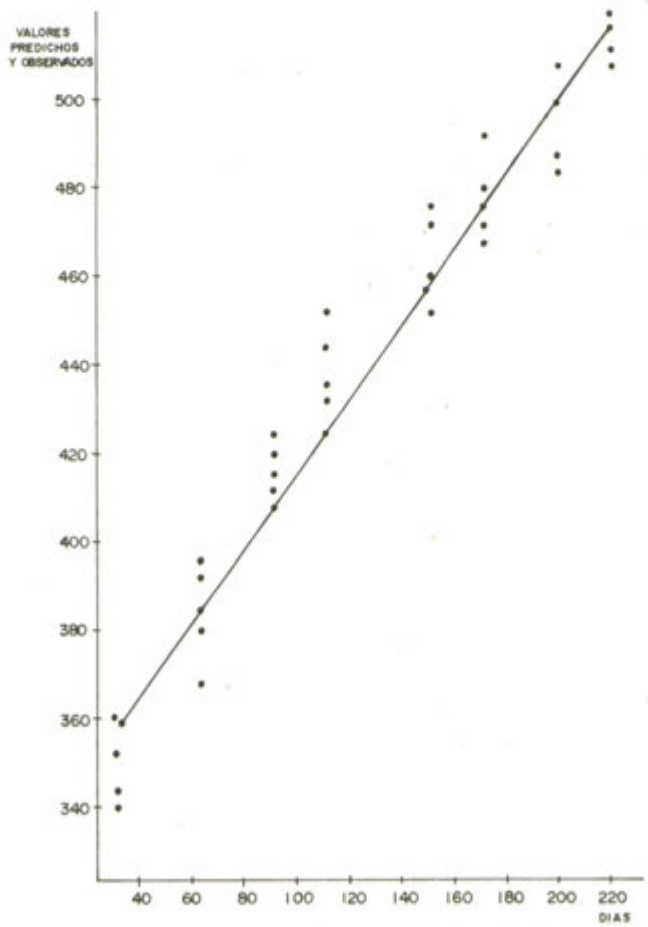


Figura 8.2.—Ajuste a la regresión lineal simple.

Comparando Fc con el valor 7,35 correspondiente a una tabla F con 1 y 38 grados de libertad y un nivel de significación del 0,01, se declara altamente significativo el modelo de regresión lineal simple. El valor del coeficiente de determinación es:

$$R^2 = \frac{\text{SC Regresión}}{\text{SC Total}} = \frac{103020,17}{108499,10} = 0,95$$

lo que parece indicar que el modelo se halla bien explicado por la linearidad, concluyendo, quizá erróneamente, que el modelo está correctamente especificado por obtenerse un R^2 alto.

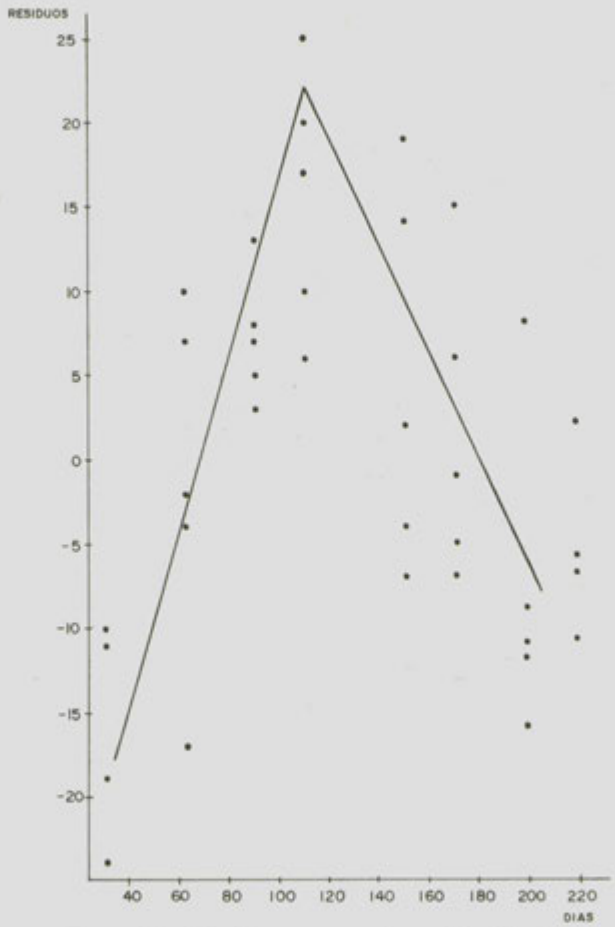


Figura 8.3.—Residuos para el modelo lineal.

Para estudiar posibles problemas no aparentes se investiga la tabla de residuos representándolos gráficamente frente al valor de la variable regresora. En las figuras 8.2 y 8.3 se observa una disposición parabólica de los residuos, por tanto no aleatoria, con una alternancia en los signos del tipo

negativo, positivo, negativo, lo que hace inmediatamente sospechar que el modelo es incorrecto y debería incluir un término cuadrático u otros de orden superior (situación ésta que debería haberse intuido al representar gráficamente los datos según la figura 8.1, o incluso conocerse por la naturaleza intrínseca del problema).

Dado que se dispone de medidas repetidas, otra herramienta que se puede emplear en este caso es desdoblar la variación residual en sus componentes de error puro y falta de ajuste tal como se indicó en 6.3.

Calculando $\sum_u (y_{iu} - \bar{y}_i)^2$ para cada grupo de datos, y sumando los resultados, se obtiene que

$$\text{S.C. (error puro)} = \sum_i \sum_u (y_{iu} - \bar{y}_i)^2 = 2180,8$$

Reconstruyendo la tabla:

Fuentes de variación	*S.C.*	*g.l.*	*C.M.*	*Fc*
Debido a regresión	103020,17	1	103020,17	
Desviaciones	5478,93	38		
Falta de ajuste	3298,13	6	549,68	8,07
Error puro	2180,80	32	68,15	

Comparando F_c al 1 por 100 con el valor 3,42 de las tablas F con 6 y 32 grados de libertad, se declara altamente significativa la falta de ajuste, corroborando lo que indicaba el estudio de los residuos.

Con estos resultados se puede comprobar que la primera conclusión alcanzada era falsa. Por tanto, debe de tenerse muy en cuenta que una cosa es la significación del modelo y otra su bondad o validez.

TABLA 8.3

Valores residuales para el ajuste parbólico de los datos de la tabla 8.1

	Ternero				
Edad	1	2	3	4	5
33	1,40	2,40	−5,60	3,40	−10,60
62	7,97	10,97	−3,03	−1,03	−16,03
92	−3,19	6,81	0,81	−1,19	1,81
111	10,81	7,81	0,81	15,81	−3,18
151	−6,09	5,91	10,91	−12,09	−15,09
172	−5,58	10,42	1,42	−9,58	−11,58
201	−5,53	−8,53	11,47	−12,53	−7,53
220	1,00	1,00	14,00	5,00	6,00

Como consecuencia de todo lo anterior es recomendable probar un modelo polinomial de segundo grado introduciendo en el modelo una nueva variable $x_3 = x^2$. Volviendo a estimar los valores de los parámetros y a veri-

ficar la hipótesis de normalidad (no incluida aquí), la ecuación resulta ser:

$$Y = 303,08 + 1,46x - 0,00254x^2$$

En la tabla 8.3 se encuentran los valores residuales de acuerdo con esta ecuación. En la figura 8.3 se aprecia que el ajuste de los puntos al modelo parabólico es mejor al presentado en la figura 8.2. En la figura 8.5 aparece la gráfica de los residuos, que denota un comportamiento más «normal», es decir, más aleatorio. De todas formas, se comprobará esta indicación mediante la tabla del análisis de la varianza:

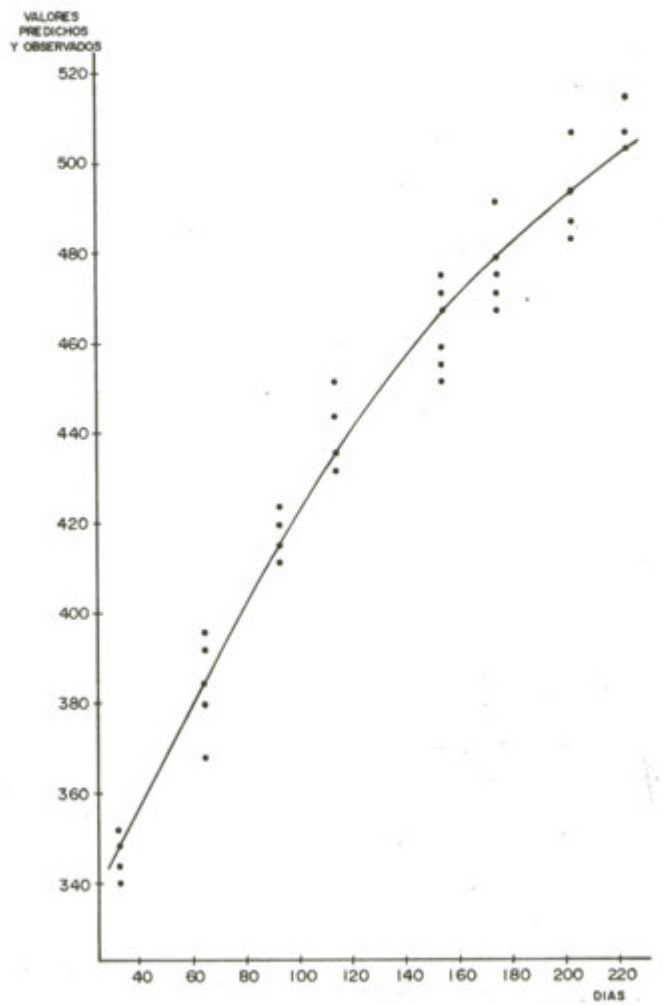


Figura 8.4.—Ajuste a la regresión parabólica.

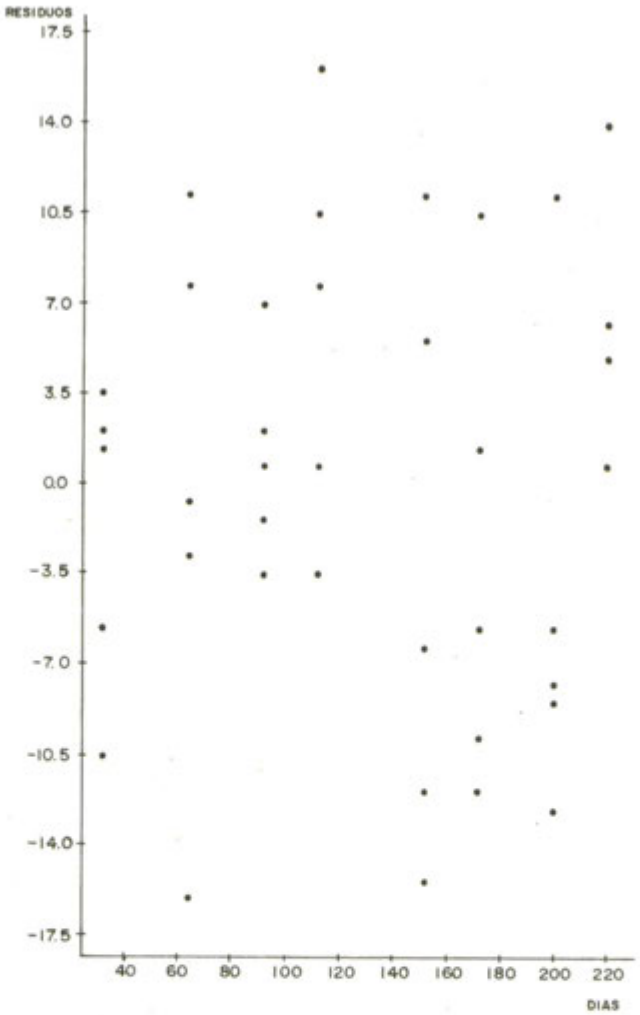


Figura 8.5.—Residuos para el modelo parabólico.

Fuentes de variación	*S.C.*	*g.l.*	*C.M.*	*Fc*
Debido a regresión	105744,60	2	52872,33	710,23
Desviaciones	2754,44	37	74,44	
Falta de ajuste	573,64	5	114,73	1,68
Error puro	2180,80	32	68,15	

El valor $F_c = 1,68$ no resulta significativo al compararlo con $F\ (6;32;0,05) = 2,51$, por lo que se puede pensar que el modelo ajustado es admisible para describir el comportamiento del crecimiento, naturalmente siempre dentro de la óptica de un modelo lineal.

Puesto que la falta de ajuste no es significativa, la mejor estima del error experimental viene dada por 74,44. Este valor actuará como denominador en el cálculo de F_c para verificar la significación de la regresión. Es de observar que el valor $F_c=710{,}23$ es menor que el que se obtuvo anteriormente para el modelo incorrecto ($F_c=714{,}51$), lo que a primera vista puede parecer ilógico; ahora bien, esta comparación carece de fundamento, pues F_c está basado en diferente número de grados de libertad en ambos casos. Realmente, la base do comparación debería ser la significación de los F_c. Para el modelo parabólico, $p=1{,}41\times10^{-10}$, y para el incorrecto, $p=3{,}70\times10^{-10}$; es decir, el modelo parabólico tiene una mayor significación, puesto que la probabilidad de rechazar la hipótesis nula cuando es correcta (dada por el valor p), es menor. Igualmente, el R^2 para el modelo parabólico sufre una ligera mejoría pasando de 0,95 a 0,97.

Dado por bueno ya el modelo, se debe obtener un intervalo de confianza para los coeficientes de la regresión, estudiándose así la precisión de las estimas.

Puesto que se dispone de un modelo parabólico en el que intervienen dos variables regresoras, $x_1=x$ y $x_2=x^2$, los cálculos se realizarían muy rápidamente por medio de matrices. Como el aparato matemático para matrices de orden tres es sencillo, se presentarán las fórmulas detalladas.

Por 2.1.3. se tiene que

$$\text{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \sigma^2(X'X)^{-1}$$

Como para este caso

$$(X'X) = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}$$

el determinante de la matriz será:

$$\text{Det}(X'X) = \sum x_i^2\left(n\sum x_i^4 + 2\sum x_i\sum x_i^3\right) - \left[\left(\sum x_i^2\right)^3 + n\left(\sum x_i^3\right)^2 + \left(\sum x_i\right)^2\sum x_i^4\right]$$

Entonces,

$$\sigma^2_{\hat{\beta}_0} = \sigma^2\frac{A_{11}}{\text{Det}(X'X)}\ ;\quad \sigma^2_{\hat{\beta}_1} = \sigma^2\frac{A_{22}}{\text{Det}(X'X)}\ ;\quad \sigma^2_{\hat{\beta}_2} = \sigma^2\frac{A_{33}}{\text{Det}(X'X)}$$

siendo A_{11}, A_{22}, A_{33} los adjuntos de los elementos [1,1], [2,2] y [3,3] de la matriz X'X, respectivamente. Sus expresiones para el cálculo son:

$$A_{11} = \sum x_i^2 \sum x_i^4 - (\sum x_i^3)^2$$

$$A_{22} = n\sum x_i^4 - (\sum x_i^2)^2$$

$$A_{33} = n\sum x_i^2 - (\sum x_i)^2$$

Sustituyendo los valores pertinentes y σ^2 por su estima (cuadrado medio del residuo), se tiene una estima de la varianza de los coeficientes de regresión. Las correspondientes desviaciones típicas son:

$$\hat{\sigma}_{\hat{\beta}_0} = 6,12 \qquad \hat{\sigma}_{\hat{\beta}_1} = 0,11 \qquad \hat{\sigma}_{\hat{\beta}_2} = 0,00042$$

El intervalo de confianza del 95 por 100 se obtendrá con la fórmula

$$\hat{\beta}_j \pm t_{37;0,05}\ \hat{\sigma}_{\hat{\beta}_j} \text{ (para } j=0,1,2)$$

que, particularizando para cada coeficiente de regresión y empleando el valor de $t_{37;05}=2,03$, proporciona los siguientes valores:

$$\beta_0 : (290,67;\ 315,49)$$

$$\beta_1 : (1,240;\ 1,686)$$

$$\beta_2 : (0,00216;\ 0,00384)$$

Dado que se ha supuesto el modelo parabólico como correcto, podría pensarse que las estimas de los coeficientes de regresión son más precisas (menor varianza) que las que se obtendrían a partir del modelo de regresión lineal, puesto que en el cálculo de las desviaciones típicas interviene el cuadrado medio del residuo y éste se encontraba hinchado por la falta de ajuste del modelo. Hallando las correspondientes desviaciones típicas, se obtiene

$$\hat{\sigma}_{\hat{\beta}_0} = 4,38 \quad ; \quad \hat{\sigma}_{\hat{\beta}_1} = 0,03$$

que, curiosamente, son menores. La explicación a este comportamiento se encuentra en la presencia de una alta colinearidad entre las dos variables regresoras x y x^2, que produce, según se explicó en 3.2.1, una inestabilidad en la estimación de los coeficientes de regresión.

Muchas veces el investigador está interesado en predecir el valor que tomará la esperanza de la variable dependiente para uno predeterminado de las regresoras (χ_w). La precisión con que se efectúa la predicción está en función de su varianza, quien, a su vez, depende del valor χ_w según la fórmula

$$\text{var}(\hat{Y}_w) = x_w'(X'X)^{-1}x_w'\sigma^2$$

Se tomarán tres valores diferentes de χ_w para mostrar que la precisión disminuye (aumentando el intervalo de confianza) a medida que χ_w se separa de su media.

Por ejemplo, para el valor $x=40$ ($x^2=1600$), χ'_w : (1 40 160).

El valor predicho para la variable dependiente se obtendrá sustituyendo en la ecuación de predicción x por 40 y x^2 por 1600; es decir, $y=303{,}08+1{,}46x-0{,}00254x^2=356{,}80$.

El resultado de $(X'X)^{-1}$, del que previamente se había obtenido su diagonal principal, es:

$$(X'X)^{-1} = \begin{pmatrix} 0{,}5026 & -0{,}0084 & 0{,}29 \times 10^{-6} \\ -0{,}0084 & 0{,}00016 & -0{,}61 \times 10^{-6} \\ 0{,}29 \times 10^{-6} & -0{,}61 \times 10^{-6} & 0{,}24 \times 10^{-8} \end{pmatrix}$$

como $\hat{\sigma}^2=74{,}44$, sustituyendo

$$\hat{\sigma}^2_{\hat{y}_w} = 8{,}34 \text{; por tanto, } \hat{\sigma}_{\hat{y}_w} = 2{,}89$$

El intervalo de confianza del 95 por 100 se hallaría por la siguiente fórmula:

$$\hat{y}_w \pm t_{37;0{,}05}\, \hat{\sigma}_{\hat{y}_w} \text{ que, sustituyendo, se obtendría}$$

$$\hat{y}_w : (356{,}79 \pm 5{,}83) = (350{,}96;\ 362{,}64)$$

Repitiendo los cálculos para, por ejemplo, $x=100$, la desviación típica de la estima sería ahora:

$$\hat{\sigma}_{\hat{y}_w} = 1{,}99$$

lo que proporcionaría el intervalo de confianza siguiente:

$$\hat{y}_w : (419{,}38 \pm 4{,}04) = (415{,}34;\ 423{,}42)$$

Si se efectuara la predicción para $x=210$, los valores correspondientes serían:

$$\hat{\sigma}_{\hat{y}_w} = 2{,}45 \qquad \hat{y}_w : (478{,}38 \pm 4{,}97) = (473{,}41;\ 483{,}35)$$

Como el valor medio de la x para el conjunto de observaciones del experimento es 132,8, queda claramente demostrado que, para x=100, que corresponde al punto estudiado más próximo a la media x=132,8, el intervalo es menor, deteriorándose la predicción cuando se efectúa lejos de los valores centrales, tanto en un sentido como en el otro. Por otro lado se podría tener intención de estimar el valor de la Y para un punto fuera del intervalo utilizado en la estimación del modelo. En la figura 8.6 se muestra gráficamente el error cometido al pretender extrapolar los resultados al punto x=390. Si se empleara el modelo que, aparentemente, era el correcto (con un R^2 de 0,95) el error cometido sería de 120,68, Aún empleando el modelo parabólico, este valor es de 43,85.

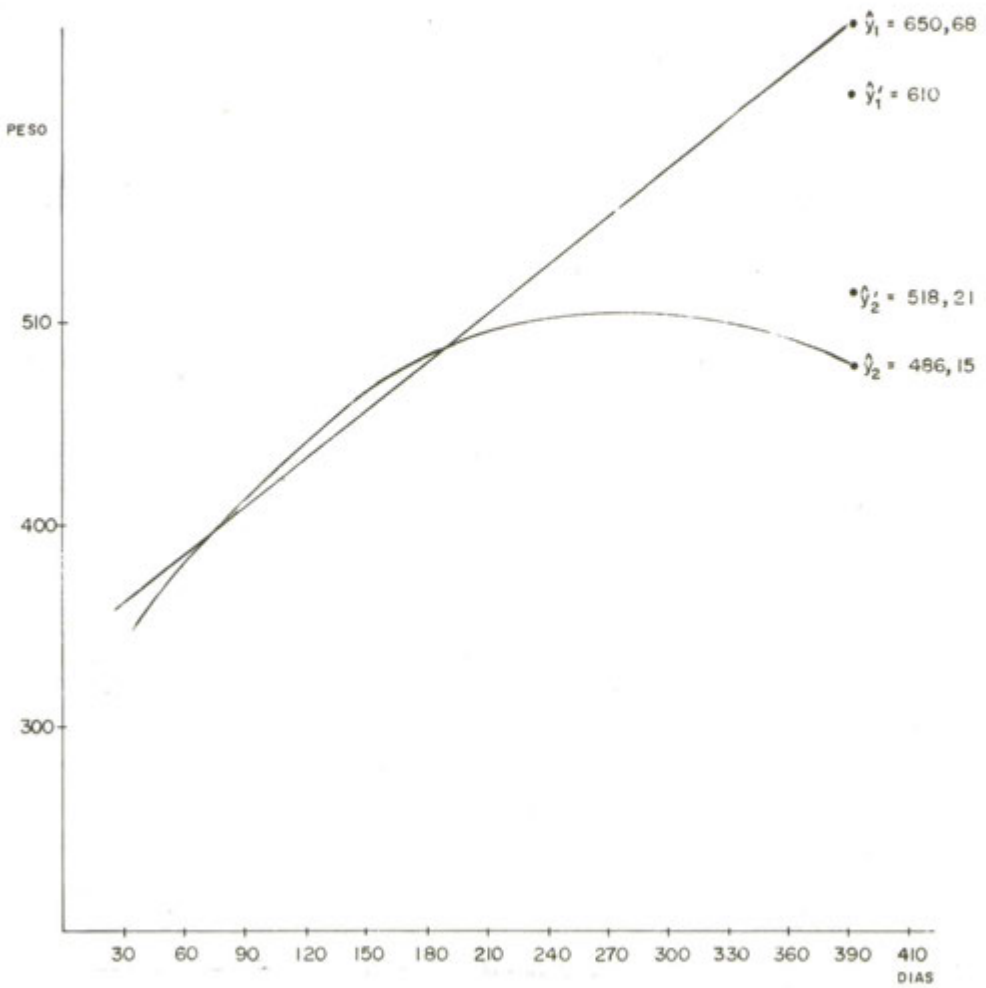


Figura 8.6.—Peligros de la extrapolación. (Subíndice 1, modelo lineal; subíndice 2, modelo parabólico. La «prima» indica la ecuación ajustada considerando el valor x=400.)

Si, por ejemplo, se hubiera incluido la pareja de valores x=400, y=530, como datos experimentales, la ecuación para el modelo parabólico sería

$$y = 309,95 + 1,314x - 0,002x^2$$

de tal manera que el error hubiera sido de 11,79. Con el modelo lineal el error sería de 80,40 puesto que la ecuación estimada es:

$$\hat{y} = 352,22 + 0,662x$$

Con todo ello, se pretende poner de manifiesto que no es suficiente con obtener una serie de valores estadísticos más o menos importantes (como el coeficiente de determinación o la significatividad del modelo), para concluir que el modelo de regresión elegido es el correcto, sino que es necesario efectuar un estudio más profundo del problema. Una vez realizado, se podría estar en condiciones de afirmar que el modelo empleado describe adecuadamente los datos obtenidos, pero solamente dentro del rango de variación de las variables regresoras que fue empleado en el experimento. No se debe extrapolar las conclusiones del experimento a puntos fuera de ese rango, a no ser que se esté seguro de que el modelo mantiene su validez en ellos (contando por otro lado con el deterioro de la predicción ya repetidamente indicado). De la misma forma, no se pueden sacar conclusiones en el sentido contrario, es decir, afirmar taxativamente que una regresora no influye en la dependiente porque se obtuvo una mala significación. Puede ocurrir que el modelo empleado al describir esa influencia no sea el adecuado.

8.3. Ejemplo 2: Colinearidad y selección de variables

El objetivo de este ejemplo es indicar cómo se detecta la colinealidad y cómo actuar consecuentemente antes de realizar una selección de variables. Se incluye también un estudio de estimación de parámetros por medio del método «ridge» (en el caso de que el investigador no esté interesado en seleccionar variables manteniéndolas todas en el modelo e intentando explicar el efecto individual de cada una de ellas).

Dada una muestra de peces del género Rutilus (tabla 8.4) se pretende establecer una ecuación de predicción de la longitud total del pez (LT) en función del menor número de variables relacionadas con las longitudes de la parte delantera del pez. Este objetivo se basa en que algunas enfermedades atacan a las partes distales del pez y a las aletas. Las variables regresoras consideradas son: Longitud predorsal (LPRD), longitud prepectoral (LPRP), longitud de la cabeza (LC), longitud postorbitaria (LPSO), longitud del ojo (LO) y longitud preorbitaria (LPRO).

En primer lugar se calcula el ajuste a un modelo lineal (en parámetros y variables) del tipo:

$$LT = \beta_0 + \beta_1 LPRD + \beta_2 LPRP + \beta_3 LC + \beta_4 LPSO + \beta_5 LO + \beta_6 LPRO + \varepsilon$$

El valor de los coeficientes de regresión, sus varianzas y la significación de los coeficientes se incluyen en la tabla 8.5. Según esta tabla, la significación de cada una de las variables cuando el resto están presentes en la ecuación, indica que, en principio, la LPRD es la más informativa y las LO y LPRO las menos. De todas maneras, antes de eliminar sin más estas dos últimas variables, es conveniente estudiar la presencia de colinearidad pues podría falsear los resultados de la selección. En la tabla 8.6 se presenta la matriz de correlaciones entre todas las variables. En ella se aprecia que las

TABLA 8.4

Datos del ejemplo 2

LT	*LPRD*	*LPRP*	*LC*	*LPSO*	*LO*	*LPRO*
972	433	180	190	92	45	53
1.080	466	199	227	118	56	53
908	405	169	180	73	48	60
871	384	166	184	92	46	42
1.079	481	212	227	108	56	63
933	418	180	190	87	45	58
923	421	172	191	95	42	54
921	393	167	181	88	42	51
965	414	172	188	86	43	59
920	412	159	195	83	45	67
1.019	460	194	210	107	50	53
973	439	177	200	105	50	45
925	411	151	183	85	45	53
1.051	455	198	204	100	50	54
950	407	176	195	97	46	42
968	410	190	197	90	49	58
956	433	180	186	83	48	55
941	415	170	190	90	46	54
950	438	194	206	112	48	46
1.006	470	183	212	95	53	64
888	416	187	177	82	45	50
835	384	164	179	86	43	50
888	382	171	179	83	43	53
886	399	181	184	85	47	52
932	406	178	188	88	46	54
963	442	186	198	90	47	61
939	416	170	191	93	46	62
880	380	172	188	90	46	52
869	395	167	175	88	43	44
923	395	172	174	82	46	46

TABLA 8.5

Coeficiente de regresión estimados, varianzas y significación en el modelo lineal

Variable	*Coeficiente*	*Varianza*	*Probabilidad*
Constante	86,92		
LPRD	1,04	0,398	0,015
LPRP	0,47	0,618	0,451
LC	2,03	2,082	0,339
LPSO	−0,66	2,114	0,757
LO	0,43	3,330	0,898
LPRO	−0,29	1,827	0,876

correlaciones simples tienen valores bastante altos, claro indicio de que podría existir colinearidad. De todos modos, para profundizar en esta suposición es necesario obtener los autovalores de la matriz de correlación de las regresoras. En la tabla 8.7 aparecen los autovalores ordenados y el tanto por ciento de varianza absorbida por cada uno, así como el tanto por ciento

acumulado. Se observa que el primer autovalor es responsable de una gran cantidad de varianza, llegándose al 87,29 por 100 del total con solamente dos autovalores. El cociente entre el menor autovalor y el número de ellos (que equivale al tanto por uno de la varianza absorbida, según se indicó en 3.3.1) es 0,002825 lo que indica la presencia de una alta colinearidad (con este mismo objeto, otros autores estudian el cociente entre el máximo autovalor y el mínimo).

TABLA 8.6

Matriz de correlaciones simples entre todas las variables

	LT	*LPRD*	*LPRD*	*LC*	*LPSO*	*LO*	*LPRO*
LT	1,000						
LPRD	0,893	1,000					
LPRP	0,748	0,739	1,000				
LC	0,878	0,869	0,746	1,000			
LPSO	0,700	0,683	0,659	0,845	1,000		
LO	0,799	0,791	0,744	0,849	0,645	1,000	
LPRO	0,343	0,344	0,138	0,328	−0,150	0,248	1,000

TABLA 8.7

Autovalores y porcentajes de varianza absorbida

Autovalor	*Porcentaje*	*Porcentaje acumulado*
4,099	68,32	68,32
1,138	18,97	87,29
0,351	5,85	93,14
0,239	3,98	97,12
0,156	2,60	99,72
0,017	0,28	100,00

TABLA 8.8

Autovectores de la matriz de correlaciones

Variables	*Autovectores*					
	1	2	3	4	5	6
LPRD	0,456	−0,144	0,095	0,135	0,862	−0,003
LPRP	0,424	0,099	−0,746	0,463	−0,197	0,000
LC	0,478	−0,025	0,335	−0,016	−0,294	−0,756
LPSO	0,407	0,430	0,493	0,231	−0,233	0,547
LO	0,445	−0,030	0,242	−0,827	−0,083	0,225
LPRO	0,142	−0,885	0,139	0,170	−0,264	0,280

Los vectores propios asociados se encuentran en la tabla 8.8. El estudio de estos vectores indica que en el primero tiene muy poca importancia la variable LPRO ya que el coeficiente asociado a esa variable es el menor, siendo LPRD y LC los más relacionados con él. Dado que este eje absorbe una gran cantidad de inercia (68 por 100) siendo el resto bastante menos importantes y además la correlación de LPRO y LT es muy pequeña, se podría eliminar del análisis la variable LPRO incluso antes de hacer un proceso de selección de variables. Más aún, el autovector asociado al último (y casi nulo) autovalor proporciona la combinación lineal entre las variables regresoras (ver 3.1.3.1). En el ejemplo presente se aprecia que, naturalmente, la longitud de la cabeza es una función de las longitudes postorbital, del ojo y de la preorbital, no teniendo demasiado que ver en esta combinación lineal la longitud predorsal y la prepectoral. Por lo tanto, una de aquellas longitudes debería ser eliminada por redundante.

Detectada y explicado el origen de la colinearidad, es de esperar que la varianza de los coeficientes de regresión sea anormalmente elevada. Una regla práctica para facilitar el cálculo numérico y la interpretación del efecto individual de las variables, es utilizar las variables originales tipificadas, que se denotan con el subíndice t (MARQUARDT y SNEE, 1975). Además resulta interesante tipificar por si posteriormente se pretende hacer una estimación de los parámetros mediante la regresión «ridge» ya que en este caso es obligatorio trabajar con variables tipificadas. Los coeficientes de regresión y sus errores típicos para este caso, se encuentran en la tabla 8.9.

TABLA 8.9

Coeficientes de regresión y errores típicos al emplear las variables tipificadas

Variable	*Coeficiente*	*Errores típicos*
$LPRD_t$	0,47	0,181
$LPRP_t$	0,10	0,136
LC_t	0,47	0,481
$LPSO_t$	−0,11	0,357
LO_t	0,03	0,202
$LPRO_t$	−0,03	0,199

La colinearidad entre las variables resulta lógica por tratarse de medidas longitudinales muy relacionadas unas con otras, luego no es extraña su presencia. Debido a que es posible la existencia de errores aleatorios en la medición de las longitudes, conviene estudiar si la variable dependiente está correlacionada con los ejes principales importantes, pues es posible que la colinearidad pueda ser incluso beneficiosa. Recuérdese que las componentes principales son combinaciones lineales de las variables originales de varianza máxima (ver 3.1.3.2). Los coeficientes de las variables originales obtenidos a partir de los de regresión de los componentes principales (es decir, deshaciendo el cambio de variables), la correlación de éstos con la variable dependiente LT y sus coeficientes de regresión se reflejan en la tabla 8.10. Igualmente, se muestra en la tabla los valores de la suma de cuadrados residuales (SCE), estadístico F para la entrada del componente (Fe), el R^2 y la F de significación del modelo. El orden de entrada de los componentes ha sido fijado por la magnitud de su correlación con la variable dependiente. Al entrar sucesivamente en la ecuación cada uno de los seis ejes principales,

Tabla 8.10

Coeficientes de las variables originales obtenidos de la regresión sobre componentes principales, valor absoluto de la correlación de éstos con la variable dependiente LT_t y sus coeficientes de regresión

Variables	*Componentes introducidos en el modelo*					
	1	5	2	6	3	4
$LPRD_t$	0,204	0,448	0,461	0,462	0,488	0,475
$LPRP_t$	0,189	0,133	0,124	0,124	0,080	0,104
LC_t	0,214	0,131	0,133	0,451	0,471	0,470
$LPSO_t$	0,182	0,116	0,077	−0,153	−0,124	−0,112
LO_t	0,199	0,175	0,178	0,083	0,069	0,026
$LPRO_t$	0,063	−0,011	0,069	−0,046	−0,040	−0,031
LT_t	0,905	0,112	0,097	0,055	0,035	0,026
β_i	0,447	0,283	−0,091	−0,421	0,059	0,052
SCE	5,22	4,86	4,58	4,50	4,46	4,45
Fe	127,4	2,0	1,5	0,5	0,2	0,1
R^2	0,82	0,83	0,84	0,84	0,84	0,84
F	127,4	67,0	46,1	34,0	26,4	21,2

los coeficientes de las variables originales, van siendo distintos. Cuando la ecuación incluye los seis, aquellos son iguales a los obtenidos con la regresión original (tabla 8.9). Sin embargo, dado que el segundo eje ya no es significativo, no hay pérdida importante de información al dar como buena la ecuación basada en un solo eje principal, es decir:

$$LT_t = 0,204LPRD_t + 0,189LPRP_t + 0,214LC_t + 0,182LPSO_t + 0,199LO_t + 0,63LPRO_t$$

Además, como la correlación de la variable original con el primer eje principal es muy alta, la colinearidad puede considerarse como beneficiosa, si se teme que las variables regresoras están medidas con error. Por tanto, se procederá, siguiendo los procedimientos tradicionales de regresión múltiple.

Si el interés del investigador fuera, sin embargo, eliminar variables, se procedería subsiguientemente a utilizar cualquier método de selección, por ejemplo el de paso a paso. Como previamente ya se habría eliminado la LPRO que, además de tener una significación baja según se demostraba en la tabla 8.4, estaba poco relacionada con los ejes principales (tabla 8.7), no se tendrá en cuenta en el proceso. Utilizando el programa BMDP2R descrito en el anejo 5, con límites de 4 y 3,9 para la F_0 de seleccionar y eliminar variables, respectivamente, se obtuvieron los resultados mostrados en las tablas 8.11 a 8.13. En ellas se incluye, además del coeficiente de correlación parcial de cada variable regresora con la dependiente, la tolerancia, el valor F* para seleccionar la variable, la variable seleccionada (señalada con un asterisco), el F_c, el R^2 del modelo, la suma de cuadrados del residuo (SCE) con las variables seleccionadas y las F* para eliminar las variables ya incluidas en el modelo. En el primer paso se introduce la $LPRD_t$ pues ya presentaba una mayor significación en la tabla 8.5 y, por otro lado, su inclusión es interesante por estar muy relacionada con el primer eje principal según la

TABLA 8.11

Resultados en el paso 0 de la selección paso a paso de variables

Variable	*Coef. corr. parcial*	*Tolerancia*	*F* (para seleccionar)*
*$LPRD_t$	0,8928	1,0000	110,09
$LPRP_t$	0,7480	1,0000	35,57
LC_t	0,8779	1,0000	94,15
$LPSO_t$	0,7002	1,0000	26,94
LO_t	0,7985	1,0000	49,29

$R^2=0,80$; $F_c=110,09$; SCE=5,88; F* ($LPRD_t$)=110,09

TABLA 8.12

Resultados en el paso 1 de la selección paso a paso de variables

Variable	*Coef. corr. parcial*	*Tolerancia*	*F**
$LPRP_t$	0,2896	0,4532	2,47
*LC_t	0,4581	0,2450	7,17
$LPSO_t$	0,2747	0,5333	2,20
LO_t	0,3349	0,3742	3,41

$R^2=0,84$; $F_c=70,77$; SCE=4,64; F* ($LPRD_t$)=11,63; F* (LC_t)=7,17

TABLA 8.13

Resultados en el paso 2 de la selección paso a paso de variables

Variable	*Coef. corr. parcial*	*Tolerancia*	*F**
$LPRP_t$	0,1736	0,4091	0,81
$LPSO_t$	−0,0701	0,2742	0,13
LO_t	0,1206	0,2679	0,38

tabla 8.10. En el siguiente paso, se introduce la variable LC deteniéndose el proceso, pues ninguna variable más presenta una F* para seleccionar superior a 4, ni tampoco una F* para eliminar menor de 3,9 (este valor no se presenta en la tabla). Por tanto, el modelo de regresión elegido viene dado por

$$LT_t = 0,531LPRD_t + 0,417LC_t$$

Con varianzas de los coeficientes de regresión igual a 0,156 en ambos.

Si se hubiera forzado a entrar todas las variables la suma de cuadrados del residuo disminuirá aunque esta disminución no quiera decir que la significación de la regresión aumente.

TABLA 8.14

Suma de cuadrados del residuo para los distintos modelos de regresión

Variables incluidas en el modelo	*Suma de cuadrados*	
	Valor	*Decremento*
LPRD	5,888	
LPRD-LC	4,646	1,242
LPRD-LC-LPRP	4,506	0,140
LPRD-LC-LPRP-LPSO	4,464	0,042
LPRD-LC-LPRP-LPSO-LPRO	4,450	0,014
LPRD-LC-LPRP-LPSO-LPRO-LO	4,447	0,003

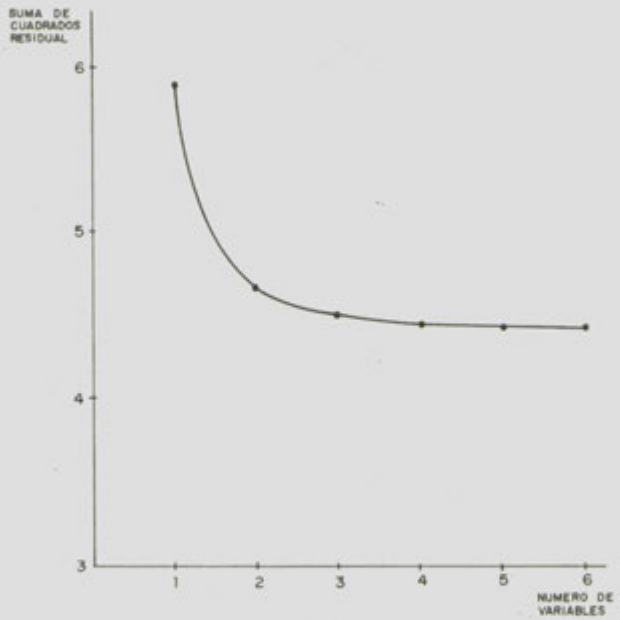


Figura 8.7.—Evolución de la suma de cuadrados residual al aumentar el número de variables en la ecuación.

Dicha disminución se incluye en la tabla 8.14 y en la figura 8.7. Se aprecia en la figura la evolución de la SCE al ir introduciendo variables sucesivamente, constatando que existe una gran disminución al incluir la segunda variable, pero que, a partir de entonces, cada vez va siendo menor, lo que indica que la información adicional aportada por las nuevas variables es despreciable.

Por último, si el investigador deseara mantener todas las variables en el modelo y obtener varianzas de los coeficientes de regresión más pequeñas, considerando como poco importante la correlación con los dos ejes principales, podría optar por seguir un camino de estimación más sofisticado, como la regresión «ridge». El método operativo de esta técnica se presenta a continuación.

Se calculan diferentes valores de los coeficientes de regresión, haciendo variar f en la expresión.

$$b^*(f) = (X'X+fI)^{-1}X'y$$

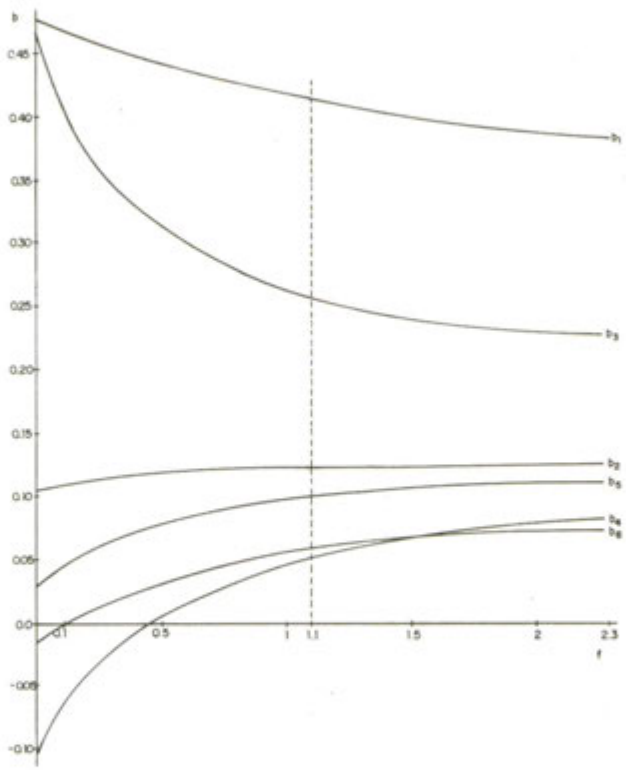


Figura 8.8.—«Ridge trace».

En la figura 8.8 se representa gráficamente la evolución de b (f) al variar f. A partir de f=1,1, se puede considerar que las curvas se encuentran bastante estabilizadas, por lo que se elige este valor para sustituirlo en la ecuación. Los valores de los coeficientes y sus errores típicos se encuentran en la tabla 8.15. Los valores de las varianzas de los estimadores «ridge» se obtuvieron como la diagonal de la matriz $\hat{\sigma}^2 (\chi'\chi+fI)^{-1}$ tomando σ^2 como el cuadrado medio del residuo de la regresión completa y f=1,1. En la tabla se aprecia que los coeficientes de regresión son bastante similares a las estimaciones mínimo cuadráticas (es decir, cuando f=0) pero los errores típicos han disminuido (comparar con tabla 8.9), aunque es necesario reconocer que estos estimadores son sesgados, tal como se indicó en 3.3.2.2.

TABLA 8.15

Coeficientes de regresión basados en la estimación «ridge» y sus errores típicos

Variable	*Coeficiente*	*Errores típicos*
$LPRD_t$	0,42	0,161
$LPRP_t$	0,11	0,127
LC_t	0,26	0,272
$LPSO_t$	0,06	0,209
LO_t	0,10	0,155
$LPRO_t$	0,06	0,131

8.4. Ejemplo 3: Selección de variables

El ejempo que sigue, presenta la problemática de los métodos de selección de variables de una forma intuitiva, sugiriendo al investigador ciertas gráficas que puedan ayudarle a tomar o revisar decisiones sobre el camino a seguir para modelizar su experimento e interpretar los resultados.

Debe tenerse presente la relación de compensación que existe entre la firmeza de las hipótesis que el investigador desea mantener, y la potencia de los métodos estadísticos para sugerir interpretaciones. Si se conoce el esquema que se quiere seguir, dudando solamente de algún ajuste parcial (por ejemplo, estudiar la significación de una determinada variable), el proceso estadístico es relativamente sencillo, pues se reducirá a la utilización de una simple prueba estadística (generalmente F o t). Ahora bien, si se desea dar una cierta flexibilidad a la investigación permitiendo a los métodos estadísticos sugerir caminos de interpretación, el procedimiento que se muestra seguidamente puede ser útil.

TABLA 8.16

Datos del ejemplo 8.3

% de enjambres enfermos (Y)	*Temperaturas medias en °F*	
	enero (x_1)	*junio* (x_2)
49	35	53
40	35	53
41	38	50
46	40	64
52	40	70
59	42	68
53	49	59
61	46	73
55	50	59
64	50	71

En este ejemplo se desea investigar la influencia del clima (concretamente la temperatura) en la incidencia de una enfermedad de las abejas. Los datos

obtenidos se encuentran en la tabla 8.16. El modelo «completo» de regresión que se pretende ensayar es

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

y el objetivo es escoger un submodelo que sea el óptimo de acuerdo con unos ciertos criterios, comparando dichos criterios de optimización.

A efectos de cuadros y gráficos se denotará la variable x_1, por 1, la x_2 por 2, la x_1^2 por $\dot{1}$, la x_2^2 por $\dot{2}$ y el producto $x_1 x_2$ por P.

8.4.1. *Método ascendente*

Como punto de partida para el análisis se calcula la matriz de correlaciones simples incluida en la tabla 8.17. Para la discusión de la matriz, conviene

TABLA 8.17

Matriz de correlaciones simples de los datos del ejemplo 8.3

	Y	1	$\dot{1}$	2	$\dot{2}$	P
Y	1,000					
1	0,796	1,000				
$\dot{1}$	0,789	0,998	1,000			
2	0,804	0,581	0,568	1,000		
$\dot{2}$	0,801	0,572	0,549	0,998	1,000	
P	0,907	0,882	0,870	0,893	0,889	1,000

separarla en dos partes: Por un lado, la primera columna representa las correlaciones de la variable dependiente con cada una de las regresoras; por otro lado el resto, que indica los valores de las correlaciones simples de las variables regresoras entre sí (realmente deberían llamarse «seudocorrelaciones», ya que las regresoras no son variables aleatorias), en el primer grupo hay que destacar que los valores son, en general, bastante altos (0,796—0,907), lo que implica una gran influencia de cada una de las regresoras sobre la variable dependiente. La ordenación de mayor a menor de las variables con respecto a su correlación con la Y es P, 2, $\dot{2}$, 1, $\dot{1}$, lo cual indica que, en caso de querer un modelo con una sola variable debería elegirse como más informativo aquel que incluyera la P (es decir, $x_1 x_2$), y en cualquier proceso de selección ascendente, sería la primera variable que entraría en el modelo.

En el segundo grupo de correlaciones sobresalen dos valores muy elevados que corresponden, como resulta natural, a las correlaciones de $\dot{1}$ con 1 y de $\dot{2}$ con 2. Por otro lado, existen cuatro correlaciones comparativamente bajas que son, por orden de mayor a menor, 1 con 2 y con $\dot{2}$ y $\dot{1}$ con 2 y con $\dot{2}$, todo lo cual parece sugerir dos bloques (1, $\dot{1}$) y (2, $\dot{2}$) con correlaciones bajas interbloque y altas intrabloque; no obstante, no se puede concluir ningún tipo de «independencia» entre los bloques, ya que P es la variable más significativa.

Finalmente, queda la fila de las correlaciones de la P con las demás variables de cuyos valores (0,87—0,89) podría intuirse que una vez incluida la P en el modelo, no interesa añadir ninguna variable más, pues aportarían poca información adicional. Sin embargo, esta intuición puede no ser cierta, tal como se demostrará a continuación.

Parece que un coeficiente que tiene interés para estudiar si una vez introducida P en el modelo ($Y=\beta_0+\beta_{12}P+\varepsilon$) debe incluirse otra variable x_j, es $r_{Yxj.P}$ (es decir, correlación entre Y y x_j, corregidas ambas por la influencia de P). Este coeficiente indica qué parte de Y no explicada (linealmente) por P, puede ser explicada (linealmente) por x_j.

La intuición a la que se hacía referencia, se basaba en el valor r_{xjP} obtenido de la matriz de correlaciones deduciendo que, por ser alto, $r_{Yxj.P}$ sería pequeño, teniendo en cuenta que ya el r_{YP} era también alto; pero ¿cuál es la relación entre r_{xjP} y $r_{Yxj.P}$? Por definición de correlación parcial, la expresión es:

$$r_{Yx_j.P} = \frac{r_{Yx_j} - r_{YP}r_{x_jP}}{\sqrt{(1-r^2_{YP})(1r^2_{x_jP})}}$$

Para el ejemplo presente y particularizando para $x_j=x_1$ se tiene la siguiente ecuación:

$$r_{Y1.P} = \frac{r_{Y1}-r_{YP}r_{1P}}{\sqrt{(1-r^2_{YP})(1-r^2_{1P})}} = \frac{0,796 - 0,907\ r_{1P}}{\sqrt{(1-0,907^2)(1-r^2_{1P})}} =$$

$$= \frac{0,796 - 0,907\ r_{1P}}{0,4211\ \sqrt{1 - r^2_{1P}}}$$

Dando valores a r_{1P} se puede observar cómo varía $r_{Y1.P}$. Esta gráfica se representa en la figura 8.9 acotando el gráfico para $r_{Y1.P}$ entre -1 y 1. En ella

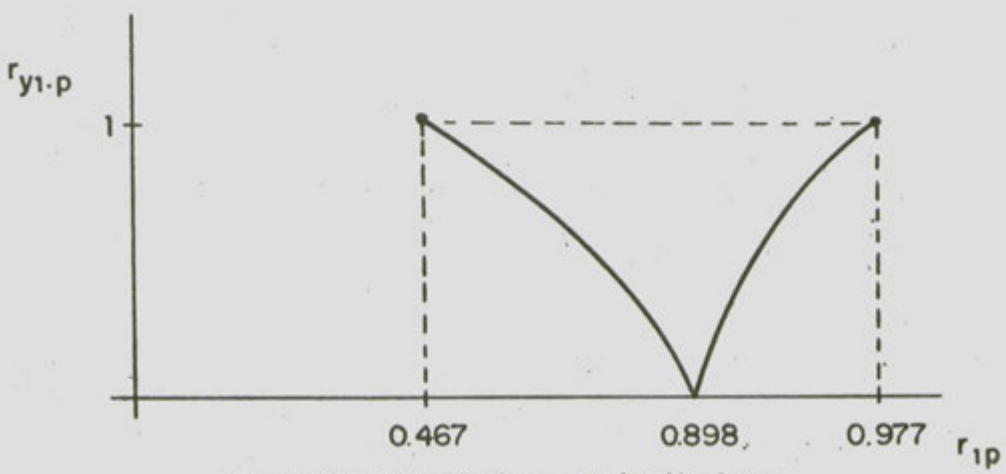


Figura 8.9.—Variación de $r_{Y1.P}$ en función de r_{1P}

se muestra cómo para valores de r_{1P} entre 0,467 y 0,898 (correspondientes a $r_{Y1.P}$ entre 1 y 0), la función es decreciente en valor absoluto; es decir, a valores altos de r_{1P} corresponden valores bajos de $r_{Y1.P}$, tal como la intuición indicaba; sin embargo, a partir de $r_{1P}=0{,}898$, el valor absoluto de la correlación $r_{Y1.P}$ empieza a crecer y en el punto $r_{1P}=0{,}977$, la correlación $r_{Y1.P}$ vale exactamente -1 (es decir, correlación perfecta). Por tanto, para valores de r_{1P} cercanos a 0,977 resulta que $r_{Y1.P}$ es negativo pero alto, lo que indica que la intuición habría llevado al investigador a cometer un temendo error.

De todas maneras, estas conclusiones deben mirarse con cautela, pues en el desarrollo a través de correlaciones se supone implícitamente que Y, x_j y P son tres variables aleatorias (para que en rigor se pueda hablar de correlación).

El enfoque a través de la correlación parcial es similar al presentado en 3.2.2 y en la figura 3.1, estudiando el tanto por ciento de variación explicado por x_j cuando P está ya en la ecuación; sin embargo, no coinciden plenamente. Recuérdese que la fórmula que allí se explicó era:

$$\%(x_j/P) = \frac{100\ (r_{YP}-r_{x_jP})^2}{(1-r^2_{x_jP})}$$

que comparada con la de $r_{Yxj.P}$ se observa que en aquélla, además de estar elevada al cuadrado, no aparece el término $1-r^2_{YP}$.

Esta diferencia de resultados se puede comprobar fácilmente recurriendo a la interpretación geométrica de la figura 7.16.

Para ayudar en la decisión de qué variable debe incluirse en el modelo, pueden efectuarse diagramas y gráficos que visualicen la situación concreta en la que se encuentra un proceso. Un ejemplo de estos diagramas se encuentra en la figura 8.10.

Además ,este método de selección ascendente de variables puede seguirse a través de matrices de correlaciones parciales. Por ejemplo, una vez seleccionada la variable P, la nueva matriz de correlaciones sería la presentada en la tabla 8.18.

Sin embargo, existen métodos alternativos para el proceso de selección de variables, que fueron abordadas en el capítulo 4. Para este ejemplo concreto, los resultados comparativos se encuentran en la figura 8.11, cuya terminología se explica a continuación. Los números situados a la derecha de cada conjunto de variables (submodelo) son el coeficiente de determinación expresado en tanto por ciento (100 R^2), y la probabilidad de significación de la F de Snedecor, también en tanto por ciento (100p). Un modelo será tanto mejor cuanto mayor R^2 y, sobre todo, cuanto menor p posea.

Los caminos seguidos por este método se simbolizarán por ($\rightarrow\rightarrow\rightarrow$) en la figura 8.11.

Como se deduce de la matriz de correlaciones simples (tabla 8.17) la primera variable que entra en el modelo es la P y, según la matriz de correlaciones parciales de la tabla 8.18 la siguiente variable a ser incluida sería la 2 (si bien no parece interesante incluirla, se ha realizado el proceso ascendente hasta el final para poderlo comparar con el descendente). Claramente, a un

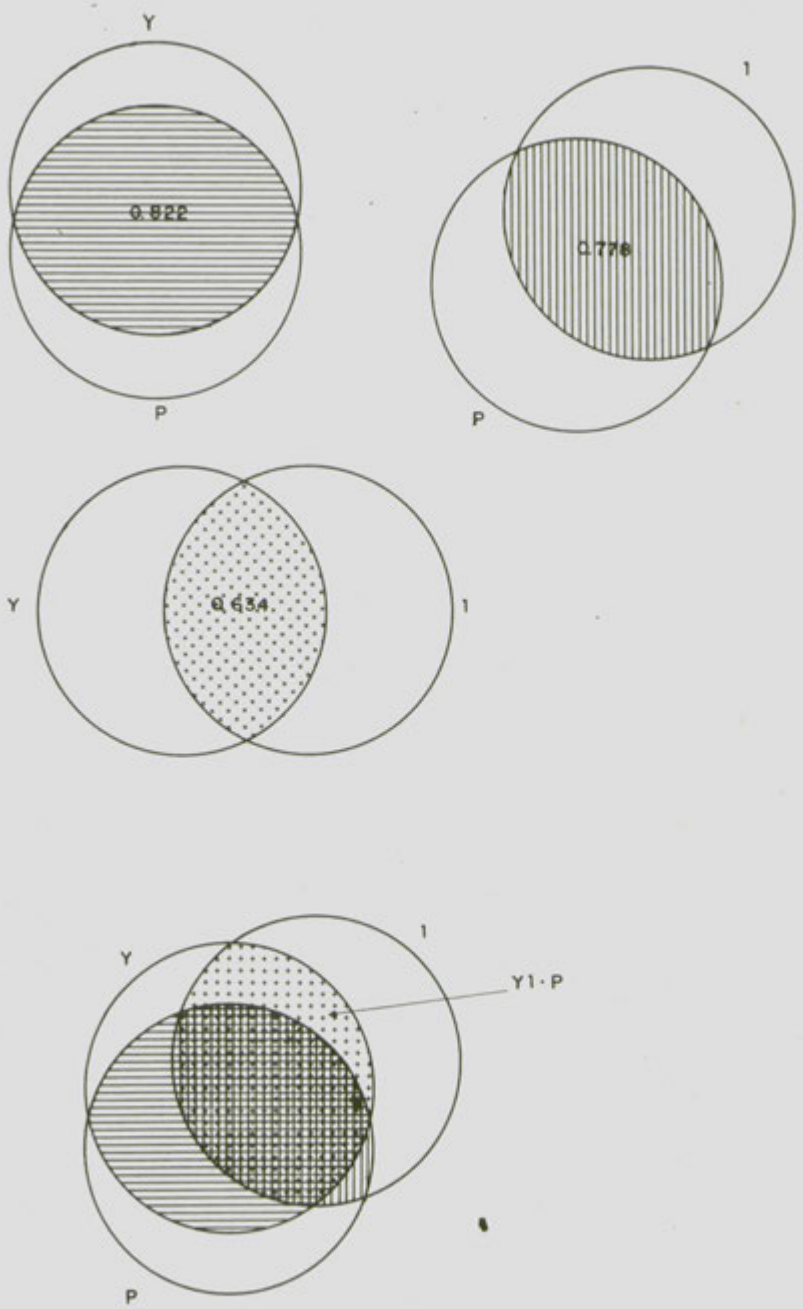


Figura 8.10.—Visualización de la matriz de correlalciones totales y parciales.

determinado nivel de significación, el proceso se detendría en un punto; para los valores más usuales del error α el modelo elegido sería $Y=\beta_0+\beta_m P+\epsilon$.

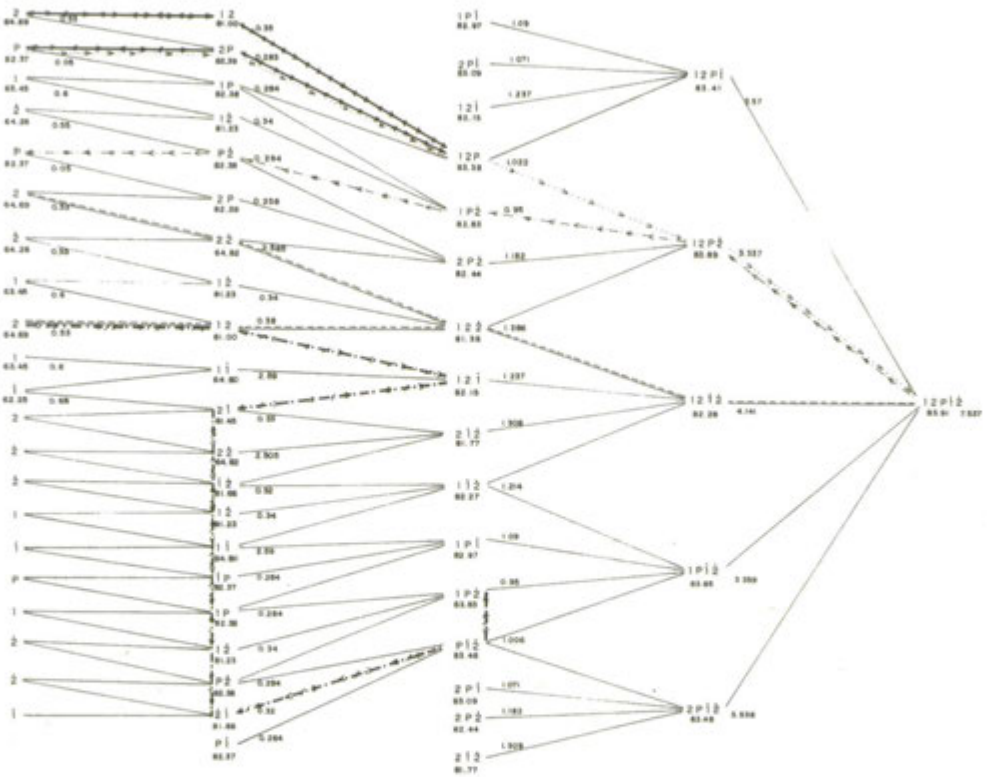


Figura 8.11.—Diagrama de caminos según distintos métodos de selección de variables.

TABLA 8.18

Matriz de correlaciones parciales cuando la variable P se encuentra en el modelo

	Y	1	1	2	2
Y	1,000				
1	−0,021	1,000			
1	−0,003	−0,914	1,000		
2	−0,033	−0,977	−0,942	1,000	
2	−0,270	−0,967	−0,994	−0,991	1,000

Analizando el método con los datos de la figura 8.11 se aprecia que, de todos los modelos con una variable, aquel que incluye la P es el mejor porque obtiene el mayor R^2 (82,37%) y mínima probabilidad (0,05%). Es necesario reconocer que la utilidad de añadir la probabilidad en el cuadro (y no el valor F_c) es para comparar modelos con distinto número de variables (y por tanto, distintos grados de libertad para la prueba F), ya que el R^2 aumenta siempre al añadir variables en el modelo, aunque la significación, que viene indicada por la probabilidad, disminuya (un p más pequeño indica una mejor significación). Obsérvese, por ejemplo, que el modelo con P se pasaría al modelo con 2 y P aumentando el R^2 al 82,39%, pero con significación menor (0,28%), por lo que sería recomendable detener el proceso en el primer paso.

De todas formas, si lo que se persiguiera fuera un aumento del R^2, aún a costa de la significación del modelo, podría proseguirse y el compromiso que el investigador estableciera entre R^2 y p sería quien determinase el punto en donde se detendría el proceso.

8.4.2. *Método descendente*

En la figura 8.11 se designa el camino seguido por este método por (←←←).

Los parámetros del modelo completo son R^2=83,91% y p=7,537%, que indican un modelo altamente explicativo pero poco significativo. Esta aparente contradicción se debe a que con un modelo de menos variables puede obtenerse un R^2 muy cercano, pero con una mejor significación. Así, con cuatro variables, 83,89%; con tres, 83,83%; con dos, 82,38% y con una, 82,37%. Todo ello indica que «no compensa» la introducción de todas las variables en el modelo.

El método descendente presentado en 4.5 no merece aquí más comentarios que el de su comparación con el ascendente: obsérvese que a nivel de una, cuatro o cinco variables ambos caminos, en este ejemplo particular, coinciden; pero con tres variables es mejor el descendente y con dos el ascendente (medido por el valor p de la combinación de variables considerada). Este hecho lleva a la conclusión de sugerir que el método ascendente es más conveniente cuando se desee un modelo con pocas variables y, por el contrario, el descendente es recomendable cuando se pretenda mantener un modelo lo más completo posible. En el presente ejemplo y, suponiendo que el investigador deja flexibilidad a la estadística, sería el ascendente claramente mejor.

8.4.3. *Método paso a paso*

El método de paso a paso, tal como se presentó en 4.7, consiste en un avance y retroceso dinámico que proporciona una mayor flexibilidad en la búsqueda del modelo óptimo, independizándose más con respecto a las soluciones iniciales; es decir, sea el método descendente en su primer paso (12P$\hat{1}\hat{2}$), claramente 12P$\hat{2}$ es el mejor modelo de cuatro variables, pero ¿estará el óptimo de tres variables entre los modelos que derivan de él (12P, 12$\hat{2}$, 1P$\hat{2}$, 2P$\hat{2}$) y que son los únicos a los cuales puede conducir el proceso en el siguiente paso? Teniendo en cuenta el carácter de la regresión, en el que una variable muy importante a un nivel puede dejar de serlo en el siguiente, no se puede dar una respuesta afirmativa; en el ejemplo sí se cumple para tres variables pero ya no para dos. Este es, pues, un inconveniente común tanto al método ascendente como al descendente.

La solución parece evidente: construir todas las regresiones posibles para

cada número de variables, elegir la mejor y luego compararlas según un determinado criterio. Para el ejemplo tratado, la lista de los cinco «mejores» modelos para cada número de variables junto al coeficiente C_p (ver 4.9) se incluyen en la Tabla 8.19. Ahora bien, el seguir este camino presenta dos inconvenientes importantes. Primero, el costo que puede suponer el cálculo de gran cantidad de regresiones y, segundo, el grafo del camino seguido puede resultar inconexo y, por ello, menos claro para algunas interpretaciones. Esta es la razón de que pueda resultar útil el método paso a paso, pues al ser dinámico, puede recuperarse alguna combinación interesante que se hubiera perdido al seguir métodos tan inflexibles como el ascendente o descendente.

TABLA 8.19

Relación de los «mejores» modelos según el número de variables incluidas

1 VARIABLE		2 VARIABLES		3 VARIABLES		4 VARIABLES		5 VARIABLES	
Modelo	*Cp*	*Modelo*	*Cp*	*Modelo*	*Cp*	*Modelo*	*Cp*	*Modelo*	*Cp*
P	−1,62	P2	0,38	1P2	2,02	12P2	4,00	12p12	6,0
2	2,78	P2	0,38	12P	2,13	1P12	4,02		
2	2,88	1P	0,38	2P1	2,20	2P12	4,10		
1	3,08	P1	0,38	2P2	2,36	11P1	4,12		
1	3,38			121	2,44	1212	1,40		

Si se emplea al método paso a paso de forma usual, dada la fuerza explicativa de la variable P, será introducida en el modelo deteniéndose ahí el proceso. Por esta razón no queda reflejado el proceso en la figura 8.11.

8.4.4. *Paso a paso con sugerencia de variables*

En la figura 8.11 se representa el camino seguido en la selección de variables usando este método por (⟷ ⟷ ⟷).

Supóngase que el investigador está «ligeramente» interesado en introducir en el modelo las variables simples 1 y 2 (pues le son de más fácil interpretación) antes que cualquier otra. El término «ligeramente» indica que no quiere forzar la introducción de esas variables, sino sólo sugerirlas, estando dispuesto a eliminarlas si los resultados estadísticos lo consideran pertinente.

El método actuará de la forma siguiente:

1.º Selecciona (2) porque es mejor que (1).

2.º Añade 1 pasando a (1 2).

3.º Termina la «sugerencia» de variables y al actuar libremente introduce P pasando a (1 2 P).

4.º Elimina la 1 y el modelo (2 P) mejora en significación al (1 2 P) (p de 0,283% frente a 1,022%).

5.º Elimina la 2 y el modelo (P) mejora en significación al (2 P) (p de 0,05% frente a 0,283%).

Así pues, se aconseja al investigador que no insista en introducir «1 y 2» y se queda únicamente con P.

Los pasos pueden resumirse así:

Resumen de la operación	*Carácter de la operación*	% R^2	*Variación del % P*
1.º Introduce en el modelo la {2}	——→	64,69	0,55
2.º Añade 1 quedando {1, 2} ...	——→	81,00	0,35
3.º Añade P quedando {1 2 P}.	——→	83,38	1,022
4.º Elimina 1 quedando {2 P}.	←——	82,39	0,283
5.º Elimina 2 quedando {P} ...	←——	82,37	0,05

8.4.5. *Paso a paso con salto*

Este es el método que añade la opción SWAP al paso a paso (ver 4.8 y Anejo 5).

La diferencia con el anterior consiste en que, antes de detenerse el proceso, se prueban todos aquellos modelos que difieran con el actual en una sola variable, dentro de los del mismo número de variables; es decir, actúa en el último paso como un método de todas las regresiones posibles. Si algún modelo se declara como «mejor», sustituye al previamente obtenido reiniciándose un paso a paso. En el ejemplo presente, este «salto» o cambio de modelo final ocurrirá dos veces.

8.4.6. *Paso a paso con salto y forzamiento de variables*

Este método se representa en la figura 8.11 por (← →← →← →).

Se ha supuesto para representar este caso que el investigador está, como en 8.4.4 interesado fundamentalmente en las variables simples 1 y 2, prefiriendo luego las cuadráticas 1 y 2 a la P. En los primeros pasos, orientado por el investigador, el proceso se comporta así:

Resumen de la operación	*Carácter de la operación*	% R^2	*Variación del % P*
1.º Toma la variable {2}	——→	64,49	0,55
2.º Introduce 1 pasando a {1, 2}	——→	81,00	0,35

En esta primera etapa ha introducido las dos variables simples.

En este punto, el proceso se pararía. Si el investigador, consciente de que ha forzado el modelo, disminuye el valor F de comparación para seleccionar una variable, favoreciendo su entrada con el fin de añadir un nuevo dinamismo al proceso, para que a través de modelos «malos» pueda conducirle a uno mejor (idea que en este ejemplo se revela como provechosa) ocurrirán los siguientes pasos:

Resumen de la operación	*Carácter de la operación*	% R^2	*Variación del % P*
do por 1 P 2	salto	83,83	0,95
3.º Introduce $\dot{1}$ pasando a $\{1\ 2\ \dot{1}\}$	——→	82,15	1,237
4.º Elimina 1 pasando a $\{2\ \dot{1}\}$.	←——	81,45	0,33
5.º Cambia 2 por $\dot{2}$ saltando a $\{\dot{1}\ \dot{2}\}$	salto	81,66	0,32

En la segunda etapa pasa de las variables simples a las variables cuadráticas.

Si el investigador «libera» el proceso de su encauzamiento en este punto, el método se apresurará a introducir P y la evolución se produce así:

Resumen de la operación	*Carácter de la operación*	% R^2	*Variación del % P*
6.º Introducción de P pasando a P $\dot{1}$ $\dot{2}$	——→	83,48	1,006
7.º Cambio de $\dot{1}$ por 1 pasando por 1 P $\dot{2}$	salto	83,83	0,95

Es necesario recordar que 1P$\dot{2}$ con p=0,95% es mucho menos significativo que P con 0,5%, lo cual representa, en este caso, una seria desventaja para el método, debido a que la F_0 necesaria para que el modelo se «desprendiera» de 1 y 2 es superior a la F_0 que se introdujo en el paso 3 para que entrara $\dot{1}$.

No conviene abusar de este método, aunque resulta muy revelador para sugerir caminos en la obtención de modelos suficientemente buenos, debido a la elevación de el error α global al emplear una y otra vez los mismos datos experimentales para probar modelos alternativos.

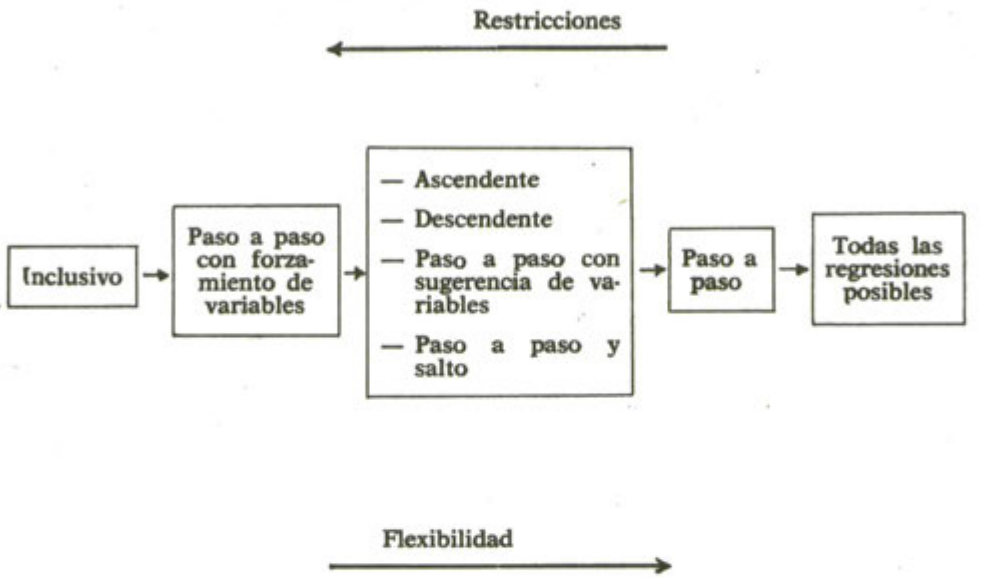
8.4.7. *Modelos inclusivos*

Este método ,explicado en el capítulo 5, queda representado en la figura 8.11 por (— — —). Es el método más rígido de todos desde el punto de vista estadístico. Aquí, el investigador establece uno o varios caminos, habitualmente coincidentes en principio y fin, y la aportación de los resultados estadísticos consiste únicamente en determinar en qué punto es conveniente terminar; en cierto sentido, puede tratarse como ascendente o descendente pero, dado que se considera como modelo «verdadero» el completo, debería utilizarse el descendente.

En la figura 8.11 se ha seguido el mismo conjunto de inclusiones que el explicado en 5.4.1, aplicado a los datos de la tabla 8.16.

7.5.8. *Comentario final*

Como se explicó al comienzo de este ejemplo, han sido desarrollados métodos distintos entre los que el investigador puede elegir teniendo en cuenta la beligerancia que está dispuesto a conceder a la estadística, siempre sobre la base de la idea general que se quiera dar a los diferentes métodos. Clasificándolos de izquierda a derecha por orden decreciente de imposición del investigador y creciente, por tanto, de beligerancia estadística, se podrían representar de la forma siguiente:



ANEJOS

ANEJO 1

Valores crticos para la prueba de Durbin-Watson

n	k' = 1		k' = 2		k' = 3		k' = 4		k' = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

n = número de observaciones
k' = número de variables explicativas

Nivel de significación al 5%.
Esta tabla está reproducida de J. Johnston, «Métodos Econométricos», Vicens-Vives, 1979, con permiso del autor y editor.

ANEJO 1 (Conclusión)

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

n = número de observaciones
k' = número de variables explicativas

Nivel de significación al 1%.
Esta tabla está reproducida de J. Johnston, «Métodos Econométricos», Vicens-Vives, 1979, con permiso del autor y editor.

ANEJO 2

Valores críticos para la prueba de Burr-Foster

	ν = 1		ν = 2		ν = 3		ν = 4	
p	.99	.999	.99	.999	.99	.999	.99	.999
3	*	*	.863	*	.757	.919	.684	.828
4	.920	*	.720	.898	.605	.754	.549	.675
5	.828	*	.608	.773	.512	.644	.443	.552
6	.744	.949	.539	.690	.430	.546	.369	.461
7	.671	.865	.469	.606	.372	.471	.318	.394
8	.609	.793	.412	.537	.325	.411	.276	.342
9	.576	.750	.371	.481	.287	.363	.244	.300
10	.528	.694	.333	.433	.257	.324	.218	.267
12	.448	.598	.276	.358	.211	.265	.179	.217
14	.391	.522	.234	.303	.178	.222	.151	.181
15	.365	.490	.217	.280	.165	.205	.140	.167
16	.343	.460	.202	.261	.154	.190	.130	.155
18	.304	.409	.178	.228	.135	.165	.114	.135
20	.273	.367	.158	.202	.120	.146	.101	.119
22	.246	.332	.142	.180	.108	.130	.090	.106
24	.224	.302	.129	.162	.098	.117	.082	.096
26	.206	.276	.118	.148	.090	.107	.075	.087
28	.190	.254	.108	.135	.082	.098	.069	.080
30	.176	.234	.100	.124	.075	.090	.064	.074
32	.163	.218	.093	.115	.070	.083	.060	.068
36	.143	.189	.082	.100	.062	.072	.052	.060
40	.127	.167	.072	.088	.055	.064	.047	.053
45	.111	.145	.063	.076	.048	.055	.041	.046
50	.098	.127	.056	.067	.043	.049	.037	.042
60	.080	.102	.045	.053	.035	.039	.030	.033
64	.074	.094	.042	.049	.033	.037	.028	.031

Esta tabla está reproducida de V. L. Anderson y R. A. McLean, «Design of experiments: A realistic approach», Marcel Dekker, 1974, con permiso de los autores y editores.

ANEJO 2 (Conclusión)

	ν = 5		ν = 6		ν = 8		ν = 10	
p	.99	.999	.99	.999	.99	.999	.99	.999
3	.631	.760	.593	.708	.539	.633	.512	.596
4	.498	.608	.461	.558	.413	.490	.383	.446
5	.399	.490	.368	.446	.328	.388	.303	.351
6	.334	.407	.307	.368	.271	.318	.250	.288
7	.284	.345	.261	.311	.230	.268	.212	.242
8	.246	.298	.226	.268	.199	.231	.184	.209
9	.217	.261	.199	.235	.176	.202	.162	.183
10	.194	.232	.178	.208	.157	.179	.145	.163
15	.123	.145	.113	.131	.101	.113	.094	.103
20	.090	.104	.083	.094	.074	.082	.069	.075
30	.058	.065	.053	.059	.048	.052	.045	.048
40	.042	.047	.039	.043	.035	.038	.033	.035
50	.033	.036	.031	.033	.028	.030	.026	.028
60	.027	.029	.025	.027	.023	.024	.022	.023

	ν = 12		ν = 14		ν = 16		ν = 20	
p	.99	.999	.99	.999	.99	.999	.99	.999
3	.486	.558	.466	.530	.451	.508	.429	.476
4	.362	.415	.347	.393	.335	.375	.319	.351
5	.287	.326	.275	.308	.265	.295	.252	.276
6	.236	.267	.227	.253	.219	.242	.209	.226
7	.201	.225	.192	.213	.186	.204	.178	.191
8	.174	.194	.167	.184	.162	.176	.154	.166
9	.154	.170	.148	.162	.143	.155	.136	.146
10	.137	.152	.132	.144	.128	.138	.122	.130
15	.089	.097	.086	.092	.083	.089	.080	.084
20	.066	.070	.063	.067	.062	.065	.059	.062
30	.043	.045	.042	.043	.040	.042	.039	.040
40	.032	.033	.031	.032	.030	.031	.029	.030
50	.025	.026	.024	.025	.024	.025	.023	.024
60	.021	.022	.020	.021	0.20	.020	.019	.020

Esta tabla está reproducida de V. L. Andersson y R. A. McLean, «Design of experiments: A realistic approach», Marcel Dekker, 1974, con permiso de los autores y editores.

ANEJO 3

Coeficientes para la prueba de Shapiro-Wilk

i \ n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2		.0000	.1677	.2413	.2806	.3031	.3164	.3244	.3291
3				.0000	.0875	.1401	.1743	.1976	.2141
4						.0000	.0561	.0947	.1224
5								.0000	.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	.3315	.3325	.3325	.3318	.3306	.3290	.3273	.3253	.3232	.3211
3	.2260	.2347	.2412	.2460	.2495	.2521	.2540	.2553	.2561	.2565
4	.1429	.1586	.1707	.1802	.1878	.1939	.1988	.2027	.2059	.2085
5	.0695	.0922	.1099	.1240	.1353	.1447	.1524	.1587	.1641	.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			.0000	.0240	.0433	.0593	.0725	.0837	.0932	.1013
8					.0000	.0196	.0359	.0496	.0612	.0711
9							.0000	.0163	.0303	.0422
10									.0000	.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	.3185	.3156	.3126	.3098	.3069	.3043	.3018	.2992	.2968	.2944
3	.2578	.2571	.2563	.2554	.2543	.2533	.2522	.2510	.2499	.2487
4	.2119	.2131	.2139	.2145	.2148	.2151	.2152	.2151	.2150	.2148
5	.1736	.1764	.1787	.1807	.1822	.1836	.1848	.1857	.1864	.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	.1092	.1150	.1201	.1245	.1283	.1316	.1346	.1372	.1395	.1415
8	.0804	.0878	.0941	.0997	.1046	.1089	.1128	.1162	.1192	.1219
9	.0530	.0618	.0696	.0764	.0823	.0876	.0923	.0965	.1002	.1036
10	.0263	.0368	.0459	.0539	.0610	.0672	.0728	.0778	.0822	.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12			.0000	.0107	.0200	.0284	.0358	.0424	.0483	.0537
13					.0000	.0094	.0178	.0253	.0320	.0381
14							.0000	.0084	.0159	.0227
15									.0000	.0076

Esta tabla está reproducida de V. L. Andersson y R. A. McLean, «Design of experiments: A realistic approach», Marcel Dekker, 1974, con permiso de los autores y editores.

ANEJO 3 (Conclusión)

i \ n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	.2921	.2898	.2876	.2854	.2834	.2813	.2794	.2774	.2755	.2737
3	.2475	.2463	.2451	.2439	.2427	.2415	.2403	.2391	.2380	.2368
4	.2145	.2141	.2137	.2132	.2127	.2121	.2116	.2110	.2104	.2098
5	.1874	.1878	.1880	.1882	.1883	.1883	.1883	.1881	.1880	.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	.1433	.1449	.1463	.1475	.1487	.1496	.1505	.1513	.1520	.1526
8	.1243	.1265	.1284	.1301	.1317	.1331	.1344	.1356	.1366	.1376
9	.1066	.1093	.1118	.1140	.1160	.1179	.1196	.1211	.1225	.1237
10	.0899	.0931	.0961	.0988	.1013	.1036	.1056	.1075	.1092	.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	.0585	.0629	.0669	.0706	.0739	.0770	.0798	.0824	.0848	.0870
13	.0435	.0485	.0530	.0572	.0610	.0645	.0677	.0706	.0733	.0759
14	.0289	.0344	.0395	.0441	.0484	.0523	.0559	.0592	.0622	.0651
15	.0144	.0206	.0262	.0314	.0361	.0404	.0444	.0481	.0515	.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17			.0000	.0062	.0119	.0172	.0220	.0264	.0305	.0343
18					.0000	.0057	.0110	.0158	.0203	.0244
19							.0000	.0053	.0101	.0146
20									.0000	.0049

i \ n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	.2719	.2701	.2684	.2667	.2651	.2635	.2620	.2604	.2589	.2574
3	.2357	.2345	.2334	.2323	.2313	.2302	.2291	.2281	.2271	.2260
4	.2091	.2085	.2078	.2072	.2065	.2058	.2052	.2045	.2038	.2032
5	.1876	.1874	.1871	.1868	.1865	.1862	.1859	.1855	.1851	.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	.1531	.1535	.1539	.1542	.1545	.1548	.1550	.1551	.1553	.1554
8	.1384	.1392	.1398	.1405	.1410	.1415	.1420	.1423	.1427	.1430
9	.1249	.1259	.1269	.1278	.1286	.1293	.1300	.1306	.1312	.1317
10	.1123	.1136	.1149	.1160	.1170	.1180	.1189	.1197	.1205	.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	.0891	.0909	.0927	.0943	.0959	.0972	.0986	.0998	.1010	.1020
13	.0782	.0804	.0824	.0842	.0860	.0876	.0892	.0906	.0919	.0932
14	.0677	.0701	.0724	.0745	.0765	.0783	.0801	.0817	.0832	.0846
15	.0575	.0602	.0628	.0651	.0673	.0694	.0713	.0731	.0748	.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	.0379	.0411	.0442	.0471	.0497	.0522	.0546	.0568	.0588	.0608
18	.0283	.0318	.0352	.0383	.0412	.0439	.0465	.0489	.0511	.0532
19	.0188	.0227	.0263	.0296	.0328	.0357	.0385	.0411	.0436	.0459
20	.0094	.0136	.0175	.0211	.0245	.0277	.0307	.0335	.0361	.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22			.0000	.0042	.0081	.0118	.0153	.0185	.0215	.0244
23					.0000	.0039	.0076	.0111	.0143	.0174
24							.0000	.0037	.0071	.0104
25									.0000	.0035

Esta tabla está reproducida de V. L. Andersson y R. A. McLean, «Design of experiments: A realistic approach», Marcel Dekker, 1974, con permiso de los autores y editores.

ANEJO 4

Valores críticos para la prueba de Shapiro-Wilk

	Level								
n	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	.687	.707	.748	.792	.935	.987	.992	.996	.997
5	.686	.715	.762	.806	.927	.979	.986	.991	.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	.730	.760	.803	.838	.928	.972	.979	.985	.988
8	.749	.778	.818	.851	.932	.972	.978	.984	.987
9	.764	.791	.829	.859	.935	.972	.978	.984	.986
10	.781	.806	.842	.869	.938	.972	.978	.983	.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	.805	.828	.859	.883	.943	.973	.979	.984	.986
13	.814	.837	.866	.889	.945	.974	.979	.984	.986
14	.825	.846	.874	.895	.947	.975	.980	.984	.986
15	.835	.855	.881	.901	.950	.975	.980	.984	.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	.851	.869	.892	.910	.954	.977	.981	.985	.987
18	.858	.874	.897	.914	.956	.978	.982	.986	.988
19	.863	.879	.901	.917	.957	.978	.982	.986	.988
20	.868	.884	.905	.920	.959	.979	.983	.986	.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	.878	.892	.911	.926	.961	.980	.984	.987	.989
23	.881	.895	.914	.928	.962	.981	.984	.987	.989
24	.884	.898	.916	.930	.963	.981	.984	.987	.989
25	.888	.901	.918	.931	.964	.981	.985	.988	.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	.894	.906	.923	.935	.965	.982	.985	.988	.990
28	.896	.908	.924	.936	.966	.982	.985	.988	.990
29	.898	.910	.926	.937	.966	.982	.985	.988	.990
30	.900	.912	.927	.939	.967	.983	.985	.988	.900
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	.904	.915	.930	.941	.968	.983	.986	.988	.990
33	.906	.917	.931	.942	.968	.983	.986	.989	.990
34	.908	.919	.933	.943	.969	.983	.986	.989	.990
35	.910	.920	.934	.944	.969	.984	.986	.989	.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	.914	.924	.936	.946	.970	.984	.987	.989	.990
38	.916	.925	.938	.947	.971	.984	.987	.989	.990
39	.917	.927	.939	.948	.971	.984	.987	.989	.991
40	.919	.928	.940	.949	.972	.985	.987	.989	.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	.922	.930	.942	.951	.972	.985	.987	.989	.991
43	.923	.932	.943	.951	.973	.985	.987	.990	.991
44	.924	.933	.944	.952	.973	.985	.987	.990	.991
45	.926	.934	.945	.953	.973	.985	.988	.990	.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	.928	.936	.946	.954	.974	.985	.988	.990	.991
48	.929	.937	.947	.954	.974	.985	.988	.990	.991
49	.929	.937	.947	.955	.974	.985	.988	.990	.991
50	.930	.938	.947	.955	.974	.985	.988	.990	.991

Esta tabla está reproducida de V. L. Andersson y R. A. McLean, «Design of experiments: A realistic approach», Marcel Dekker, 1974, con permiso de los autores y editores.

ANEJO 5
Descripción de los programas de la serie BMDP

Descripción resumida de los programas de la serie BMDP existentes en la Sección de Proceso de Datos del INIA relacionados con el Análisis de Regresión.

5.1. BMDP1R: Regresión lineal múltiple

Este programa calcula una regresión lineal múltiple a partir de las pruebas de una matriz de datos con un máximo de 147 variables. No tiene incorporadas las pruebas para la verificación de las hipótesis de base, si bien la normalidad se puede comprobar de una manera gráfica a través del estudio de los residuos. La ecuación de regresión puede incluir o no el término independiente β_0. Si se especifica una variable agrupadora para delimitar subconjuntos de datos, el programa efectúa una regresión para cada subconjunto comparando entre sí los diversos modelos obtenidos. Algunas observaciones pueden venir afectadas de unos pesos o incluso ser eliminadas del análisis.

La salida del programa incluye, para cada conjunto o subconjuntos de datos, la siguiente información:

— Medias, desviaciones típicas, valores máximos y mínimos para todas las variables consideradas.

— Coeficentes de correlación múltiple y de determinación.

— Tabla completa del Análisis de la Varianza excluyendo la media.

— Coeficientes de regresión, sus desviaciones típicas, significatividad de la variable regresora cuando el resto están presentes en la ecuación y coeficientes de regresión tipificados.

Como salida opcional se puede obtener:

— Matriz de varianzas-covarianzas y de correlaciones de los datos.

— Gráfica del ajuste comparando los valores observados con los esperados por el modelo para cada variable regresora.

— Gráficas para el estudio de los residuos.

5.2. BMDP2R: Regresión paso a paso según varios criterios

Selecciona un subconjunto de variables regresoras mediante el método de paso a paso (Stepwise), basándose en cuatro algoritmos diferentes que serán discutidos más adelante (F, FSWAP, R, RSWAP). A pesar de ser un paso a paso, permite que se fuerce la entrada obligatoria de alguna variable o que se sugiera algún orden de preferencia de entrada asignando diversos niveles. El método de regresión puede tener o no término independiente o incluso tratarlo como una variable más en el proceso de la selección, incluyéndolo en la ecuación solamente si es significativa su contribución.

Los algoritmos de selección se basan en los siguientes condicionantes o reglas:

a) Si hay por lo menos dos variables regresoras en la ecuación y una o o más de ellas tienen un valor F* menor que el valor predeterminado para eliminar variables, F_0, se eliminará de la ecuación aquella que tenga el menor F*.

b) Si dos o más variables están presentes en la ecuación, aquella con el menor F* se eliminará, si su exclusión proporciona un coeficiente de determinación mayor que el que se obtuvo previamente para el mismo número total de variables regresoras.

c) Si dos o más variables están en la ecuación, una de ellas será intercambiada por otra, no presente todavía en el modelo, si el cambio produce un aumento en el coeficiente de determinación.

d) Si una o más variables no están aún en la ecuación, aquella con el mayor F* superior al límite marcado para introducir una variable, será incluida en el modelo.

Estas cuatro reglas se combinan para dar lugar a los algoritmos de selección. Así:

F: Regla *a)*, seguida de la regla *d)*.

FSWAP: *Regla a)*, seguida de la regla *c)* y de la *d)*.

R: *Regla b)*, seguida de la regla *d)*.

RSWAP: Regla *b)*, seguida de la *c)* y de la *d)*.

Estas reglas se modifican ligeramente cuando existe forzamiento o sugerencia de variables por parte del usuario.

Este programa, aunque básicamente sigue el mismo procedimiento que el descrito en 4.7, tiene algunos matices que conviene señalar.

En el apartado 4.7 se menciona únicamente un valor F_0 utilizado indistintamente en el proceso de seleccionar una variable o para proceder a su eliminación; en este programa, por el contraio, se especifican dos valores, uno para incluir y otro para extraer una variable del modelo. Lógicamente, el valor F_0 para eliminar una variable debe ser menor o igual que el empleado para incluirla. Por defecto, si no se especifica nada en contra, los valores que automáticamente toma el programa son 4.0 y 3.9.

Igualmente en la descripción del método se indicó que se seleccionaba un valor F_0 (n_1, n_2; α) que dependería de n_1 y n_2 (grados de libertad) y α (nivel de significación). Al disponer de más o menos variables en el modelo, n_1 y n_2 variarán a lo largo del proceso. Sin embargo, en el programa, los valores F_0 para introducción o exclusión de variables se fijan al principio y permanecen invariables durante toda la ejecución. Por otro lado, los algoritmos R y RSWAP no se describieron en 4.7.

La salida del programa está constituida por:

— Medias, desviaciones típicas y otros estadísticos elementales para cada variable.

— Tabla del Análisis de la Varianza, coeficentes de correlación múltiple y de determinación para cada paso.

Como opcional, también se puede disponer de:

— Matriz de varianzas-covarianzas y correlaciones.

— Coeficientes de regresión, sus desviaciones típicas, significatividad, F* y tolerancia en cada paso.

— Tabla resumen con la ecuación seleccionada.

— Gráficos de residuos.

5.3. BMDP4R: Regresión sobre componentes principales

Este programa realiza una regresión de la variable dependiente sobre un conjunto de componentes principales calculados a partir de las variables regresoras originales. Los componentes principales se calculan utilizando los datos originales (matriz de varianzas-covarianzas) o tipificados (matriz de correlaciones). El análisis se efectúa siguiendo un procedimiento «paso a paso» y los coeficientes de regresión obtenidos se expresan en términos de los componentes principales o en términos de las variables regresoras originales.

La salida del programa incluye, entre otras, las informaciones siguientes:

— Estadísticos elementales y la matriz de covarianzas o de correlaciones.

— Valores y vectores propios.

— Coeficientes de regresión y suma de cuadrados de residuo.

5.4. BMDP9R. Todas las regresiones posibles

Este programa estima la ecuación de regresión para los «mejores» subconjuntos de variables predictoras y hace un completo análisis de residuos.

El mejor subconjunto se define en función del R-cuadrado, R-cuadrado ajustado, o empleando el criterio Cp de Mallows.

Por ejemplo, si se elige el R-cuadrado ajustado, los mejores subconjuntos son los que maximizan el R-cuadrado ajustado. Puede identificar además un número predeterminado ($M < 10$) de los mejores subconjuntos; esto es, no sólo el mejor, sino también el segundo mejor, el tercero, etc., proporcionando con ello varias alternativas buenas.

El criterio R-cuadrado identifica los M ($M < 10$) mejores subconjuntos para cada tamaño (el tamaño de un subconjunto es el número de variables regresoras incluidas en la ecuación).

Cuando se elige el criterio de R-cuadrado ajustado o el criterio Cp de Mallows, se identifican los M mejores subconjuntos en general, es decir, sin tener en cuenta el tamaño del subconjunto. Por ejemplo, si se desean los 10 mejores subconjuntos de un total de 20 variables independientes, puede que no se incluya ningún subconjunto de tamaño 1 ó de tamaño 20.

El BMDP9R puede analizar problemas que contengan hasta 27 variables y el costo del cálculo es comparable al costo de la regresión paso a paso. El número de variables analizadas puede ampliarse a 100 cuando se efectúa una regresión lineal múltiple sin selección de un subconjunto de variables.

La salida incluye:

- Estadísticos elementales para cada variable.
- Matriz de correlaciones.
- Para cada subconjunto de variables independientes:
  - R^2, R^2 ajustado y el valor Cp de Mallows.
  - Un análisis completo para el mejor subconjunto seleccionado siguiendo el criterio establecido, incluyendo:
    - R^2: correlación múltiple al cuadrado.
    - R^2 ajustado.
    - Cuadrado medio del error.
    - Estadístico F_c.
- Para cada variable en el mejor subconjunto, escribe:
  - El coeficiente de regresión.
    - Su error típico $s(b_i)$.
    - Coeficietne de regresión tipificador: $b_i \times s(b_i)/S(y)$.
    - El estadístico $t: b_i/s(b_i)$.
    - Nivel de significación en una distribución de dos colas para el estadístico t.
  - Resultados opcionales.
    - Matriz de varianzas covarianzas de las variables.
    - Valor residual y estimado.
    - Matriz de correlaciones de los coeficientes de regresión.
    - Salidas gráficas.

Para una mejor descripción del programa, consultar: CALVO, R. (1980).

REFERENCIAS BIBLIOGRAFICAS

ABRAHAM, B., y BOX, G. E. P., 1978. «Linear models and spurious observations». *J. Royal Statist. Soc. C.*, **27**, 131-138.

ANDERSON, T. W., 1958. *Introduction to Multivariate Statistical Analysis*, 374 p., Wiley, New York.

ANDERSON, V. L., y MCLEAN, R. A., 1974. *Design of Experiments: A realistic approach*, 418 p., Marcel Dekker, New York.

ANDREWS, D. F., y PREGIBON, D., 1978. «Finding the outliers that matter». *J. Royal Statist. Soc. B.*, **40**, 85-93.

BARTLETT, M. S., 1977. «Some examples of statistical methods of research in agriculture and applied biology». *J. Royal Statist. Soc., Supp.* **4**, 158-159.

BICKEL, P. J., 1978. «Using residuals robustly. I. Test for heteroscedasticiy, nonlinearity». *Ann. Statist.*, **6**, 266-291.

BOX, G. E. P., y COX, D. R., 1964. «An analysis of transformations». *J. Royal Statist. Soc. B*, **26**, 211-243.

BURR, I. W., y FOSTER, L. A., 1972. «A test for equality of variances». *Dept. of Statistic Mimeo Series N.° 282. Purdue University.*

CAILLEZ, F., y PAGES, J. P., 1976. *Introduction a l'Analyse des Données*, 616 p., SMASH.

CALVO, R. M., 1980. «Todos los posibles subconjuntos de regresión». *Nota Técnica número 6. Sección de Proceso de Datos. INIA.*

COOK, R. D., 1977. «Detection of influential observations in linear regression». *Technometrics*, **19**, 15-17.

DANIEL, C., y WOOD, F. S., 1980. *Fitting Equations to Data*, 458 p., Wiley, New York.

DRAPER, N. R., y SMITH, H., 1981. *Applied Regression Analysis*, 709 p., Wiley, New York.

DURBIN, J., 1970. «An alternative to the bounds test for testing for serial correlation in least-squares regression». *Econometrica*, **38**, 422-429.

DURBIN, J., y WATSON, G. S., 1951. «Testing for serial correlation in least-squares regression». *Biometrika*, **38**, 159-178.

EFROYMSON, M. A., 1960. «Multiple regression analysis». In A. Ralston and H. S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, vol. 1:191-203.

GALTON, F., 1886. «Family likeness in stature». *Proc. Royal Soc. London*, **40**, 42-72.

GUNST, R. F., y MASON, R. L., 1977. «Advantages of examining multicolinearities in regression analysis». *Biometrics*, **33**, 249-260.

HILL, R. C.; JUDGE, G. G., y FOMBY, T. B., 1978. «On testing the adecuacy of a regression model». *Technometrics*, **20**, 491-494.

HILL, R. C.; JUDGE, G. C., y FOMBY, T. B., 1980. «Is the regression equation adequate? A reply». *Technometrics*, **22**, 127-128.

HOCKING, R. R., 1976. «The analysis and selection of variables in linear regression». *Biometrics*, **83**, 1-50.

HOERL, A. S., y KENNARD, R. W., 1970a. «Ridge regression: biased estimation for non-orthogonal problems». *Technometrics*, **12**, 55-67.

HOERL, A. E., y KENNARD, R. W., 1970b .«Ridge regression: applications to non-orthogonal problems». *Technometrics*, **12**, 69-82.

JOHN, J. A., y DRAPER, N. R., 1980. «An alternative family of transformations». *Applied Statistics*, **29**, 190-197.

LEHMANN, E. I., 1975. *Nonparametrics: Statistical Methods based on ranks*, 457 p., McGraw-Hill, New York.

MALLOWS, C. L., 1964. «Choosing variables in a linear regression: a graphical aid». *Presented at the Central Regional Meeting of the IMS*, Manhattan, Kansas.

MALLOWS, C. L., 1966. «Choosing a subset regression». *Presented at Joint Statistical Meetings*. L. A. California.

MALLOWS, C. L., 1973. «Some comments on Cp». *Technometrics*, **15**, 661-675.

MARQUARDT, D. W., y SNEE, R. D., 1975. «Ridge regression in practice». *The American Statistitian*, *29*, 3-20.

NARULA, S. C., y WELLINGTON, J. F., 1977. «Prediction, linear regression and the minimum sums of relative errors». *Technometrics*, **19**, 185-190.

OSTLE, B., y MENSING, R. W., 1975. *Statistics in Research*, 596 p., The Iowa State Unversity Press.

PEARSON, E. S.; D'AGOSTINO, R. B., y BOWMAN, K. O., 1977. «Test for departure from normality. Comparison of process». *Biometrika*, **64**, 231-246.

SEARLE, S. R., 1971. *Linear models*, 532 p., Wiley, New York.

SHAPIRO, S. R., y WILK, M. B., 1965. «An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.

SIEGEL, S., 1956. *Non-parametric statistics for the Behavioral Sciences*, 312 p., McGraw-Hill, New York.

SNEE, R. D., 1977. «Validation of regression models: Methods and examples». *Technometrics*, **22**, 125-126.

SPRENT, P., 1969. *Models in Regression and related topics*, 173 p., Methuen, London.

SUICH, R., y DERRINGER, G. C., 1977. «Is the regression equation adequate? One criterion». *Technometrics*, **22**, 213-216.

SUICH, R., y DERRINGER, G. C., 1980. «Is the regression equation adequate? A further note». *Technometrics*, **22**, 125-126.

TAYLOR, L., 1974. «Estimation by minimizing the sums of absolute errors». In *Frontiers in Econometrics*, Academic Press.

THOMPSON, M. L., 1978a. «Selection of variables in multiple regression. I. A review and evaluation». *Int. Statist. Rev.*, **46**, 1-19.

THOMPSON, M. L., 1978b. «Selection of variables in multiple regression. II. Chosen procedures, computatations and examples». *Int. Statist. Rev.*, **46**, 129-146.

WEBSTER, J. T.; GUNST, R. F., y MASON, R. L., 1974. «Latent root regression analysis». *Technometrics*, **16**, 513-522.

WETZ, J. M., 1964. «Criteria for judging adequacy of estimation by approximating response functions». *Ph. D. Thesis. University of Wisconsin.*

WONNACOTT, T. H., y WONNACOTT, R. J., 1981. *Regression: A second course in Statistics*, 556 p., Wiley, New York.

Servicio de Publicaciones Agrarias
Ministerio de Agricultura, Pesca y Alimentación
P°. Infanta Isabel, 1 - Madrid - 7