"REGRESIÓN Y CORRELACIÓN"

6.1 Parte básica

6.1.1 Introducción

Regresión es una palabra un tanto rara. La utilizan los biólogos, los médicos, los psicólogos... y suena como "ir hacia atrás", "volver al pasado", y realmente este es verdadero significado del vocablo.

Fue un biólogo y estadístico inglés, SIR FRANCIS GALTON*, quien introdujo en 1889 el término ***regresión*** en Estadística. Empleó este concepto para indicar la relación que existía entre la estatura de los niños de una muestra y la estatura de su padre. Observó, que si los padres son altos, los hijos generalmente también lo son, y si los padres son bajos los hijos son también de menor estatura. Pero ocurría un hecho curioso: cuando el padre es muy alto o muy bajo, aparece una perceptible "*regresión*" hacia la estatura media de la población, de modo que sus hijos *retroceden* hacia la media de la que sus padres, por cierto, están muy alejados. Hoy día, el término no se utiliza en ese sentido.

En muchas ocasiones, se desea conocer algo acerca de la relación o dependencia entre dos características cuantitativas, o másde una, consideradas sobre la misma población objeto de estudio (por ejemplo la talla y el peso). Hay muchos casos en los que ya de antemano se "sospecha" que puede existir algún tipo de relación, y por consiguiente, se pretende saber por ejemplo, en el caso de que tengamos únicamente dos variables:

1.- Si ambas variables están realmente relacionadas entre sí o si, por el contrario, pueden considerarse independientes.

2.- Si existe dependencia, es necesario conocer el "grado de relación", así como el "tipo" de relación entre ambas.

3.- Si puede predecirse la variable que es considerada como dependiente a partir de los valores de la otra, que es considerada independiente, y si es así, con qué precisión.

* GALTON, F. (1889). Natural Inheritance. London. Mcmillan & Co.

6.1.2 ¿Cuándo existe regresión?

De una forma general, lo primero que suele hacerse para ver si dos variables aleatorias están relacionadas o no (de ahora en adelante las llamaremos X e Y, denotando con Y a la variable dependiente, y X a la variable independiente o regresora), consiste en tomar una muestra aleatoria. Sobre cada individuo de la muestra se analizan las dos características en estudio, de modo que para cada individuo tenemos un para de valores (x_i, y_i) $(i=1,...,n)$.

Seguidamente, representamos dichos valores en unos ejes cartesianos, dando lugar al diagrama conocido como diagrama de dispersión o nube de puntos. Así, cada individuo vendrá representado por un punto en el gráfico, de coordenadas, x_i, y_i.

De esa forma, podremos obtener una primera idea acerca de la forma y de la dispersión de la nube de puntos.

Al dibujar la nube de puntos, podemos encontrarnos, entre otros, los casos a los que hace referencia la figura 6.1.

En primer lugar deberemos distinguir entre **dependencia funcional** y **dependencia estocástica**. En el primer caso la relación es perfecta: Y=f(X) (ver figura 6.1d y e); es decir, los puntos del diagrama de dispersión correspondiente, aparecen sobre la función Y=f(X). Por ejemplo, el caso de la figura 6.1d sería Y=a+bX.

Sin embargo, lo que suele ocurrir es que no existe una dependencia funcional perfecta, sino otra dependencia o relación menos rigurosa que se denomina dependencia estocástica (figura 6.1b y c); entonces, la relación entre X e Y, podríamos escribirla (en el caso de la figura 6.1.b) de la forma Y=a+bX+e, donde e es un error o un residual, debido por ejemplo, a no incluir variables en el modelo que sean importantes a la hora de explicar el comportamiento de Y, y cuyos efectos sean diferentes a los de X; errores aleatorios o de medida, o simplemente a que estamos especificando mal el modelo (por ejemplo, que en lugar de ser una recta, sea una parábola).

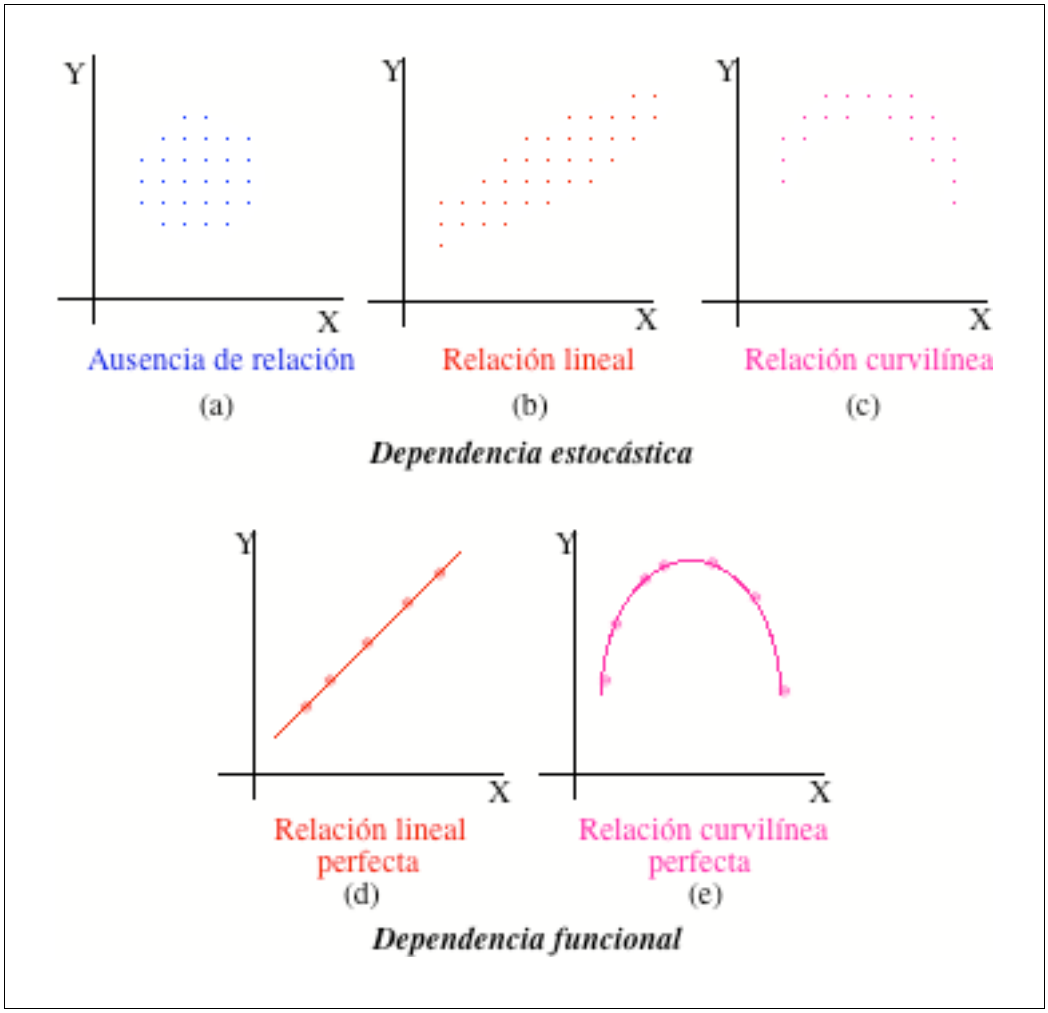


Figura 6.1: Tipos de relación entre dos variables X e Y

El caso de la figura 6.1a se corresponde con el de ausencia de relación, o independencia.

En la dependencia estocástica, se distinguen dos tipos de técnicas:

1.- Análisis de Regresión
2.- Análisis de Correlación*

El Análisis de correlación, tiene como fin dar respuesta a las preguntas:

a.- ¿Existe dependencia estocástica entre las variables?

b.- ¿Cuál es el grado de dicha dependencia?

* El orden de exposición de los dos Análisis es arbitrario. El orden para su estudio puede invertirse.

El Análisis de regresión, :

a.- ¿Cuál es el tipo de dependencia entre las dos variables?

b.- ¿Pueden estimarse los valores de Y a partir de los de X?. ¿Con qué precisión?.

De modo general, diremos que **existe regresión** de los valores de una variable con respecto a los de otra, cuando hay alguna línea, llamada **línea de regresión** que se ajusta más o menos claramente a la nube de puntos.

Si existe regresión, a la ecuación que nos describe la relación entre las dos variables la denominamos ***ecuación de regresión***.

Por ejemplo: $Y=a+bX$

$Y=a+bX+cX^2$

En general, la variable X se conoce como variable ***independiente***, y la Y como variable ***dependiente.***

Evidentemente puede ser arbitrario el determinar la existencia de regresión así como el tipo de la misma, ya que depende del autor o del estado de ánimo de la persona en un momento determinado.

Por lo tanto, se hacen necesarios métodos estadísticos objetivos, independientes del investigador, para determinar la existencia o no de relación y el tipo de la misma.

6.1.3 Tipos de regresión

Si las dos variables X e Y se relacionan según un modelo de línea recta, hablaremos de ***Regresión Lineal Simple***: $Y=a+bx$.

Cuando las variables X e Y se relacionan según una línea curva, hablaremos de ***Regresión no lineal o curvilínea***. Aquí podemos distinguir entre *Regresión parabólica, Exponencial, Potencial*, etc.

Cuando tenemos más de una variable independiente (X_1, X_2,..., X_p), y una sola variable dependiente Y, hablaremos de ***Regresión múltiple***, que se estudiará en detalle

en el apartado 6.2. A las variables X_i, se las denomina, regresoras, predictoras o independientes.

6.1.3.1 Consideraciones previas

En el Primera Unidad Didáctica, hemos analizado cómo varía cada una de las variables por separado. Sería interesante también tener idea de cómo varían dichas variables conjuntamente, es decir, cómo "*covarían*".

> Se dice que dos variables están ***variando conjuntamente, y en el mismo sentido,*** cuando al crecer los valores de una de las variables también aumentan los de la otra.
>
> En cambio, están ***variando conjuntamente, pero en sentido contrario,*** cuando al aumentar los valores de una, los de la otra disminuyen.

Definiremos como covarianza de dos variables X e Y, y denotaremos por S_{XY}, el estadístico que nos permite analizar la variación conjunta de dos variables. Viene dado por la siguiente expresión:

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

Si cada pareja de observaciones (x_i,y_i) se repitiese un número de veces, deberíamos introducir en la expresión anterior la correspondiente frecuencia, análogamente a como se hace en la expresión de la varianza.

La covarianza, puede ser utilizada como una medida inicial de la asociación lineal entre las dos variables. Para ello, observaremos detenidamente el gráfico de la figura 6.2.

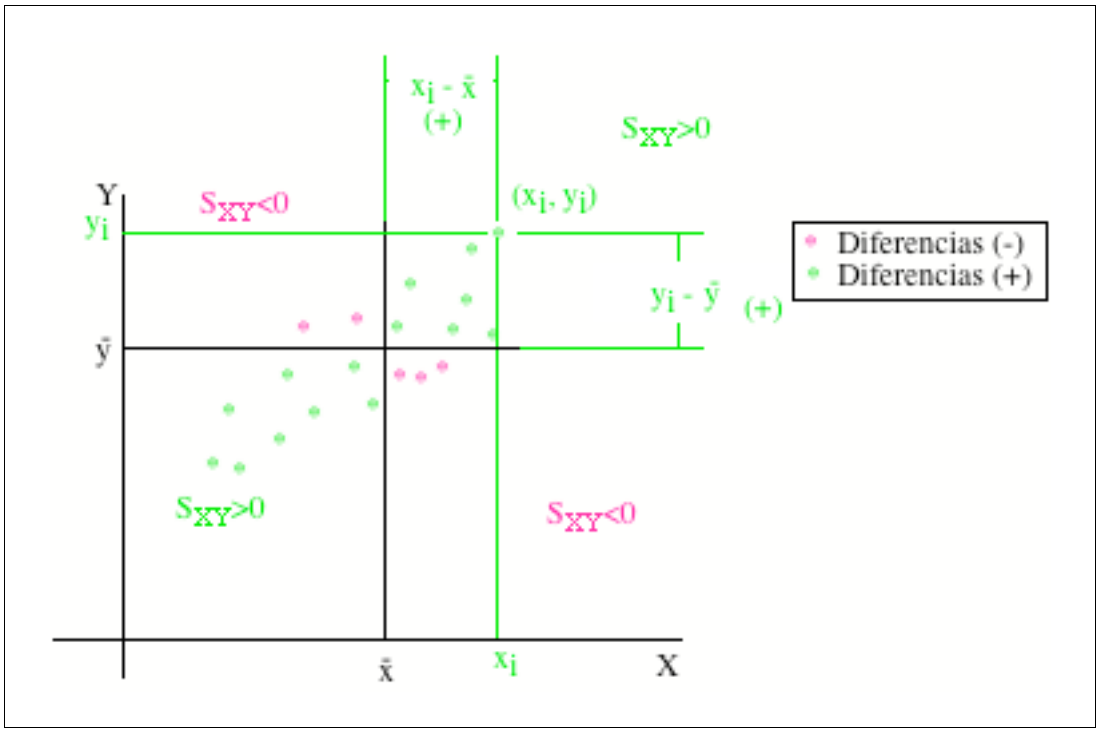


Figura 6.2: Gráfico que pone de manifiesto la importancia de la covarianza como medida de la asociación lineal

En ella aparece la nube de puntos para un par de variables X e Y. Se pone de manifiesto cómo aquellos pares de valores que ocupan el cuadrante superior derecho (tomando como origen el punto de medias) nos dan como resultado sumandos positivos en la expresión de la covarianza. Lo mismo ocurre con aquellos que se encuentran en el cuadrante inferior izquierdo. Sin embargo, los del cuadrante superior izquierdo e inferior derecho, nos dan sumandos negativos. Ello tiene como consecuencia, que dependiendo del número de observaciones situado en cada uno de dichos cuadrantes, obtendremos un signo diferente en la covarianza, de modo que si predominan las diferencias positivas, esta será positiva, y si predominan las negativas, la covarianza también lo será.

Esto nos lleva a utilizar la covarianza como una medida de la asociación lineal entre las variables, de modo que si ésta es positiva, nos indica una relación directa entre ellas y si es negativa, nos indica una relación inversa. Si las variables son independientes, entonces la covarianza es aproximadamente 0. Un ejemplo, de este último caso se correspondería con la figura 6.3a.

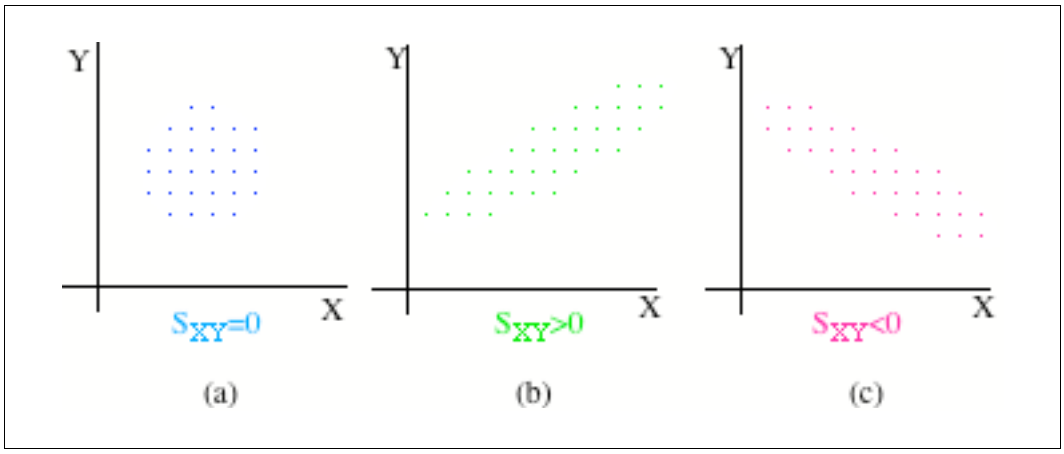


Figura 6.3: Diferentes casos para la covarianza

6.1.3.2 Regresión Lineal Simple

Nos centraremos en primer lugar, en el caso de que la función que relaciona las dos variables X e Y sea la más simple posible, es decir, una línea recta.

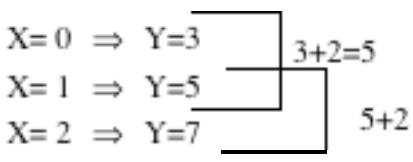
Por ello pasaremos a interpretar los coeficientes que determinan una línea recta.

Toda función de la forma Y=a+bX determina, al representarla en el plano una línea recta, donde X e Y son variables y a y b son constantes. Por ejemplo: Y=3+2X.

SIGNIFICADO DE a y b

a es la ordenada en el origen, es decir, es la altura a la que la recta corta al eje Y. Se denomina también ***término independiente.***

b, también denominada ***pendiente*** es la inclinación de la recta, es decir, es el incremento que se produce en la variable Y cuando la variable X aumenta una unidad. Por ejemplo, en el caso anterior Y=3+2X:



En la recta de regresión -como ya veremos- ***b*** recibe el nombre de ***Coeficiente de regresión.***

Si b>0, entonces cuando X aumenta Y también lo hace (relación directa).
Si b<0, entonces, cuando X aumenta Y disminuye (relación inversa).

Ver figura 6.4a y b respectivamente.

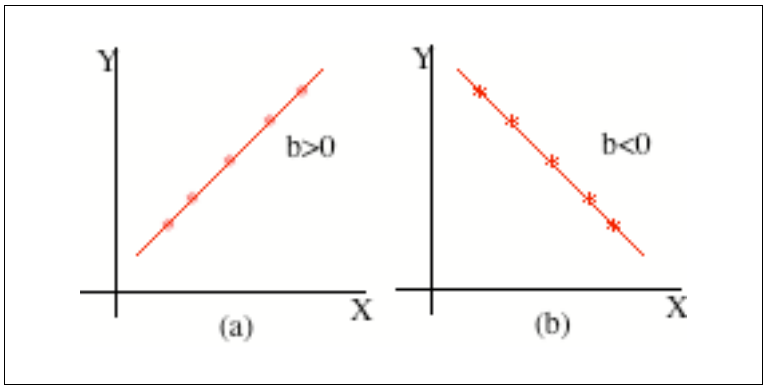


Figura 6.4: Signo de la pendiente en una recta de regresión

ESTIMACIÓN DE LA RECTA DE REGRESIÓN POR EL MÉTODO DE LOS MÍNIMOS CUADRADOS

Sean X e Y dos variables aleatorias medidas sobre los mismos individuos, y sean (x_i,y_i) los pares de observaciones sobre dichos individuos.

En primer lugar procederemos a representar el diagrama de dispersión, o nube de puntos. Supongamos que es la obtenida en la figura 6.5. Aunque la nube revele una gran dispersión, podemos observar una cierta tendencia lineal al aumentar X e Y (tendencia que no es del todo exacta; por ejemplo si suponemos que X es la edad e Y es la talla, obviamente, la talla no sólo depende de la edad, además también puede haber errores de medida).

Por esa nube de puntos podemos hacer pasar infinitas rectas. De todas ellas debemos elegir una ¿cual?... Obviamente elegiremos la mejor de todas en algún sentido.

La recta de regresión debe tener carácter de línea media, debe ajustarse bien a la mayoría de los datos, es decir, pasar lo más cerca posible de todos y cada uno de los puntos.

Llamaremos a la mejor de todas Y*=a+bX (Y* para distinguir los valores de la tabla de los que se habrían producido con la recta si la relación fuese funcional).

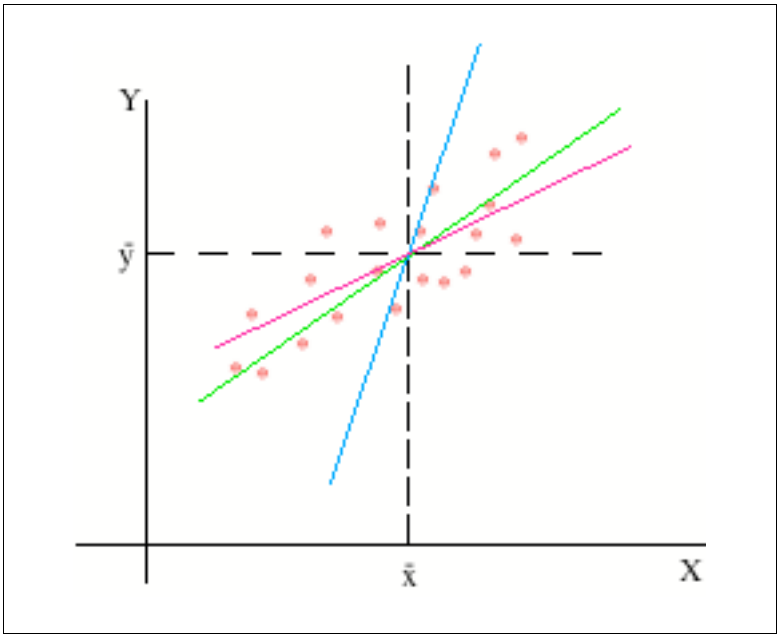


Figura 6.5: Nube de puntos y posibles rectas que pueden pasar por ella.

Que pase lo más cerca posible de todos los puntos, es decir que diste poco de todos y cada uno de ellos significa que hemos de adoptar un criterio particular que en general se conoce como MÍNIMOS CUADRADOS. Este criterio significa que la suma de los cuadrados de las distancias verticales de los puntos a la recta debe ser lo más pequeña posible (ver figura 6.6). (Obviamente, este es uno de los posibles criterios a adoptar, pero es el más utilizado).

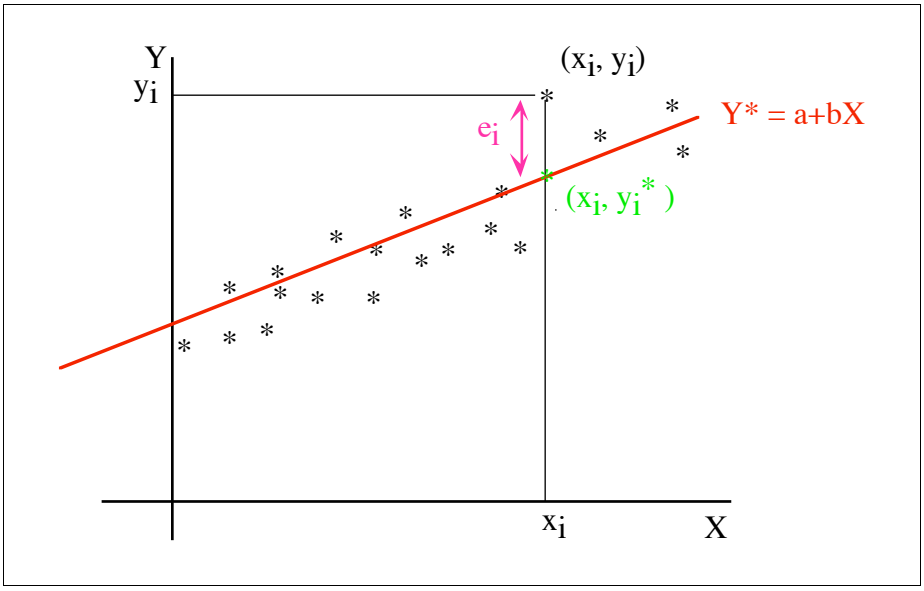


Figura 6.6: Recta de regresión mostrando los residuos o errores que se minimizan en el procedimiento de ajuste de los Mínimos cuadrados.

Estas distancias verticales se denominan errores o residuos.

Entonces el criterio puede expresarse:

$$D = \sum_{i=1}^{n} e_i \qquad \text{mínima}$$

Dado que la recta de regresión deberá tener carácter de línea media, esa suma de distancias deberá anularse (lo mismo que sucedía, como veíamos en la primera unidad didáctica al tratar de hallar la suma de las diferencias con respecto a la media aritmética). Por las mismas razones que entonces, para evaluar la dispersión, trabajaremos con esas distancias, pero al cuadrado, de modo que la función que deberemos minimizar será:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

donde y_i^* son los valores estimados según el modelo Y=a+bX

En la anterior expresión lo conocemos todo, excepto **a** y **b**. Para encontrar dichos valores, con la condición de que D sea mínima, deberemos hallar las derivadas parciales de D con respecto a **a** y a **b**, y resolver el sistema resultante, al igualar las ecuaciones obtenidas a 0. Es decir, el problema se reduce a un problema de mínimos.

Así, obtendremos:

$$\frac{\partial D}{\partial a} = 2 \sum_{i=1}^{n} \left(y_i - a - bx_i\right)(-1) = 0$$

$$\frac{\partial D}{\partial b} = 2 \sum_{i=1}^{n} \left(y_i - a - bx_i\right)(-x_i) = 0$$

Adecuando convenientemente las ecuaciones anteriores, obtenemos:

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right)(x_i) = 0$$

Operando y reorganizando términos, obtenemos las denominadas ***Ecuaciones Normales de Gauss***:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Resolviendo el sistema, obtenemos las expresiones para a y b:

$$a = \overline{y} - b\overline{x}$$

$$b = \frac{S_{XY}}{s_X^2}$$

La interpretación de **a** y **b**, es análoga a la que comentábamos en el apartado 6.1.3.2, sólo que como ya dijimos entonces, **b** recibe el nombre de ***Coeficiente de Regresión*.**

Como podemos observar, en el numerador de **b**, aparece la covarianza, y en el denominador la varianza de la variable independiente. Esto hace que el signo de **b** sea el mismo signo que el de la covarianza, por lo que si b>0, entonces, existe una relación directa entre las variables, y si b<0 entonces la relación es inversa.

En nuestro ejemplo de talla y edad, **b** sería el incremento medio que se produce en la talla, por cada incremento unitario de edad; si la edad está en años, por cada año aumente la edad.

Si queremos predecir un valor y_i a partir de un valor concreto de x_i, utilizaremos la expresión de la ecuación donde ahora ya, a y b son conocidos. No olvidemos que ese era uno de los objetivos del análisis, tratar de conocer valores de Y a partir de los de X:

$$y^*_i = a+bx_i$$

REPRESENTATIVIDAD DE LA RECTA DE REGRESIÓN.

- **Poder explicativo del modelo**

La recta de regresión, tiene carácter de línea media, como ya se ha señalado con anterioridad, tratando por lo tanto de resumir o sintetizar la información suministrada por los datos.

Si tiene carácter de linea media (de promedio, en definitiva), deberá ir acompañada *siempre* de una medida que nos hable de su representatividad, es decir, de lo buena que es la recta, ya que el haber obtenido la mejor de todas no da garantías de que sea buena.

Necesitamos, por tanto, una medida de dispersión, que tenga en cuenta la dispersión de cada observación con respecto a la recta, es decir, lo alejado que se encuentra cada punto de la recta.

Es decir, deberemos evaluar esas distancias verticales a la recta, es decir, los errores o residuales.

Si las dispersiones son pequeñas, la recta será un buen representante de la nube de puntos, o lo que es lo mismo, la ***bondad de ajuste del modelo será alta***. Si la dispersión es grande, la bondad de ajuste será baja.

Una forma de medir dicha bondad de ajuste es precisamente evaluando la suma de los cuadrados de los errores. Por tanto, llamaremos ***Varianza residual*** a la expresión:

$$S_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - y_i^* \right)^2}{n}$$

Si la varianza residual es grande, el modelo será malo, es decir, la recta no explicará el comportamiento general de la nube.

La fórmula práctica para el cálculo de la varianza residual, si el procedimiento de ajuste es el de los mínimos cuadrados es la siguiente:

$$S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n}$$

La cota máxima de la varianza residual es la varianza que tratamos de explicar mediante el modelo de regresión, es decir, la varianza de la variable dependiente. Por tanto, sin más que hacer relativa la varianza residual respecto de su máximo valor, y

multiplicando por 100, obtendremos el porcentaje de variaciones no explicado por el modelo:

$$\%\ \text{de var iaciones sin exp licar} = \frac{S_e^2}{s_y^2} 100$$

Ahora, ya es fácil obtener una media que nos indique el porcentaje de variaciones controladas o explicadas mediante el modelo, que se conoce como ***Coeficiente de Determinación***, que denotaremos con R^2. Su expresión en tantos por 1, será:

$$R^2 = 1 - \frac{S_e^2}{s_y^2}$$

Como puede observarse, a partir de la expresión anterior: $0 < R^2 < 1$. Por tanto:

Si $R^2=1$, entonces no hay residuos, habrá una dependencia funcional. Cuanto más se acerque dicho valor a la unidad, mayor ***poder explicativo*** tendrá el modelo de regresión.

Si $R^2=0$, X no explica en absoluto ninguna de las variaciones de la variable Y, de modo que o bien el modelo es inadecuado, o bien las variables son independientes. Cuanto más cercano a 0 esté dicho valor, menor poder explicativo.

❖ Poder explicativo vs poder predictivo

Un modelo de regresión con un alto porcentaje de variaciones explicado, puede no ser bueno para predecir, ya que el que la mayoría de los puntos se encuentren cercanos a la recta de regresión, no implica que todos lo estén, y puede ocurrir, que justamente para aquel rango de valores en el que el investigador está interesado, se alejen de la recta, y por tanto, el valor predecido puede alejarse mucho de la realidad.

La única forma de poder evaluar el poder predictivo del modelo es tras la observación y el análisis de los gráficos de residuales, es decir, de diagramas de dispersión, en los que en el eje de ordenadas se colocan los residuales, y en el eje de abscisas se colocan o bien X, Y, o Y*.

Sólo si la banda de residuales es homogénea, y se encuentran todos los puntos no demasiado alejados del 0 (aunque depende de la escala de medida), diremos, que un modelo con un alto poder explicativo, también es bueno para predecir.

Un análisis detallado de los residuales se realizará en la sección 6.2.

CAUSALIDAD

Es muy importante resaltar el hecho, de que un modelo sea capaz de explicar de manera adecuada las variaciones de la variable dependiente en función de la independiente, no implica que la primera sea causa de la segunda.

Es un error muy común confundir causalidad con *casualidad*. El hecho de que las variables estén relacionadas no implica que una sea causa de la otra, ya que puede ocurrir el hecho de que se esté dando una variación concomitante, por el simple hecho de que las dos son causa de una tercera. Por ejemplo, si realizamos un estudio en el que se analice el número de canas (X) y la presión arterial (Y), podríamos encontrar una relación lineal casi perfecta. Eso no significa que el tener canas aumente la presión arterial, lo que verdaderamente está ocurriendo es que es la edad, la causante, de que se tengan más canas y una tendencia a tener más alta la presión arterial.

EXTRAPOLACIÓN

Es importante, resaltar el hecho de que a la hora de hacer predicciones, no deben extrapolarse los resultados más allá del rango de la variable X utilizado para ajustar el modelo, ya que más allá de ese rango no sabemos qué puede estar ocurriendo.

Por todos es conocido que las plantas necesitan abono para poder crecer. Desde pequeños hemos aprendido que hay que abonarlas, de modo que en principio, cuanto más abono se les suministre más crecerán. Pero... ¿qué ocurriría si abonásemos demasiado el suelo?. Obviamente la planta moriría. Bien, esto se traduce, en que conforme aumenta la cantidad de abono, el crecimiento es más notable, pero a partir de un punto, la planta deja de crecer, y es más se muere. Esto queda reflejado en la figura 6.7. De ahí el peligro de extrapolar los resultados.

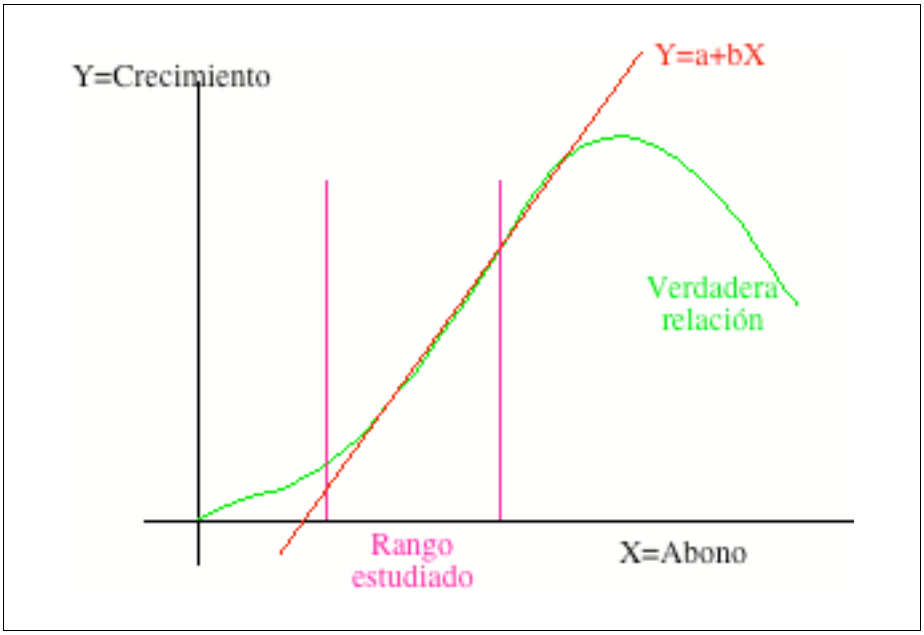


Figura 6.7: Comparación de una posible verdadera relación entre cantidad de abono y crecimiento de una planta, con los resultados de una recta de regresión obtenida mediante el estudio de un rango limitado de valores de abono.

6.1.3.3 Regresión no lineal

Supongamos que al hacer la representación gráfica correspondiente la distribución bidimensional, hemos obtenido la figura 6.1c. Se observa una clara relación entre las dos variables, pero desde luego, esa relación no es lineal.

Por tanto, debemos buscar la función que ha de describir la dependencia entre las dos variables.

Nos limitaremos al estudio de las más utilizadas: la función parabólica, la logarítmica, la exponencial y la potencial.

PARÁBOLA DE REGRESIÓN

En muchos casos, es una función de segundo grado la que se ajusta lo suficiente a la situación real dada.

La expresión general de un polinomio de 2º grado es:

$$Y=a+bX+cX^2$$

donde **a**, **b** y **c** son los parámetros.

El problema consiste, por tanto, en determinar dichos parámetros para una distribución dada. Seguiremos para ello, un razonamiento similar al que hicimos en el caso del modelo de regresión lineal simple, utilizando el procedimiento de ajuste de los mínimos cuadrados, es decir, haciendo que la suma de los cuadrados de las desviaciones con respecto a la curva de regresión sea mínima:

$$D = \sum_{i=1}^{n}(y_i - y_i^*)^2$$

donde, siguiendo la notación habitual, y_i son los valores observados de la variable dependiente, e y_i^* los valores estimados según el modelo; por tanto, podemos escribir D de la forma:

$$D = \sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2$$

Para encontrar los valores de a, b y c que hacen mínima la expresión anterior, deberemos igualar las derivadas parciales de D con respecto a dichos parámetros a cero y resolver el sistema resultante. Las ecuaciones que forman dicho sistema se conocen como ***ecuaciones normales de Gauss*** (igual que en el caso de la regresión lineal simple).

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

FUNCIÓN EXPONENCIAL, POTENCIAL Y LOGARÍTMICA

El problema de ajustar un modelo potencial, de la forma $Y=AX^b$ y uno exponencial $Y=AB^X$ se reduce al de la función lineal, con solo tomar logaritmos.

- **Modelo potencial:**

Si tomamos logaritmos en la expresión de la función potencial, obtendremos:

$$\log Y = \log A + b \log X$$

Como vemos es la ecuación de una recta: $Y=a+bX$, donde ahora $a = \log A$. De modo que el problema es sencillo, basta con transformar Y en logY y X en logX y ajustar una recta a los valores transformados. El parámetro b del modelo potencial coincide con el coeficiente de regresión de la recta ajustada a los datos transformados, y A lo obtenemos mediante el antilog(a).

- **Modelo exponencial:**

Tomando logaritmos en la expresión de la función exponencial, obtendremos:

$$\log Y = \log A + \log B\ X$$

También se trata de la ecuación de una recta $Y=a+bX$, pero ahora ajustándola a logY y a X; de modo que, para obtener el parámetro A del modelo exponencial, basta con hacer antilog(a), y el parámetro B se obtiene tomando antilog(b).

- **Modelo logarítmico:**

La curva logarítmica $Y = a + b \log X$ es también una recta, pero en lugar de estar referida a las variables originales X e Y, está referida a logX y a Y.

Hemos visto, cómo, a pesar de ser inicialmente modelos mucho más complejos que el de una recta, estos tres últimos se reducen al modelo lineal sin más que transformar adecuadamente los datos de partida.

6.1.4 Correlación

Como hemos visto con anterioridad, al analizar las relaciones existentes entre dos variables aleatorias cuantitativas, deberemos responder a las preguntas, de si existe dependencia estocástica entre ellas y de qué grado. El análisis de correlación nos dará respuesta a dichas preguntas.

6.1.4.1 Prueba de independencia de dos caracteres cuantitativos

Dos variables X e Y son independientes, es decir, no están relacionadas, cuando la variable Y tiene el mismo valor, en media, sea cual sea el valor de la variable X y viceversa. (Ver por ejemplo la figura 6.1a).

Como vimos en la sección 6.1.3.1, la covarianza podía ser un medida que nos habla de la dependencia entre las dos variables. Sin embargo, la covarianza presenta el inconveniente de que no se trata de una medida adimensional, y por lo tanto se hace necesario conocer la fuerza de la relación -si existe- así como poder realizar comparaciones entre parejas de variables que vienen medidas en unidades diferentes. Por ello, y dado que viene medida en unidades de la variable X por unidades de la variable Y, la dividimos entre las correspondientes desviaciones típicas, obteniendo así, el denominado ***Coeficiente de correlación lineal de Pearson*** y que denotamos con una r minúscula:

$$r = \frac{S_{xy}}{s_x s_y}$$

Es importante fijarnos en que hemos denominado a dicho coeficiente: coeficiente de correlación ***lineal*** de Pearson. El "apellido lineal" es conveniente utilizarlo porque dicho coeficiente solo tiene potencia para analizar si la relación entre las dos variables es o no de tipo lineal. Si las variables son independientes, es un hecho de que el coeficiente de correlación lineal debe ser cero. Sin embargo, si el coeficiente de correlación lineal es 0, no implica que las variables sean independientes, simplemente que la relación no es lineal.

Como vemos, el coeficiente de correlación lleva asociado el mismo signo que la covarianza, por lo que si éste resulta ser positivo, indicará que se trata de una relación lineal directa, mientras que si es negativo, la relación será inversa.

6.1.4.2 Relación entre r y R^2

Una propiedad sumamente importante del coeficiente de correlación **r** es que si el procedimiento de ajuste de la recta de regresión es el del criterio de los mínimos cuadrados, resulta:

$$\mathbf{r^2 = R^2}$$

En el apartado 6.1.3.2 vimos que el coeficiente de determinación era un valor acotado entre 0 y 1. Teniendo en cuenta la relación anterior, podemos asegurar que el coeficiente de correlación es un valor acotado entre -1 y +1. Si r=+1, existe una correlación positiva perfecta, y si r=-1, analogamente pero negativa (en ambos casos R^2=1, por lo tanto no hay errores, sería una dependencia funcional). A nivel muestral, es difícil encontrarnos con un valor de r = 0 aun cuando las variables sean independientes, de modo que podríamos pensar que cuanto más se acerque |r| a 1, el grado de relación entre X e Y será más fuerte. ¿Sin embargo, a partir de qué valor muestral de r decidiremos que las variables son independientes, y a partir de cuál diremos que están relacionadas?

6.1.4.3 Distribución del coeficiente de correlación muestral

Para dar respuesta a la pregunta anterior, se ha estudiado la ley de probabilidad de los coeficientes de correlación observados en muestras extraídas al azar de una población en la que se sabe que X e Y son independientes, es decir, que el coeficiente de correlación poblacional (ρ) es 0.

Al extraer muestras de dicha población, los coeficientes de correlación muestral obtenidos, fluctúan alrededor de cero en forma simétrica, lo cual no ocurre si ρ es distinto de cero. Por ello, se ha construido una tabla en la que aparece el valor de r, que sólo era superado en el 5% (o el 1%) de las muestras extraídas de la población con ρ=0; En la primera columna de la tabla aparece el tamaño de muestra n -2.

grados de libertad (n-2)	5%	1%	grados de libertad (n-2)	5%	1%
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

Tabla del coeficiente de correlación

Realmente no se trata más que de un contraste de hipótesis. La hipótesis nula es: Ho: $\rho=0$, de modo que la hipótesis se rechaza sólo si el coeficiente de correlación muestral es, en valor absoluto, mayor que el valor crítico de la tabla, al nivel de significación elegido, y con los grados de libertad adecuados, ya que sólo rechazaremos Ho si el valor muestral encontrado es poco probable que ocurra cuando $\rho=0$.

"EL MODELO LINEAL GENERAL"

6.2 Ampliación

6.2.1 Introducción

En la investigación práctica nos encontramos frecuentemente con situaciones en las que una variable, Y, viene determinada por otra u otras variables, X_1, X_2, ... , X_k , sin que a su vez la primera determine las últimas. Podemos escribir la relación como $Y = f(X_1, X_2, ... , X_k)$.

La variable Y es denominada dependiente, respuesta ó endógena mientras que las variables X se denominan independientes, predictoras o regresoras.

Utilizaremos este tipo de relaciones para:

- Predecir los valores de la respuesta (a partir de los de las regresoras).
- Determinar el efecto de cada predictora (sobre la respuesta).
- Confirmar, sugerir o refutar relaciones teóricas.

Conocida la posible dependencia entre las variables tendremos que determinar la forma de la relación, generalmente sugerida a través de la teoría de la materia objeto de estudio o través de la revisión de experimentos anteriores.

La forma más usada en la práctica es aquella en la que podemos suponer que el modelo es lineal en sus parámetros o al menos que podemos linealizarlo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Debido a la naturaleza de los fenómenos estudiados es necesario introducir un error procedente de:

-No incluir variables importantes.
-Errores aleatorios y errores de medida.
-Especificación incorrecta de la forma de la ecuación.

En realidad solamente el segundo de los supuestos es realmente admisible como término de perturbación aleatoria.

El modelo real será entonces:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

donde ε es el error o perturbación aleatoria y los coeficientes $\beta_0, \beta_1, \beta_2, ... , \beta_k$ son los parámetros estructurales o estructura paramétrica de la relación propuesta.

El modelo propuesto en el que las X_i son variables observables no aleatorias y los β_i son constantes fijas desconocidas se denomina Modelo Lineal General (MLG).

Se presenta aquí una introducción que trata de mostrar los aspectos más relevantes de la teoría así como algunos aspectos prácticos importantes olvidados generalmente en los libros de teoría. No olvidemos que el objeto final de lo que vamos a ver es la aplicación a datos reales en la investigación aplicada en campos tan diversos como el Diseño de Experimentos o la Econometría.

Para ampliar el tema, una excelente revisión teórica puede encontrarse en SEBER (1977)* ; una versión más aplicada dirigida tanto a profesionales de la Estadística como a investigadores puede encontrarse en el libro de FOX (1984)** . En castellano podemos encontrar el tema dirigido especialmente al campo de la Economía en libros sobre Econometría, pueden consultarse, PEÑA (1994)*** .

6.2.2 Forma muestral del modelo

Normalmente supondremos que el modelo propuesto es el correcto en una población y disponemos de una muestra de n observaciones que utilizaremos para la estimación de los parámetros desconocidos. Los valores muestrales ordenados en forma de vectores y matrices son

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}$$

* SEBER, G.A.F. (1977); Linear Regression Analysis. Wiley. New York.
** FOX, J. (1984): Linear Statistical Models and Related Methods. With Applications to Social Researh. Wiley. New York.
*** PEÑA, D. (1994) Estadística: Modelos y Métodos. Vols. I y II. Alianza Universidad. Textos.

Se ha incluido una columna de unos para tener en cuenta el término independiente del modelo. El modelo para cada una de las n observaciones muestrales es:

$$\begin{aligned} y_1 &= \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \ldots + \hat{\beta}_k x_{1k} + e_1 \\ y_2 &= \hat{\beta}_0 + \hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22} + \ldots + \hat{\beta}_k x_{2k} + e_2 \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ y_n &= \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \hat{\beta}_2 x_{n2} + \ldots + \hat{\beta}_k x_{nk} + e_2 \end{aligned}$$

Escrito en forma matricial será $y_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + e_i$ para cada observación, $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$ para todas las observaciones, correspondiente al modelo poblacional $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, con

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Obsérvese que los estimadores muestrales se han denotado con $\hat{\beta}_i$, los errores aleatorios desconocidos con ε_i y los errores estimados una vez que se han estimado los parámetros (residuales) con e_i.

6.2.3 Hipótesis del modelo

Previamente a la construcción del modelo es necesario tener en cuenta algunas hipótesis que serán necesarias a la hora de determinar las propiedades de los estimadores de los parámetros de modelo. Las hipótesis siguientes convierten a la clásica regresión múltiple en un modelo estadístico más formal.

1.- La relación es estocástica (aleatoria):el término de error ε_i recoge la componente aleatoria de y_i que el modelo no puede explicar. ε_i es no observable.

2.- Ausencia de error de especificación: Aparecen en el modelo todas las variables relevantes para explicar el comportamiento de Y. Esta hipótesis está directamente relacionada con la investigación inicial ya que es el investigador aplicado quien conoce, a partir de la teoría, que variables pueden ser relevantes para explicar la respuesta. Esta hipótesis es necesaria para que el termino de perturbación aleatoria sea

error puro con media nula.

3.- Linealidad de la relación: $E(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}$. Las medias de la distribución de **Y** condicionadas a cada valor de **X** se encuentran sobre una línea (en el caso simple).

4.- Esperanza matemática nula del término de perturbación: La especificación correcta del modelo hace que no se introduzca ninguna componente sistemática en los errores al compensarse, en promedio, los positivos y negativos. Esta hipótesis es consecuencia directa de la anterior.

5.- Homocedasticidad: Varianza constante de los errores: $Var(\varepsilon_i) = \sigma^2$, para todo i.

6.- No autocorrelación: Ausencia de covarianza (o correlación) entre los errores: $Cov(\varepsilon_i , \varepsilon_j) = 0 \quad si \quad i \neq j$.

7.- Variables explicativas deterministas o no aleatorias:. Variables controladas por el investigador y medidas sin error. De esta forma el modelo lineal general está especialmente indicado en el análisis de experimentos diseñados en los que se controlan las condiciones experimentales. Esta hipótesis se puede relajar suponiendo que las variables regresoras son independientes del error aunque no sean constantes. En la mayor parte de las aplicaciones del modelo lineal las variables regresoras son aleatorias.

8.- No multicolinealidad: Es decir la variables explicativas no son linealmente dependientes. (ninguna de ellas puede obtenerse como combinación lineal de las demás). El problema será estudiado posteriormente con más detalle.

9.- Constancia de los parámetros: Debemos admitir una única estructura válida para el periodo de observación y el horizonte de predicción.

10.- Normalidad: Los errores tienen distribución normal, de media nula y desviación típica σ.

En estas condiciones iniciales, pasaremos a la estimación de los parámetros del modelo así como a la comprobación de las hipótesis básicas que permiten la validez de los resultados. Trataremos también de hacer inferencias sobre los parámetros del modelo suponiendo que disponemos de una muestra de una población más general.

6.2.4 Estimadores de los parámetros: método de los mínimos cuadrados

6.2.4.1 Interpretación de la ecuación de regresión

La figura 6.8 muestra la situación esquematizada cuando se dispone de dos variables explicativas. Se dispone de una nube de puntos en tres dimensiones y buscamos el plano que pasa lo más cerca posible de todos los puntos de la nube.

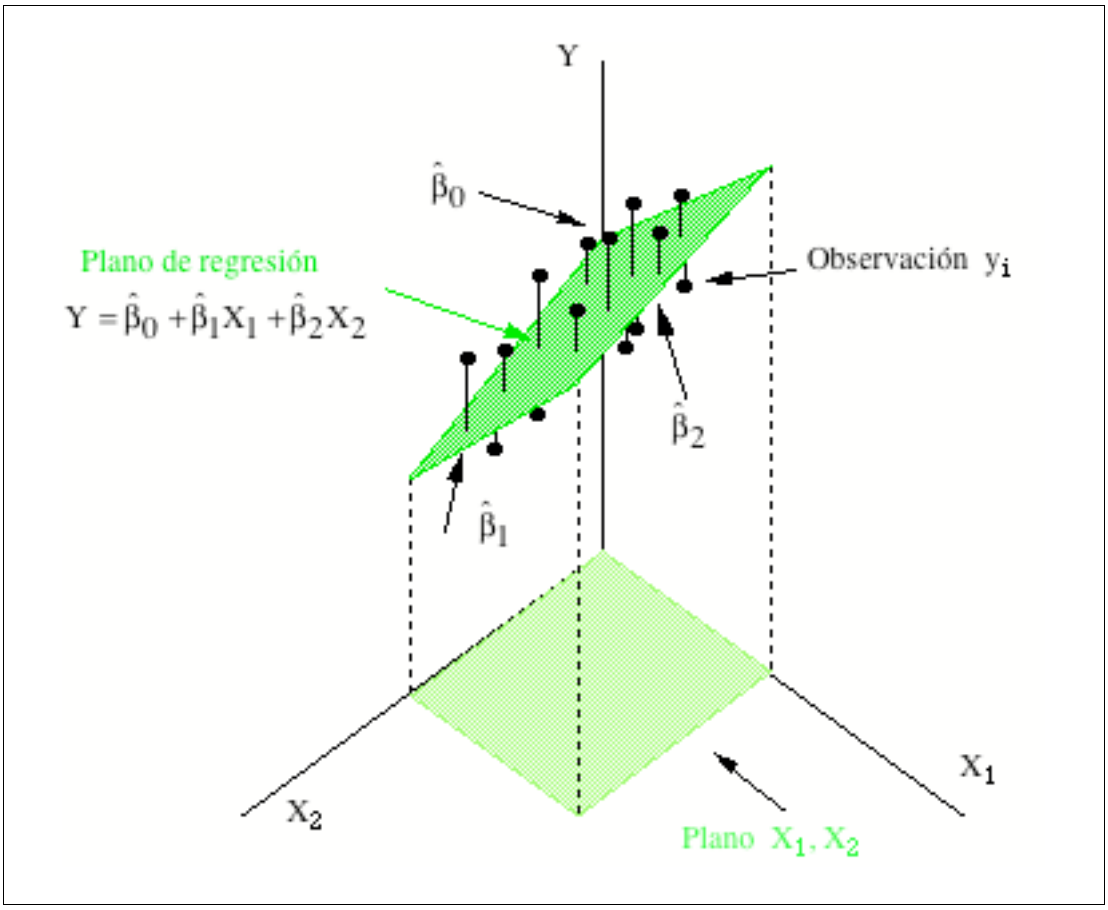


Figura 6.8: Nube de puntos e hiperplano de regresión estimado en tres dimensiones

La ecuación del plano que buscamos es de la forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ que para una muestra concreta será $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$. Los parámetros a los que tenemos que dar valor son β_0, β_1 y β_2. La interpretación es simple β_0 es lo que vale la variable dependiente cuando todas las independientes son cero y β_i es lo que aumenta la variable dependiente cuando la variable X_i aumenta en una unidad, manteniendo el resto constantes, es por esto por lo que se les denomina coeficientes de regresión parcial.

6.2.4.2 Descomposición de los valores observados en sus dos componentes.

La figura 6.9 presenta la situación esquematizada para uno de los puntos de la nube.

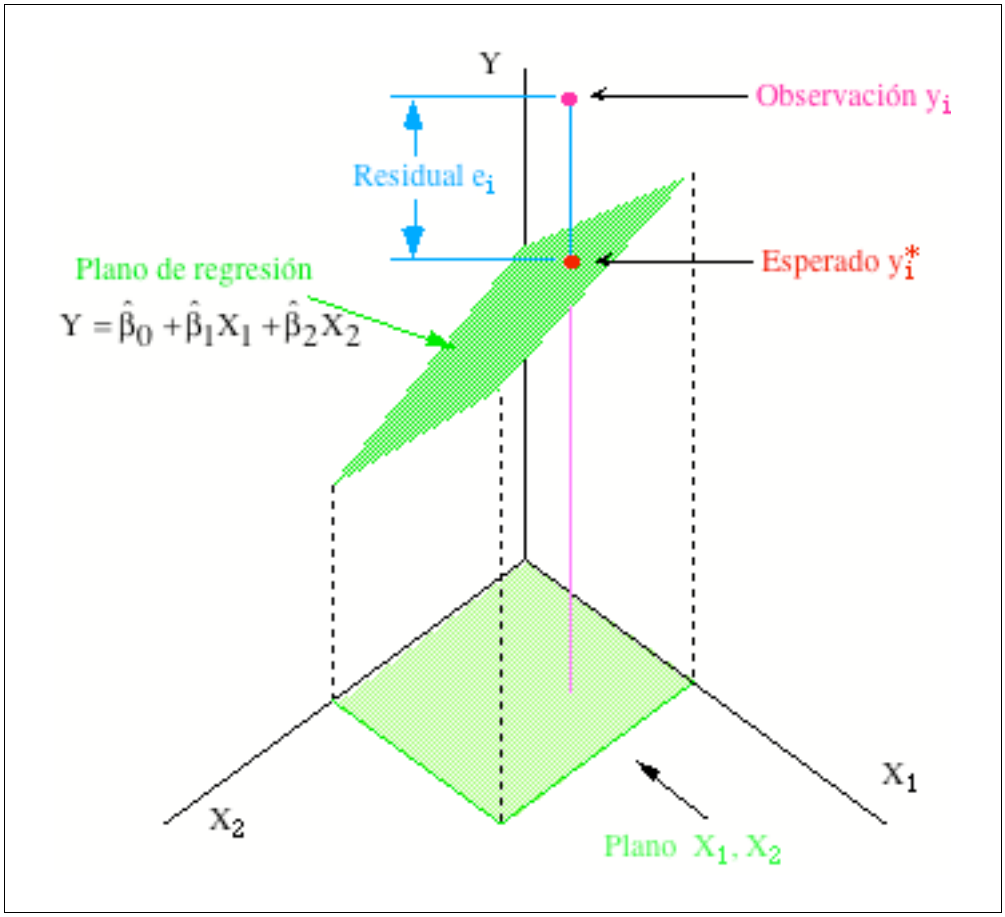


Figura 6.9: Descomposición de los valores observados en parte explicada y residual.

Llamando

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

en el modelo para los valores muestrales obtenemos

$$y_i = y_i^* + e_i$$

siendo

$$e_i = y_i - y_i^* = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$$

los residuales del modelo. Hemos descompuesto así el valor observado en dos partes, el valor esperado (o ajustado) sobre el hiperplano de regresión y_i^* que representa la parte controlada por el modelo y el residual e_i que representa la parte no controlada.

En forma matricial $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

6.2.4.3 Criterio de los mínimos cuadrados.

Se trata de buscar el hiperplano de regresión que pase lo más cerca posible de todos los puntos de la nube con algún criterio predefinido. El criterio utilizado será el de los mínimos cuadrados que consiste en minimizar la suma de cuadrados de los residuales

$$SCR = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\,\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\,\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\,\mathbf{X}'\,\mathbf{y} + \hat{\boldsymbol{\beta}}'\,\mathbf{X}'\,\mathbf{X}\hat{\boldsymbol{\beta}} =$$
$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\,\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\,(\mathbf{X}'\,\mathbf{X})\hat{\boldsymbol{\beta}}$$

Derivando con respecto a $\boldsymbol{\beta}$ e igualando a cero obtenemos

$$\frac{\partial SCR}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{0} - 2\mathbf{X}'\,\mathbf{y} + 2(\mathbf{X}'\,\mathbf{X})\hat{\boldsymbol{\beta}} = 0$$

es decir

$$(\mathbf{X}'\,\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\,\mathbf{y}$$

Resolviendo el sistema resultante obtenemos

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

La hipótesis de no multicolinealidad es necesaria para que $(\mathbf{X}'\,\mathbf{X})$ sea invertible. Obsérvese que de momento es la única de las hipótesis previas que hemos utilizado. Esto quiere decir que si lo que se pretende es simplemente ajustar un hiperplano de regresión a un conjunto de datos de forma descriptiva, puede utilizarse el criterio de los mínimos cuadrados sin ninguna suposición adicional.

6.2.5 Estimadores de los parámetros: el método de máxima verosimilitud

Se trata de buscar aquellos parámetros para los que la función de verosimilitud es máxima.

Sabemos, a partir de las hipótesis básicas, que los valores observados y_i tienen distribución normal $y_i \approx N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma)$ y que las observaciones son independientes. La función de densidad de cada valor muestral es de la forma .

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]}$$

La función de verosimilitud de los datos es la función de densidad conjunta de los valores muestrales que, como son independientes, coincide con el producto de las funciones de densidad individuales.

$$L(y_1, \ldots, y_n / \boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]} =$$

$$= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{\left[-\frac{\sum_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]} = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} e^{\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]}$$

buscamos los valores $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ que hacen mínima la verosimilitud. Como la función de verosimilitud y su logaritmo alcanzan el máximo en el mismo punto utilizamos esta última por comodidad. El logaritmo de la función de verosimilitud es

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Derivando con respecto a los parámetros $\boldsymbol{\beta}$ y σ e igualando las derivadas a cero se obtiene

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y}) = 0$$
$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

Resolviendo las ecuaciones se obtienen los valores $\hat{\boldsymbol{\beta}}$ y σ que hacen máxima la función de verosimilitud

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$\hat{\sigma}^2 = S_e^2 = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$$

Para los coeficientes de regresión se han obtenido exactamente los mismos estimadores que con el método de los mínimos cuadrados. Se ha obtenido también un estimador de la dispersión aunque se trata de un estimador sesgado.

6.2.6 Propiedades de los estimadores mínimo-cuadráticos

1.- El estimador mínimo cuadrático es un estimador lineal

El estimador es una combinación lineal de los valores observados de la respuesta

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{y} = \mathbf{M}\,\mathbf{y}$$

con $\mathbf{M} = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'$.

2.- El estimador es insesgado: La esperanza matemática del estimador coincide con el parámetro a estimar.

$$E(\hat{\boldsymbol{\beta}}) = E(\mathbf{M}\,\mathbf{y}) = \mathbf{M}\,E(\mathbf{y}) = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}$$

ya que $E(\mathbf{M}\,\mathbf{y}) = \mathbf{M}\,E(\mathbf{y})$ al ser $\mathbf{X}$, y por tanto $\mathbf{M}$, constantes controladas por el investigador. Obsérvese que aquí hemos utilizado las hipótesis $\mathbf{X}$ constante y la de linealidad.

3.- La matriz de covarianzas de los estimadores es

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] =$$
$$\mathbf{M}\, \mathrm{V}(\mathbf{y})\mathbf{M}' = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2 \mathbf{I}_n [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' =$$
$$\sigma^2 (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Hemos utilizado aquí la hipótesis de homocedasticidad.

Las varianzas de los estimadores están contenidas en la diagonal de la matriz de covarianzas.

El siguiente resultado justifica la elección de los estimados dentro de todos los estimadores lineales e insesgados. Se muestra solamente el resultado sin la correspondiente demostración que puede consultarse en los libros citados anteriormente.

Teorema de Gauss-Markov

El estimador mínimo cuadrático es entre todos los estimadores lineales insesgados el que tiene la varianza mínima (eficiente).

4.- La distribución muestral del estimador es normal

Basta tener en cuenta que una combinación lineal de variables independientes, todas con distribución normal, tiene también distribución normal.

$$\hat{\boldsymbol{\beta}} \approx \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Para cada elemento del vector (para cada estimador particular):

$$\hat{\beta}_i \approx \mathrm{N}(\beta_i, \sigma^2 a_{ii})$$

donde a_{ii} es el i-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$.

5.- Estimación de la varianza de los errores: El estimador de la varianza del error obtenido a partir del método de máxima verosimilitud era sesgado.

El estimador insesgado que utilizaremos es:

$$\hat{S}_e^2 = \frac{\mathbf{e'e}}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

que tiene distribución muestral asociada

$$\frac{(n-k-1)S_e^2}{\sigma^2} = \frac{\mathbf{e'e}}{\sigma^2} \approx \chi^2_{n-k-1}$$

Las demostraciones completas de las propiedades de los estimadores pueden revisarse en la bibliografía propuesta.

6.2.7 Contraste de significación del modelo global: análisis de la varianza en los modelos lineales

El primer paso que debemos realizar una vez ajustado el modelo es comprobar si existe realmente una relación entre las variables, lo que se traduce en que alguno de los parámetros del modelo sea distinto de cero en la población. El contraste para el ajuste global es de la forma

Es decir, comparamos el modelo reducido que tiene solamente el término independiente frente al modelo completo con todas las variables consideradas. La comparación la realizaremos comprobando si las variables regresoras consiguen explicar una parte significativa en la variabilidad de la variable dependiente. Ilustraremos el procedimiento con gráficos para el caso de una sola variable regresora.

Estudiemos primero el comportamiento del modelo reducido $Y = \beta_0$ en el que el estimador del parámetro es $\hat{\beta}_0 = \bar{y}$ la media de los valores en **y**. Luego si no tenemos ninguna información sobre las variables regresoras, la cantidad que mejor explica el comportamiento de la variable dependiente es la media de sus valores. A la suma de las desviaciones cuadráticas de cada valor con respecto a la media la denominaremos Suma de Cuadrados Total (SCT) ya que mide la dispersión máxima cuando no se tiene información sobre las regresoras.

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2$$

El valor de la suma de cuadrados total es el objetivo que trataremos de explicar al introducir la información de las variables regresoras.

Introducimos ahora las regresoras y ajustamos el modelo completo, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$. La Suma de Cuadrados de los Residuales (SCR) del modelo completo

$$SCR = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

mide la dispersión en torno al hiperplano ajustado, es decir, mide la dispersión que todavía queda después de haber introducido las variables regresoras o dispersión residual no explicada. La suma de cuadrados de los residuales mide también la dispersión intrínseca de los datos.

La figura 6.10 muestra esquemáticamente la situación descrita en los párrafos anteriores.

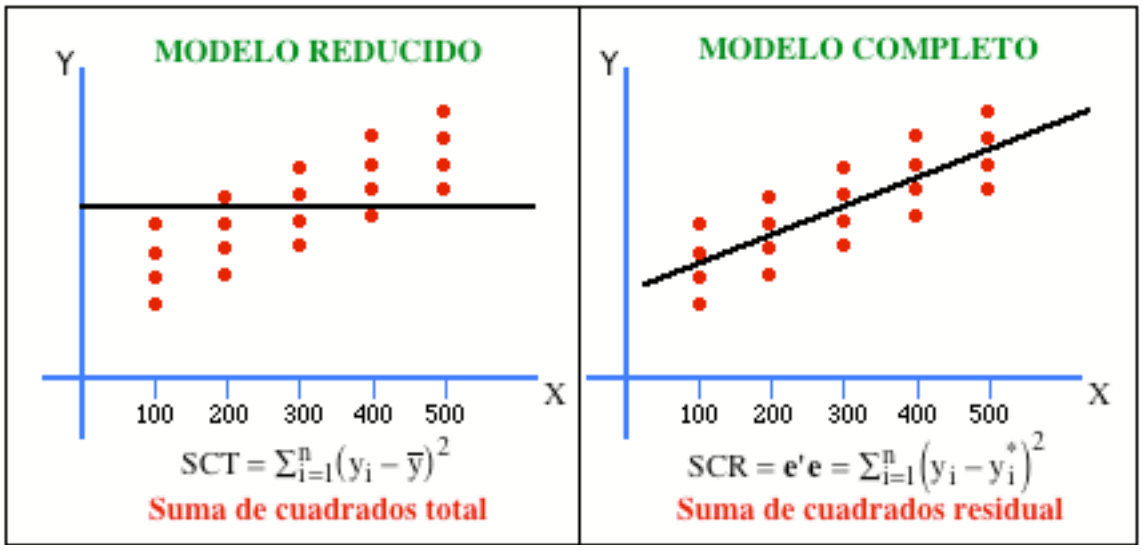


Figura 6.10: Descomposición de la suma de cuadrados en un modelo lineal.

A la vista del gráfico es claro que dispersión es mayor en torno a la media que en torno al modelo de regresión, ya que este posee mayor información. La diferencia entre ambas será la parte de la dispersión que se ha conseguido explicar mediante la introducción de las variables regresoras. Llamaremos Suma de Cuadrados Explicada (SCE) dicha diferencia (SCE = SCT - SCR). Obtenemos así la descomposición de la variabilidad total de la variable dependiente en dos partes, una parte explicada por las variables regresoras y una parte residual que todavía queda sin explicar después de haber ajustado el modelo.

$$\mathbf{y}'\mathbf{y} - n\bar{y}^2 = (\mathbf{y}^{*\prime}\mathbf{y}^{*} - n\bar{y}^2) + [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]$$

$$\mathbf{y}'\mathbf{y} - n\bar{y}^2 = (\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - n\bar{y}^2) + [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]$$

$$SCT \quad = \quad SCE \quad + \quad SCR$$

El problema es ahora saber si la dispersión explicada es lo suficientemente grande como para considerarla estadísticamente significativa. El patrón de comparación será la dispersión residual o dispersión intrínseca. Las sumas de cuadrados no son estrictamente comparables ya que están referidas a un número distinto de grados de libertad, concretamente k para la suma explicada, (n-k-1) para la residual y (n-1) para la total. Podemos construir estimadores de la variabilidad dividiendo la suma de cuadrados por los correspondientes grados de libertad, el cociente entre el estimador de la variabilidad explicada y la variabilidad residual será utilizado como medida de la importancia de la parte explicada, además dicho cociente sigue una distribución F de Snedecor con k y (n-k-1) grados de libertad en el numerados y en el denominados

respectivamente.

Obtenemos así el estadígrafo para el contraste que habíamos planteado al principio, que hemos convertido en un contraste de comparación de variabilidades.

$$F = \frac{\frac{SCE}{k}}{\frac{SCR}{(n-k-1)}} \approx F_{k,n-k-1}$$

El contraste es necesariamente unilateral superior ya que solo rechazaremos la hipótesis nula cuando la variabilidad explicada sea muy grande en comparación con la residual. Los resultados del contraste se suelen resumir en una tabla denominada tabla del Análisis de la Varianza (ANOVA), como la que se muestra en el cuadro 6.1.

Fuente de variación	**Sumas de cuadrados**	**Grados de Libertad**	**Estimadores**	**F experimental**	**Conclusión**
Explicada	SCE	k	SCE/k	(SCE/SCR)((n-k-1)/k)	n.s.= no significativo
Residual	SCR	n-k-1	SCR/(n-k-1)		* = Probablemente sign. (al 5%)
Total	SCT	n-1			** = Altamente sign. (al 1%)

Cuadro 6.1: Análisis de la varianza en un modelo de regresión.

El análisis de la varianza para el modelo de regresión forma parte de la salida estándar de cualquier programa de ordenador.

En algunos casos es posible dividir la suma de cuadrados explicada en diversas partes explicadas por una o varias variables. En general, si las variables regresoras no son independientes no es posible separar la parte explicada debida a cada una de ellas. En los experimentos diseñados es habitual tomar combinaciones de las variables explicativas con valores prefijados de forma que sean independientes para poder separar el efecto de cada una de ellas.

6.2.8 Medida de la bondad del ajuste: el coeficiente de determinación

El análisis de la varianza descrito en el caso anterior nos da un criterio para decidir si alguno de los parámetros es distinto de cero y, por tanto, si las variables regresoras explican significativamente la variabilidad de la variable independiente, sin embargo, no miden el grado de la relación existente entre la dependiente y las regresoras. Una medida descriptiva del grado de la relación existente entre las variables se denomina *Coeficiente de Determinación*, se denota con R^2 y se define como el cociente entre la suma de cuadrados explicada y la suma de cuadrados total.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Está acotado entre 0 y 1 y multiplicado por 100 representa el porcentaje de la variabilidad de la variable dependiente explicado por la introducción de las regresoras en el modelo lineal modelo lineal.

Para el modelo de regresión simple en el que se dispone de una sola variable regresora, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación de Pearson, para el modelo general, el coeficiente de determinación puede interpretarse también como el cuadrado del coeficiente de correlación entre los valores de **y** y los de **y***. A dicho coeficiente de correlación se le denomina *Coeficiente de Correlación Múltiple***.**

El coeficiente de determinación es sencillo y fácil de interpretar aunque tiene un problema importante, aumenta con el número de variables regresoras, estén o no relacionadas con la dependiente, de forma que es posible conseguir una bondad del ajuste próxima a 1 simplemente introduciendo en el modelo un número elevado de variables. Para evitar este problema se define el *Coeficiente de Determinación Ajustado*, en el que las sumas de cuadrados se dividen por sus correspondientes grados de libertad.

$$R^2_{aj.} = 1 - \frac{SCR / (n-k-1)}{SCT / n-1}$$

la interpretación es exactamente la misma que la del coeficiente de determinación.

6.2.9 Suma de cuadrados explicada por un grupo de variables: contraste para un grupo de parámetros

En algunas situaciones es importante conocer, no solo la variabilidad explicada por el conjunto total de regresoras sino también la variabilidad explicada por un subconjunto de los mismos, para contrastar si consiguen explicar significativamente parte de la variabilidad. El contraste es ahora que los coeficientes de un subgrupo de p regresoras son todos iguales a cero frente a la alternativa de que alguno es distinto de cero. Sin pérdida de generalidad podemos suponer que el subconjunto está formado por las p primeras variables y escribimos el modelo completo como

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{p+1} X_{p+1} + \ldots + \beta_k X_k$$

Las hipótesis a contrastar pueden escribirse de la siguiente manera

$$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad (Y = \beta_0 + \beta_{p+1} X_{p+1} + \ldots + \beta_k X_k)$$

$$H_a : \exists i / \beta_i \neq 0, \; i \in (1, \ldots\ldots, p) \qquad (Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{p+1} X_{p+1} + \ldots + \beta_k X_k)$$

El procedimiento será similar al del contraste global y consiste en la comparación de las sumas de cuadrados explicadas en el modelo completo y un modelo reducido en el que se eliminan las variables que se quieren contrastar. La suma de cuadrados explicada en el modelo completo (con todas las variables) la vamos a dividir en dos partes, una parte explicada por las k-p, variables no incluidas en el subconjunto a contrastar y una parte explicada por las p variables a contrastar y que no ha sido explicada por el resto.

La descomposición de la suma de cuadrados en el modelo completo es SCT=SCE+SCR, donde SCE es la variabilidad explicada por todas las variables regresoras. La descomposición al ajustar el modelo reducido la denotaremos como $SCT=SCE_0+SCR_0$, donde SCE_0 representa la parte explicada por las k-p variables que no están en el subconjunto objetivo. La diferencia entre ambas sumas de cuadrados explicados será la parte explicada por las p variables objetivo y que no ha sido ya explicada por el resto. Denotaremos esta última suma de cuadrados como $SCE_p = SCE - SCE_0$. Los grados de libertad asociados son p.

Es posible construir el contraste correspondiente teniendo en cuenta que

$$F_{Y,(1,..,p)/(p+1,..,k)} = \frac{\frac{SCE_p}{p}}{\frac{SCR}{n-k-1}}$$

sigue una distribución F de Snedecor con p y n-k-1 grados de libertad en el numerados y denominados respectivamente.

Obsérvese que se ha utilizado en el contraste la parte explicada por las p variables del subconjunto objetivo y que no ha sido ya explicada por el resto, en lugar de utilizar la suma de cuadrados explicada por las p variables sin tener en cuenta el resto. Ambas sumas de cuadrados sólo coinciden cuando las p variables y el resto son independientes.

6.2.10 El coeficiente de correlación parcial

El coeficiente de determinación (múltiple) R^2, medía la reducción proporcional en la variabilidad de Y conseguida mediante la introducción del conjunto completo de regresoras en el modelo. Es posible definir un *Coeficiente de Determinación Parcial* que mida la contribución marginal de un subconjunto de regresoras, cuando todas las demás han sido ya incluidas en el modelo. Denotaremos este coeficiente como $R^2_{Y,(1,..,p)/(p+1,..,k)}$ y lo calcularemos como

$$R^2_{Y,(1,..,p)/(p+1,..,k)} = \frac{SCE - SCE_0}{SCT - SCE_0} = \frac{SCE_p}{SCR_0} = 1 - \frac{SCR}{SCR_0}$$

Representa la parte que se ha conseguido explicar de la suma de cuadrados residual del modelo reducido al introducir el subconjunto de p variables en el modelo.

La raíz cuadrada del coeficiente de determinación parcial se denomina ***Coeficiente de Correlación Parcial***

$$r_{Y,(1,..,p)/(p+1,..,k)} = \sqrt{R^2_{Y,(1,..,p)/(p+1,..,k)}}$$

el signo de la raíz cuadrada ha de ser el mismo que el signo del coeficiente de regresión estimado. Puede interpretarse como una medida de la relación entre la variable dependiente y un subconjunto de las regresoras dadas. todas las demás. La interpretación es similar a la del coeficiente de correlación de Pearson cuando el subconjunto objetivo está formado por una única variable, aunque sólo coincide con éste cuando la variable objetivo y el resto son independientes.

6.2.11 Contrastes e intervalos de confianza para cada uno de los parámetros por separado

Hasta el momento hemos visto como realizar contrastes para el modelo completo o para un subconjunto de parámetros. Cuando el subconjunto está formado por un único parámetro existe una forma alternativa de realizar el contraste individual basándose en la combinación de la distribución normal de los estimadores de los parámetros del modelo y en la distribución ji-cuadrado asociada a la varianza de los residuales, para construir una distribución t de Student.

Las hipótesis del contraste individual son

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

y las correspondientes versiones unilaterales.

La cantidad

$$t_i = \frac{\hat{\beta}_i - \beta_i}{\hat{S}_e \sqrt{a_{ii}}}$$

donde a_{ii} es el i-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$, sigue una distribución t de Student con n-k-1 grados de libertad. La construcción del contraste es inmediata.

Es necesario hacer notar que si el número de parámetros es elevado y cada uno se realiza al nivel α, el contraste global de igualdad a cero de todos los parámetros a partir

de los contrastes individuales, tiene un considerable incremento en el riesgo tipo I. Es por esto por lo que puede ocurrir que el análisis de la varianza global resulte ser no significativo y alguno de los parámetros individuales sea significativamente distinto de cero.

El contraste, basado en la F, para un subgrupo formado por una sola variable es completamente equivalente al descrito aquí ya que se verifica que

$$t_i^2 = F_{Y,(i)/(1,\ldots,i-1,i+1,\ldots,k)}$$

Este valor es el que aparece en muchos programas de ordenador como F parcial.

Los intervalos de confianza para los parámetros por separado calculados a partir de la distribución t de Student son de la forma

$$I_{\beta_i}^{1-\alpha} = \left[\hat{\beta}_i \pm t_{n-k-1;\alpha}\hat{S}_e\sqrt{a_{ii}}\right]$$

6.2.12 Ajuste de un modelo: los programas de ordenador

El ajuste de un modelo lineal a datos reales requiere un elevado número de cálculos de forma que es necesario disponer de un ordenador para ajustar de forma eficiente distintos modelos a los datos experimentales. Los métodos de regresión múltiple se encuentran prácticamente en todos los paquetes estadísticos disponibles, desde los más simples hasta los más avanzados. La primera consideración que hemos de tener en cuenta es que el ordenador es solamente una herramienta de cálculo rápido que, en ningún momento sustituye el conocimiento del profesional sobre los métodos que está utilizando. La aplicación sistemática de métodos estadísticos sin un análisis previo de su aplicabilidad es un error común entre los investigadores aplicados. Como veremos en apartados posteriores el proceso de análisis de un modelo concreto no termina con el ajuste rápido del mismo sino que implica la comprobación de varias hipótesis de forma interactiva entre el conocimiento teórico y la herramienta de cálculo.

Las figuras 6.11, 6.12 y 6.13 muestran los elementos básicos de los resultados del ajuste de un modelo lineal a unos datos reales.

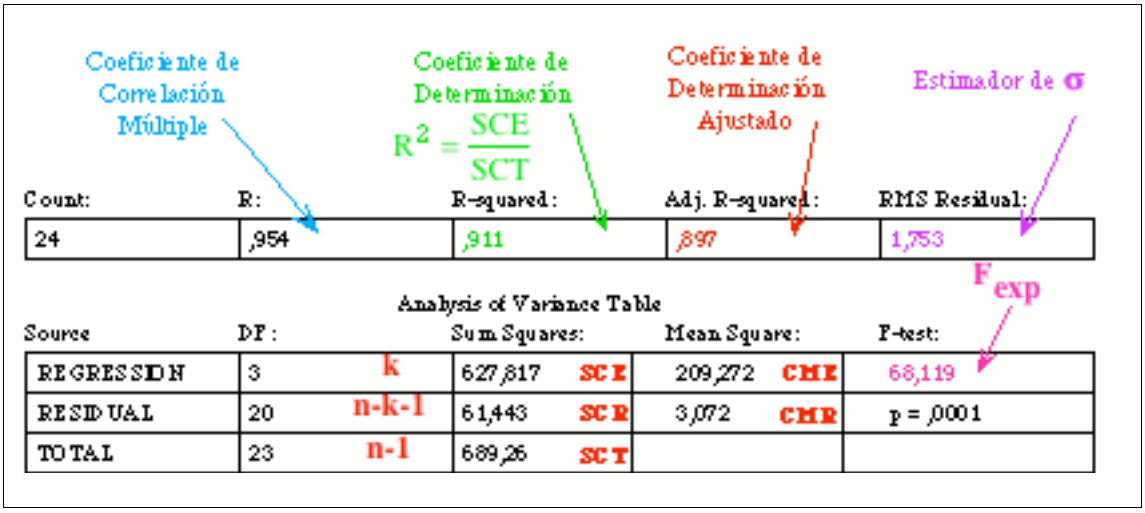


Figura 6.11: Bondad del ajuste y Análisis de la varianza para el modelo lineal.

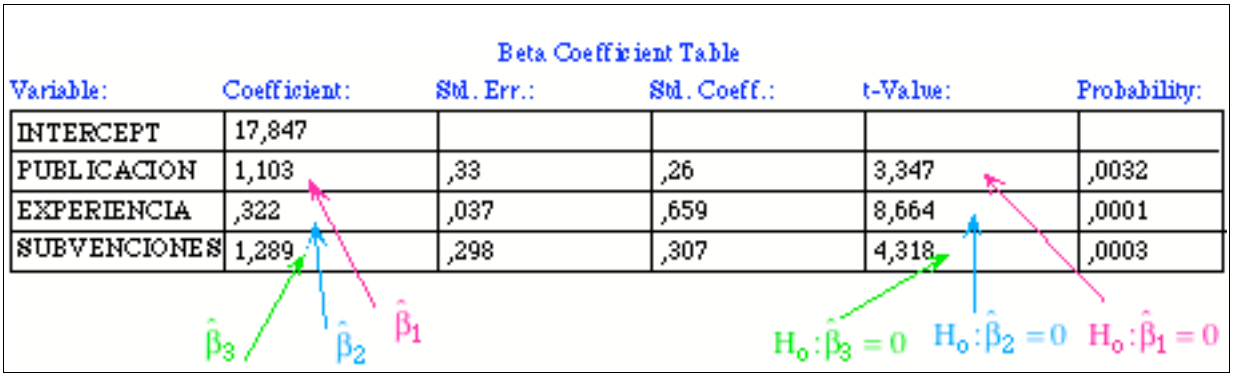


Figura 6.12: Estimadores de los parámetros y contrastes individuales para el modelo lineal

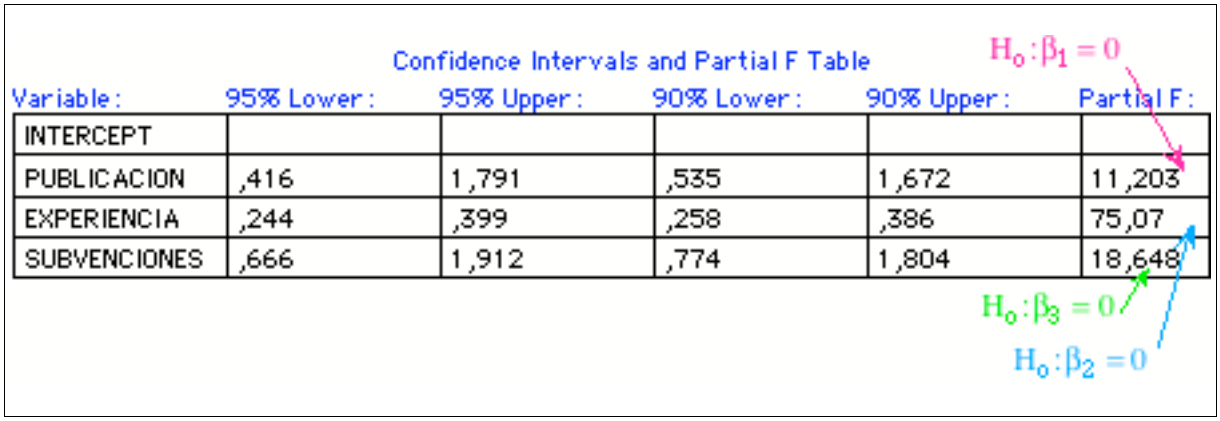


Figura 6.13: Intervalos de confianza y F parciales para el modelo lineal.

6.2.13 Selección de subconjuntos de variables (métodos paso a paso)

Los contrastes para un grupo de parámetros, basados en los incrementos de la suma de cuadrados conseguidos al introducir una o varias variables en un modelo, nos dan criterios de selección de subconjuntos de variables conocidos como **métodos paso a paso**, utilizados en la mayor parte de los paquetes estadísticos. Podemos distinguir tres tipos fundamentales:

6.2.13.1 Selección ascendente (forward selection)

-Comenzamos con un modelo sin ninguna variable.

-Introducimos aquella variable que produce un mayor incremento significativo en la suma de cuadrados explicada (El coeficiente de correlación parcial más alto).

-Repetimos el proceso de selección hasta que ninguna de las variables fuera del modelo produzca un incremento significativo en la suma de cuadrados.

6.2.13.2 Eliminación descendente (backward elimination)

-Comenzamos con el modelo completo.

-Eliminamos aquella variable que al ser sacada fuera del modelo produce la menor pérdida no significativa.

-El proceso termina cuando todas las variables dentro del modelo producen una pérdida (incremento) significativa.

6.2.13.3 Regresión paso a paso

Es básicamente un proceso de selección ascendente en el que en cada paso se permite la posibilidad de que las variables que ya están dentro del modelo puedan ser eliminadas.

VENTAJAS:

-Producen un subconjunto reducido de variables más fácil de manejar.

INCONVENIENTES:

-El subconjunto final obtenido no es óptimo, en general.

-Si las variables están relacionadas entre si (existe multicolinealidad) los procesos son muy inestables ya que no es posible separar el efecto debido a cada una de ellas.

-El orden de entrada es irrelevante.

6.2.14 Predicción en el modelo lineal general

Supongamos que disponemos de un vector de observaciones para las k variables regresoras $\mathbf{x_{(0)}} = (x_{01}, \dots, x_{0k})$ y deseamos la predicción del valor medio que tomaría la variable Y o bien de un valor concreto para Y. Distinguimos entre la predicción de un valor medio y la predicción de una observación individual porque la variabilidad es diferente. En ambos casos la predicción es la misma

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_k x_{0k}$$

Es posible calcular intervalos de confianza para la predicción en los dos casos mencionados:

Media de Y: $I_{\overline{y}_0} = \left[\hat{y}_0 \pm t_{n-k-1,\alpha}\, S_e \sqrt{\mathbf{x}'_{(0)}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_{(0)}} \right]$

Valor concreto de Y: $I_{\overline{y}_0} = \left[\hat{y}_0 \pm t_{n-k-1,\alpha}\, S_e \sqrt{1 + \mathbf{x}'_{(0)}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_{(0)}} \right]$

Han de tenerse las siguientes precauciones para la validez de la predicciones:

1.-Se supone que la estructura paramétrica no ha variado en el momento de la predicción.

2.-Las predicciones han de realizarse para valores dentro del intervalo en el que las regresoras han sido medidas, es decir no deben extrapolarse los resultados.

3.-Los intervalos de confianza para las predicciones son menos precisos a medida que nos alejamos de los valores medios de las regresoras.

4.-El hecho de que un modelo presente un alto porcentaje de variaciones controladas no implica que sea siempre un buen modelo predictivo.

Distinguiremos así entre lo que denominaremos poder explicativo, medido a través del coeficiente de determinación, y poder predictivo o capacidad de predicción. Alcanzaremos un poder predictivo aceptable cuando además de tener una explicación correcta el modelo verifique las hipótesis básicas y no se detecte la presencia de observaciones extrañas, grupos con estructuras diferentes, etc. Este punto se tratará más ampliamente en los apartados siguientes.

6.2.15 Introducción de variables cualitativas en un modelo de regresión múltiple

Todo lo que hemos visto hasta el momento se refiere a la utilización de variables de tipo continuo como regresoras, sin embargo, en la práctica es muy común encontrar situaciones en las que alguna de las regresoras es de tipo cualitativo o incluso ordinal. El problema para variables ordinales suele resolverse dando puntuaciones a los distintos valores de la variable que reproduzcan el orden de los mismos, tratándolas así como si fueran variables continuas. Las variables de tipo cualitativo son más comunes en la práctica ya que en la mayor parte de los experimentos diseñados las regresoras son

niveles de un factor cualitativo. La introducción de este tipo de variables la haremos a través de lo que denominaremos variables ficticias que describimos a continuación.

6.2.15.1 Variables ficticias (dummy)

Ilustraremos la introducción de variables ficticias con un ejemplo adaptado del libro de FOX citado anteriormente. El ejemplo se refiere a la relación entre el nivel de ingresos y el nivel de educación en dos grupos raciales en Estados Unidos. Suponemos que el nivel de ingresos (medido a través del salario) es la variable dependiente y que el nivel de educación (medido a través del número de años) es la variable independiente o regresora.

Cabe esperar que, en general, para un nivel de educación más alto el nivel de ingresos sea también más alto. Dadas las características de la sociedad americana, es de esperar también que para un mismo nivel de educación una persona de raza blanca tenga un nivel de ingresos mayor que una persona de raza negra. Luego el nivel de ingresos depende de la raza (variable cualitativa) y debería ser incluida en el modelo como regresora. La situación se ha esquematizado en la figura 6.14.

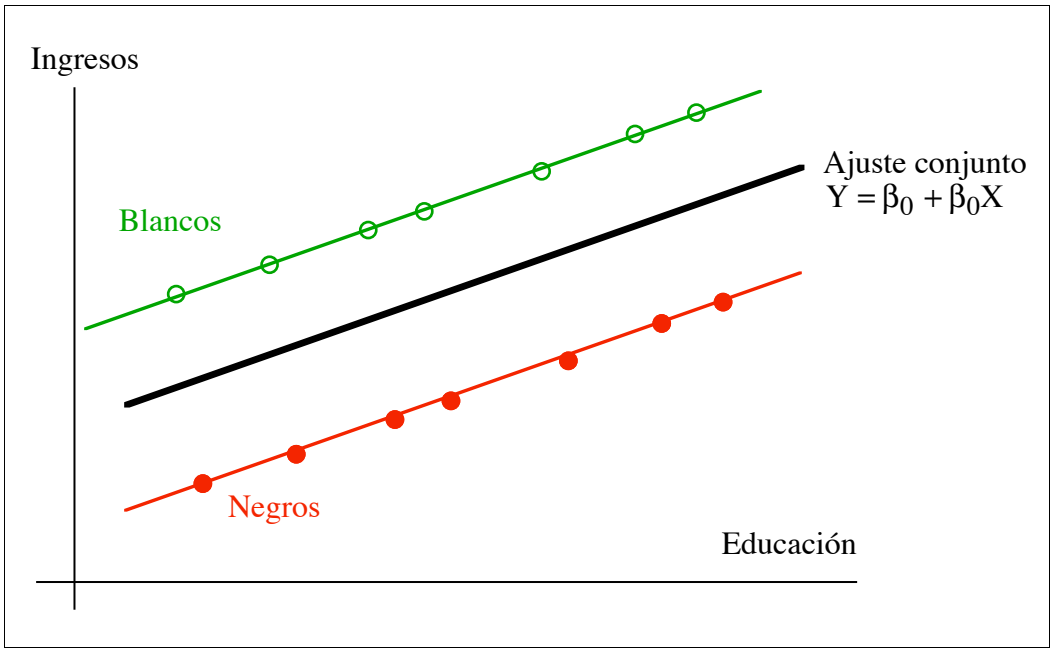


Figura 6.14: Relación entre el nivel de ingresos y el de educación para blancos y negros.

Es claro que si no incluimos la variable raza de alguna manera, el modelo conjunto que relaciona ingresos y educación no se ajusta bien ni al grupo de los blancos

ni al de los negros. Una solución sería ajustar un modelo por separado para cada grupo y compararlos, sin embrago, vamos a buscar una solución que explique correctamente la situación con un solo modelo.

En lugar de ajustar el modelo conjunto

$$Y = \beta_0 + \beta_1 X$$

podemos ajustar el modelo

$$Y = \beta_0 + \beta_1 X + \delta D$$

donde la variable D se define de la siguiente manera

$$D = \begin{cases} 1 & \text{si el individuo es blanco} \\ 0 & \text{si el individuo es negro} \end{cases}$$

la variable D se dice que es una ***variable ficticia*** ya que no ha sido media directamente de esta forma. Veamos como la variable ficticia soluciona el problema.

6.2.15.2 Interpretación del modelo con variables ficticias

La interpretación de los modelos en los que se han incluido variables ficticias es simple. Calculemos el modelo en cada uno de los grupos.

En el grupo de los negros (**D** = 0)

$$Y = \beta_0 + \beta_1 X + \delta\, 0 = \beta_0 + \beta_1 X$$

En el grupo de los blancos (**D** = 1)

$$Y = \beta_0 + \beta_1 X + \delta\, 1 = (\beta_0 + \delta) + \beta_1 X$$

Luego β_1 es la pendiente (común) de los modelos para ambos grupos. β_0 es la constante en el modelo para el grupo de los negros, $\beta_0 + \delta$ es la constante en el modelo para el grupo de los blancos y δ, por tanto, es la diferencia entre los ingresos de los blancos y los negros, sea cual sea el nivel de educación. El contraste de igualdad a cero

de δ es el contraste de que no hay diferencias en el nivel de ingresos entre los dos grupos de la raza, sea cual sea el nivel de educación. La situación esquematizada se muestra en la figura 6.15.

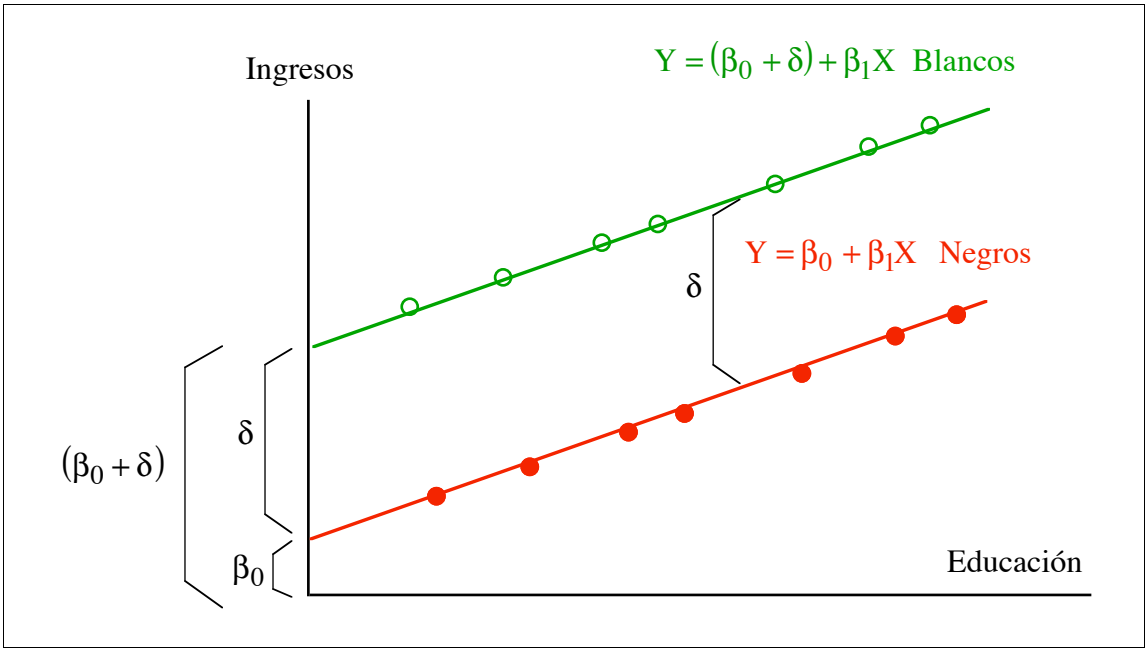


Figura 6.15: Interpretación de un modelo con variables ficticias.

Cuando la variable cualitativa tiene más de dos grupos tenemos que introducir varias variables ficticias.

6.2.15.3 Variables ficticias para mas de dos grupos

Supongamos ahora que disponemos de un grupo más, el de los hispanos y hemos de introducir los tres en el modelo que relaciona ingresos y educación. Tomaremos uno de los grupos como base de comparación, por ejemplo, el grupo de los hispanos.

El modelo será ahora

$$Y = \beta_0 + \beta_1 X + \delta_n D_n + \delta_b D_b$$

donde las variables $\mathbf{D_n}$ y $\mathbf{D_b}$ se define de la siguiente manera

$$D_n = \begin{cases} 1 & \text{si el individuo es negro} \\ 0 & \text{si el individuo no es negro} \end{cases} \qquad D_b = \begin{cases} 1 & \text{si el individuo es blanco} \\ 0 & \text{si el individuo no es blanco} \end{cases}$$

La interpretación de los parámetros y el modelo para los distintos grupos es clara a partir del gráfico de la figura 6.16.

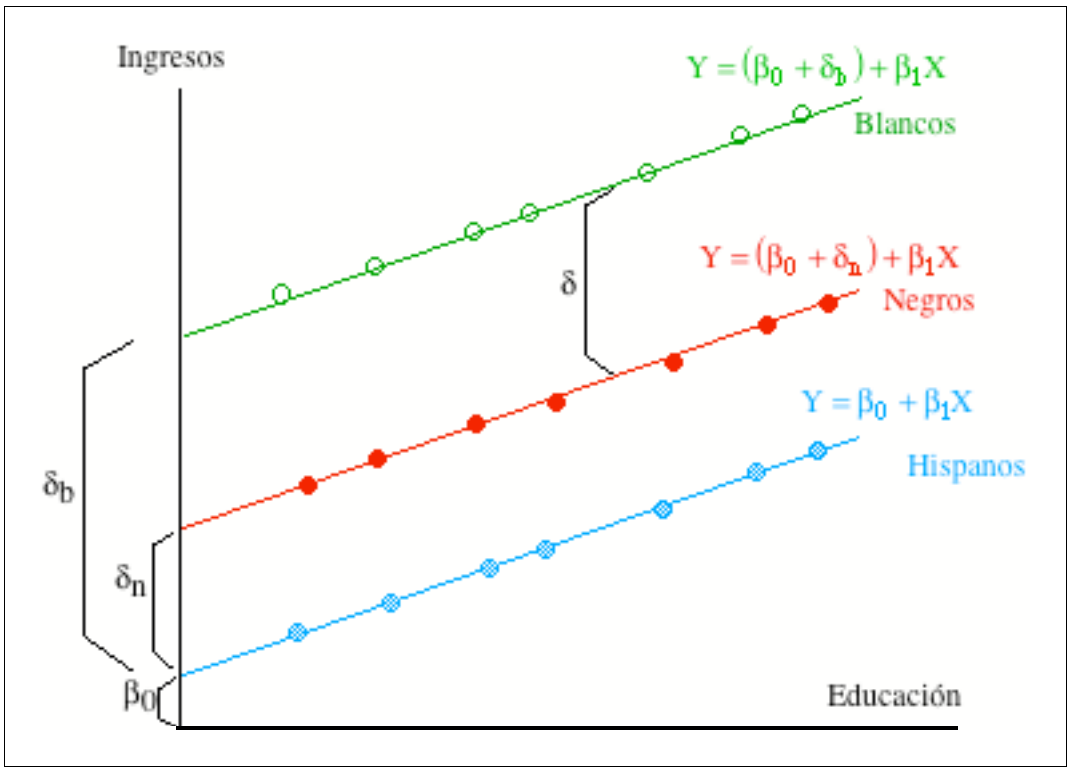


Figura 6.16: Variables ficticias para varios grupos.

Luego β_1 es la pendiente (común) de los modelos para los tres grupos. β_0 es la constante en el modelo para el grupo de los hispanos, $\beta_0 + \delta_n$ es la constante en el modelo para el grupo de los negros y $\beta_0 + \delta_b$ es la constante en el modelo para el grupo de los blancos; entonces δ_n es la diferencia entre el grupo de los negros y el de los hispanos, δ_b es la diferencia entre el grupo de los blancos y el de los hispanos y $\delta_b - \delta_n$ es la diferencia entre blancos y negros.

6.2.15.4 Variables ficticias en presencia de interacción

Supongamos ahora que, en el ejemplo anterior, las diferencias entre los ingresos para las dos razas, aumentan a medida que aumenta el nivel de educación, es decir, los efectos de la raza y del nivel de educación no son aditivos, existe lo que se denomina **interacción** entre la raza y el nivel de educación. El concepto de interacción es clave en la investigación aplicada, ya que implica que las relación de la variable dependiente con otra variable depende de los valores de una tercera. No debe confundirse

interacción con relación, en el ejemplo, raza y educación interactúan en el efecto que manifiestan sobre el nivel de educación, pero no tienen porqué estar relacionadas entre si.

La interacción se traduce en que las pendientes de las rectas para ambos grupos no son las misma. La situación se representa en la figura 6.17.

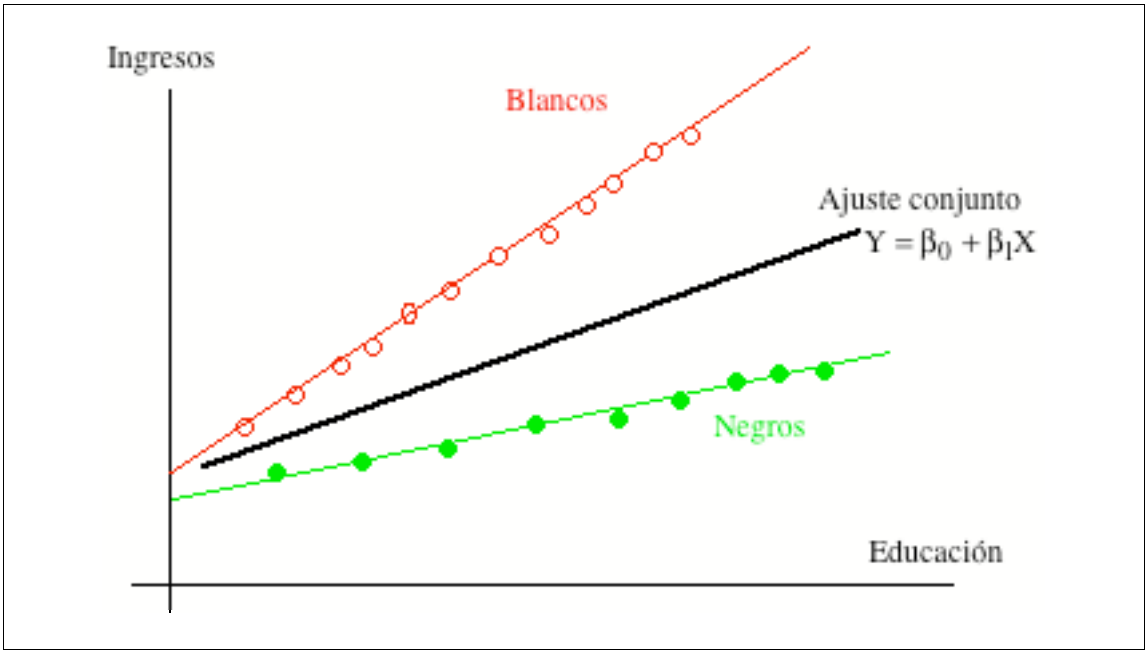


Figura 6.17: Interacción entre raza y educación.

En este caso no es válido el modelo anterior con variables ficticias, ya que, allí suponíamos que las pendientes de las rectas eran iguales y, por tanto, la diferencia entre blancos y negros era constante.

Tomaremos ahora el modelo

$$Y = \beta_0 + \beta_1 X + \delta D + \gamma DX$$

donde la variable D se define como antes, y DX es el producto de las variables D y X, es decir

$$D = \begin{cases} 1 & \text{si el individuo es blanco} \\ 0 & \text{si el individuo es negro} \end{cases}$$

y

$$DX = \begin{cases} X & \text{si el individuo es blanco} \\ 0 & \text{si el individuo es negro} \end{cases}$$

La interpretación del nuevo modelo es simple. Calculamos el modelo en cada uno de los grupos.

En el grupo de los negros (**D** = 0)

$$Y = \beta_0 + \beta_1 X + \delta\, 0 + \gamma\, 0 = \beta_0 + \beta_1 X$$

En el grupo de los blancos (**D** = 1)

$$Y = \beta_0 + \beta_1 X + \delta\, 1 + \gamma X = (\beta_0 + \delta) + (\beta_1 + \gamma) X$$

Luego β_1 es la pendiente del modelo para el grupo de los negros. $(\beta_1 + \gamma)$es la pendiente del modelo para el grupo de los blancos y, por tanto, γes la diferencia en las pendientes.

β_0 es la constante en el modelo para el grupo de los negros, $\beta_0 + \delta$ es la constante en el modelo para el grupo de los blancos. δ ya no es la diferencia entre los ingresos de los blancos y los negros, ya que esta depende del nivel de educación (ver figura 6.18).

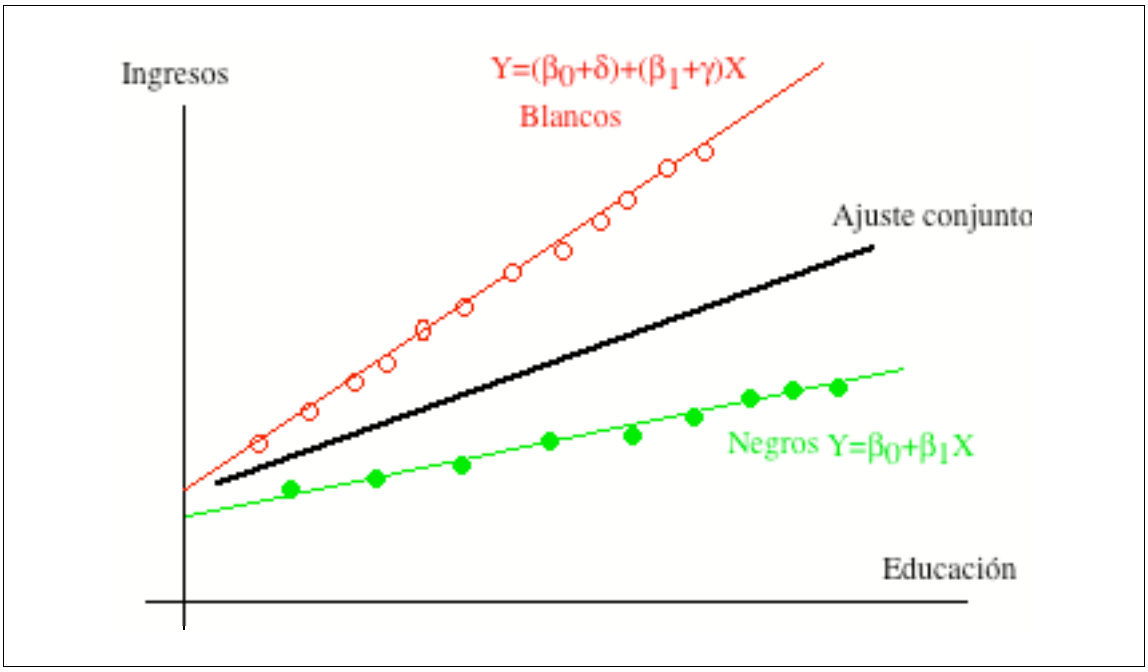


Figura 6.18: Interpretación del modelo de variables ficticias considerando interacción.

Contrastar la presencia de interacción en el modelo consiste en contrastar la nulidad del parámetro γ. Si se dispone de varios grupos es necesario introducir en el

modelo el producto de la variable continua por todas las variables ficticias. Si se dispone de dos variables cualitativas y se desea introducir la interacción de las mismas en el modelo hay que multiplicar todos los pares posibles de variables ficticias resultantes. Si se desea introducir la interacción entre dos variables continuas basta con introducir el producto de las mismas.

6.2.16 Validación de las hipótesis básicas del modelo

Hasta el momento hemos realizado todos los cálculos suponiendo que las hipótesis básicas formuladas al principio se verificaban, sin embargo, aun no hemos desarrollado ningún test de verificación. Esto es así debido a que la mayor parte de las mismas se refieren a las perturbaciones aleatorias que son variables aleatorias no observables. Las perturbaciones pueden ser estimadas mediante los residuales y, es necesario realizar el ajuste previamente.

Se han desarrollado muchos procedimientos y tests formales para detectar la posible violación de las hipótesis básicas, sin embargo nos limitaremos a realizar una aproximación básicamente descriptiva basada en los residuales, es decir, en las diferencias entre valores observados y ajustados con el modelo.

La filosofía de los apartados que siguen es que los modelos lineales que utilizaremos en la práctica necesitan de una inspección detallada una vez que han sido ajustados ya que no es suficiente con la el cálculo del poder explicativo del modelo. En el caso de la regresión simple, suele ser suficiente con el examen del diagrama de dispersión aunque es conveniente también gráficos de residuales para detectar posibles problemas que pasan inadvertidos en el diagrama de dispersión. Incluso en el caso en el que se use la regresión en forma descriptiva los residuales pueden ayudar a detectar problemas como la no linealidad o la presencia de observaciones extrañas, o la presencia de grupos diferenciados que no están necesariamente relacionados con el modelo estadístico formal.

Para ilustrar la necesidad de un análisis detallado de cada caso particular más allá de la simple bondad del ajuste, utilizaremos cuatro conjuntos de datos artificiales tomados de ANSCOMBE (1973)* . Los datos aparecen en la tabla 6.1. La primera columna de la tabla contiene los valores de X para los tres primeros conjuntos de datos.

* ASCOMBE, F.J. (1973) Graphs in Statistical Analysis. *Am. Statist.* **27**, 17-21.

X	Y1	Y2	Y3	X4	Y4
10	8,04	9,14	7,46	8	6,58
8	6,95	8,14	6,77	8	5,76
13	7,58	8,74	12,74	8	7,71
9	8,81	8,77	7,11	8	8,84
11	8,33	9,26	7,81	8	8,47
14	9,96	8,10	8,84	8	7,04
6	7,24	6,13	6,08	8	5,25
4	4,26	3,10	5,39	19	12,50
12	10,84	9,13	8,15	8	5,56
7	4,82	7,26	6,42	8	7,91
5	5,68	4,74	5,73	8	6,89

Tabla 6.1: Datos de Anscombe.

Los cuatro conjuntos de datos presentan los mismos estimadores de los parámetros y la misma bondad del ajuste. A priori parece que el modelo lineal se ajusta igualmente bien en todos los casos, sin embargo, los ajustes son muy diferentes y solamente uno de ellos está en buenas condiciones. La figura 6.19 muestra los diagramas de dispersión.

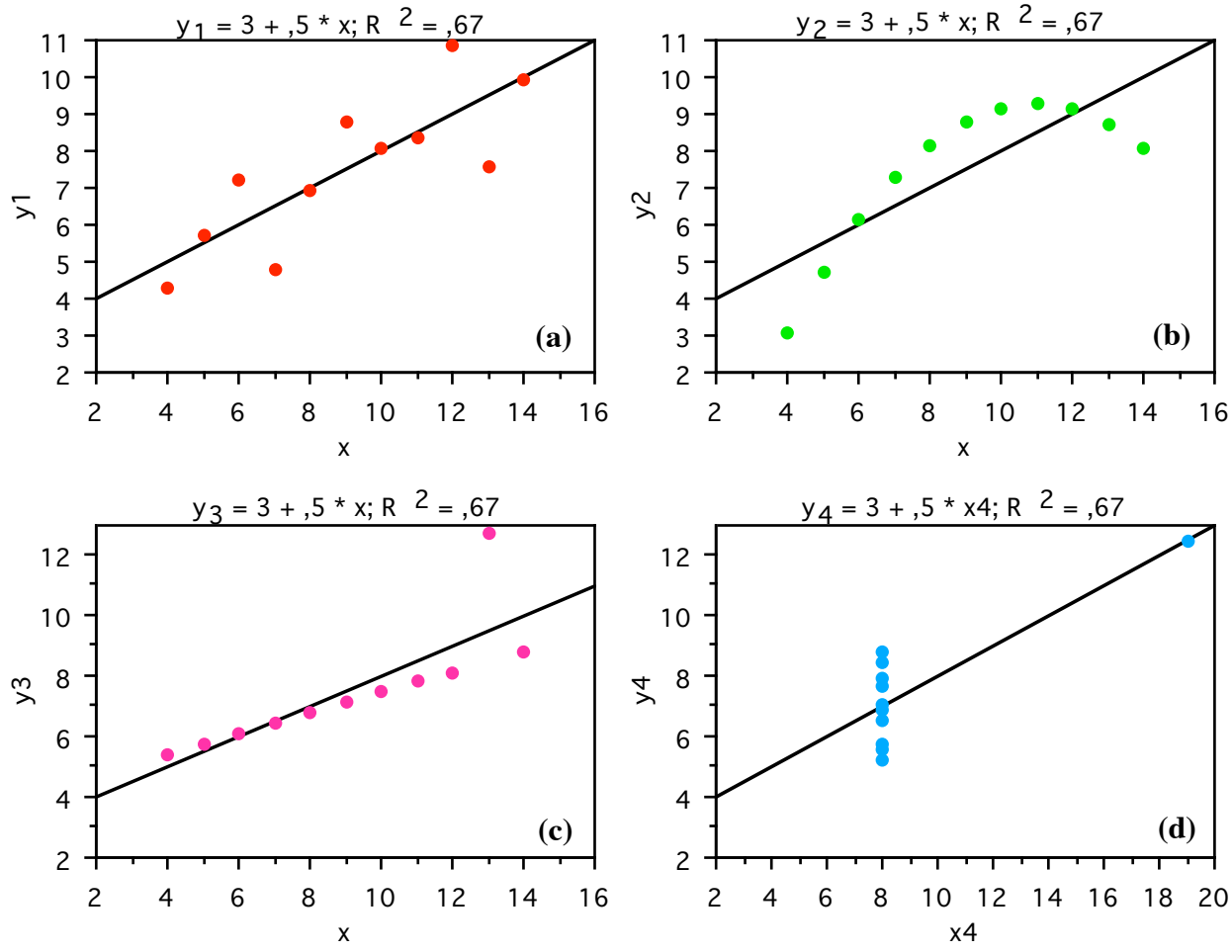


Figura 6.19: Gráficos de Anscombe.

El poder explicativo de todos los conjuntos de datos es el mismo, sin embrago, el único en el que el ajuste es razonable es en el caso (a) en el que los datos varían de forma aleatoria alrededor de la recta de regresión. En el caso (b) se observa claramente como debería ajustarse una parábola a los datos. En el caso (c) existe una relación casi perfecta entre las dos variables que está modificada por el punto aislado que, probablemente, es un outlier. En el caso (d) la relación está completamente determinada por el punto aislado, si lo suprimimos, las variables serían independientes.

Hemos descrito aquí problemas en regresión simple que pueden verse directamente sobre el diagrama de dispersión, en el caso múltiple la búsqueda es más compleja al no poder representar directamente los gráficos. Realizaremos los diagnósticos de forma indirecta utilizando gráficos de residuales en diversas versiones.

6.2.16.1 Los gráficos de residuales como herramienta de diagnóstico

Una de las herramientas fundamentales para el diagnóstico de posibles problemas en los modelos lineales son los denominados gráficos de residuales. Un gráfico de residuales no es más que un diagrama de dispersión de los residuales e_i obtenidos al ajustar el modelo, frente a los valores de la variable dependiente , a los valores esperados con el modelo o a los valores de las regresoras.

En un modelo con un poder explicativo aceptable, los residuales deben distribuirse homogéneamente alrededor del hiperplano de regresión, no deben variar de forma sistemática y la varianza ha de ser constante. La representación de los residuales frente a los valores esperados con el modelo es el gráfico más utilizado aunque son posibles muchas otras posibilidades. El aspecto del gráfico de residuales en un modelo en buenas condiciones debe presentar una nube de puntos homogénea como la que se muestra en el gráfico de la figura 6.20.

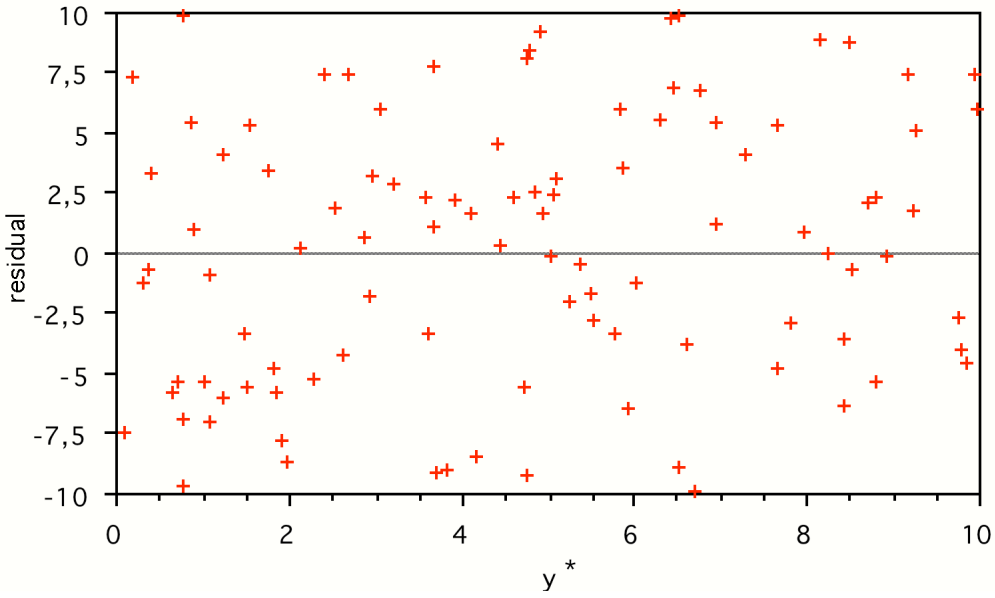


Figura 6.20: Gráfico de residuales para un modelo en buenas condiciones.

Los residuales pueden servir para detectar diversos problemas como posibles datos aberrantes (outliers), desviaciones de la linealidad, heteroscedasticidad, autocorrelación entre las observaciones, etc.

6.2.16.2 Linealidad de la relación

Es evidente que, en la práctica se dan muchas situaciones en las que los modelos que mejor se ajustan a los datos no son lineales. Se han descrito exhaustivamente relaciones de tipo lineal porque su tratamiento es muy sencillo y porque muchas de las relaciones no lineales pueden convertirse en lineales mediante transformaciones simples de las variables.

Supongamos por ejemplo, que la relación que liga a dos variables es de tipo potencial, de forma que el incremento de la variable dependiente se realiza en progresión geométrica.

$$Y = \alpha X^{\beta}$$

El ajuste de esta ecuación por mínimos cuadrados conduciría a un sistema de ecuaciones no lineales que ha de resolverse, generalmente, mediante métodos numéricos como por ejemplo el de Newton-Raphson. El problema puede tratarse de una forma mucho más simple con la transformación logarítmica. Tomando logaritmos en ambos lados de la igualdad tenemos que el modelo original se convierte en un modelo

lineal el las variables log(Y) y log(X).

$$\log(Y) = \log(\alpha) + \beta \log(X)$$

A cambio de trabajar en escala logarítmica, podemos utilizar los métodos de los modelos lineales.

Los efectos de ajustar un modelo lineal a datos que no lo siguen están relacionados con problemas de ajuste y predicción.

DETECCIÓN:

❖ Gráficos de residuales frente a variables externas en el caso de variables no incluidas. **Se observarán relaciones entre las variables externas y los residuales.**

❖ Gráficos de residuales frente a las variables regresoras, la variable dependiente o los valores ajustados. **Se observarán tendencias en los residuales. Bandas no homogéneas con tendencia curva definida** (ver figura 6.21)**.**

❖ Gráficos de residuales parciales, que representan los residuales del ajuste del modelo completo frente al residual más la componente de los valores ajustados debida a cada una de las variables regresoras $e_i + \hat{\beta}_k x_{ik}$. El gráfico se interpreta como la relación entre Y y X_k pero ajustada para el resto de las variables, es decir cuando las otras variables han sido ya consideradas en el modelo. **Se observarán tendencias en los residuales.**

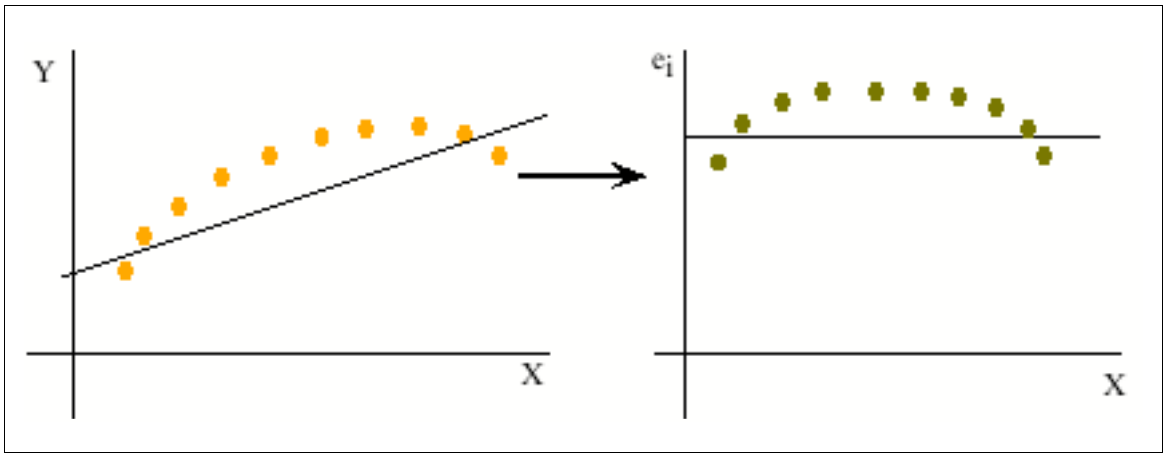


Figura 6.21: Gráfico de residuales mostrando una tendencia no lineal y diagrama de dispersión correspondiente.

TRATAMIENTO

❖ Inclusión de las variables externas que expliquen la componente no lineal.

❖Transformación de las variables regresoras causantes de la no linealidad o de la variable dependiente. En muchos casos, como el del ejemplo mencionado antes la transformación de las variables regresoras, de la dependiente o de ambas, convierte un modelo no lineal en uno que lo es. Las transformaciones más habituales son:

- ❖ Inclusión de términos de orden mayor (cuadráticos, cúbicos).
- ❖ Logaritmos de las regresoras o de la dependiente.
- ❖ Transformaciones inversas.

Para el investigador aplicado el proceso de transformación de los datos y de selección del modelo más adecuado suele ser un proceso interactivo en el que se van probando distintos modelo seleccionando aquel que proporcione un mejor ajuste, no solo en cuanto al poder explicativo sino también en cuanto al poder predictivo.

6.2.16.3 Homocedasticidad (igualdad de varianzas)

Relación entre la magnitud del error y el valor esperado de la variable dependiente o los valores de las regresoras. La varianza de los errores no es la misma para todas las observaciones.

Los efectos de la desigualdad de las varianzas son:

-Estimadores insesgados y consistentes pero no de varianza mínima.

-Problemas de variabilidad en las predicciones al ser muy diferente dependiendo del valor de la predicción..

DETECCIÓN

❖ Gráficos de residuales con bandas crecientes, decrecientes o combinaciones de ambas. En general, bandas de residuales con distintas anchuras para distintos valores ajustados (ver figura 6.22).

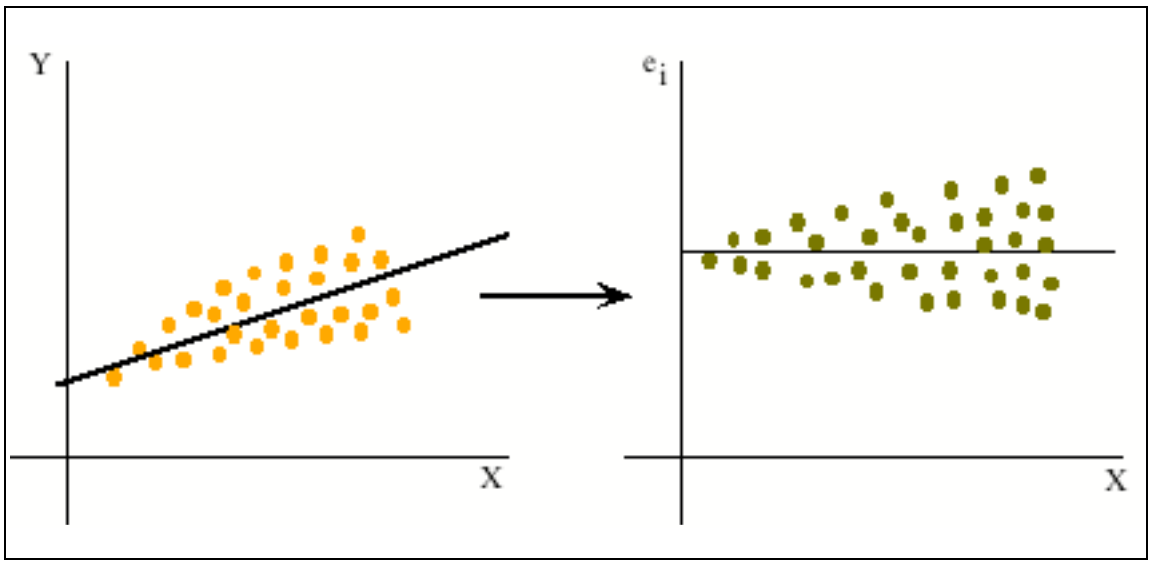


Figura 6.22: Gráfico de residuales con problemas de heteroscedasticidad y diagrama de dispersión corresponden en el caso de la regresión simple.

TRATAMIENTO

❖ **Mínimos cuadrados generalizados cuando se tiene información previa sobre la naturaleza de la heteroscedasticidad.**

Por ejemplo cuando los datos son medias de distinto número de observaciones tomadas todas ellas de una población con la misma varianza. Sabemos entonces que la varianza es inversamente proporcional al tamaño muestral para cada caso. El estudio de los mínimos cuadrados generalizados está fuera del alcance de este trabajo.

❖ **Transformaciones de la variable dependiente estabilizadoras de la varianza.**

Por ejemplo si las observaciones variable dependiente son recuentos de Poisson es claro que media y varianza coinciden, de forma que si la media aumenta linealmente con las regresoras, también lo hará la variabilidad. Tomar la raíz cuadrada de la variable dependiente en lugar de la propia variable suele estabilizar la varianza.

6.2.16.4 Autocorrelación

El problema se produce cuando los errores de las distintas observaciones no son independientes. Es frecuente cuando se trabaja con datos temporales o recogidos con un determinado orden.

EFECTOS

-Los estimadores mínimo cuadráticos son insesgados pero no tienen varianza mínima.

-Varianza del error subestimada.

-Varianza de los estimadores subestimada.

-La inferencia (t y F) no es estrictamente aplicable.

DIAGNOSTICO:

❖ **Gráficos de residuales que muestran tendencias cíclicas, tendencias lineales o no lineales o alternancia positivo-negativo** (ver figura 6.23).

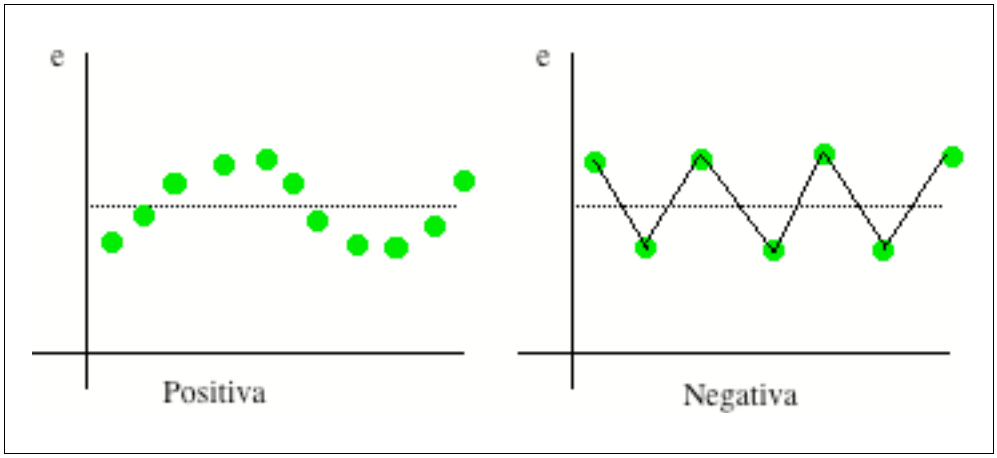


Figura 6.23: Gráficos de residuales en presencia de autocorrelación..

❖ **Gráficos de residuales para diferentes momentos de tiempo** (Residuales para cada momento del tiempo frente a residuales en el momento anterior) que mostrarán tendencias lineales (ver figura 6.24).

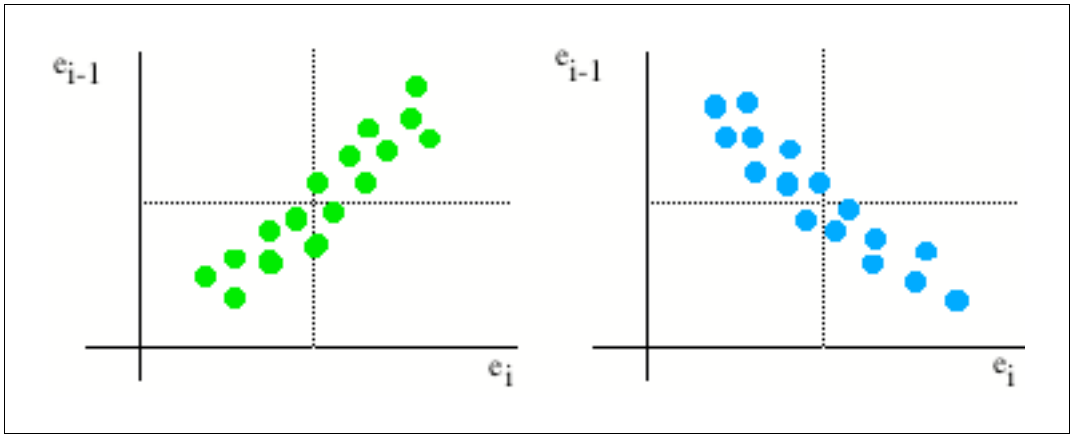


Figura 6.24: Gráficos de residuales en presencia de autocorrelación..

TRATAMIENTO:

❖ **Mínimos cuadrados generalizados cuando se tiene información previa sobre la naturaleza de la autocorrelación.**

"EL PROBLEMA DE LA COLINEALIDAD"

6.3 Ampliación

6.3.1 El problema de la colinealidad

En muchas ciencias, ocurre con frecuencia que las variables consideradas en el análisis, no son independientes. Por lo tanto los riesgos al hacer estimaciones son incalculables, y un porcentaje de variaciones explicadas muy alto por el modelo de regresión puede ser perfectamente compatible con un modelo sin ningún poder predictivo. Esta problemática se conoce con el nombre de colinealidad.

La colinealidad desde el punto de vista estadístico, no se corresponde con ninguna definición matemática concreta, ya que existen múltiples estadíos intermedios entre la ausencia total de colinealidad y la colinealidad extrema.

Se dice que hay colinealidad cuando existe relación lineal entre las regresoras y diremos que la colinealidad está ausente cuando las regresoras son ortogonales. En el caso de colinealidad extrema, es decir, si al menos dos regresoras están perfectamente relacionadas, los coeficientes de regresión mínimo cuadráticos no están definidos.

El problema surge cuando se da una colinealidad no perfecta, ya que entonces los estimadores de los coeficientes de regresión se hacen inestables, pudiendo -incluso- aparecer con signo contrario al que cabría esperar.

En esta sección vamos a llevar a cabo el estudio de los métodos para detectar y tratar de paliar este problema, que viene como consecuencia de una causalidad compleja, ya que el efecto de una variable puede ser causa de otra e incluso de más, o también se pueden afectar mutuamente.

6.3.2 Formulación matemática y gráfica del problema

El modo de visualizar gráficamente los distintos casos que nos podremos encontrar: *Ausencia de colinealidad*, *Colinealidad perfecta*, y *Fuerte colinealidad*, puede verse en las figuras 6.25; 6.26 y 6.27 respectivamente, (Para simplificar consideraremos únicamente dos variables regresoras X_1 y X_2. Tomado del FOX, 1984 pág. 139, adaptado de BELSLEY, KUH, & WELSCH, 1980).

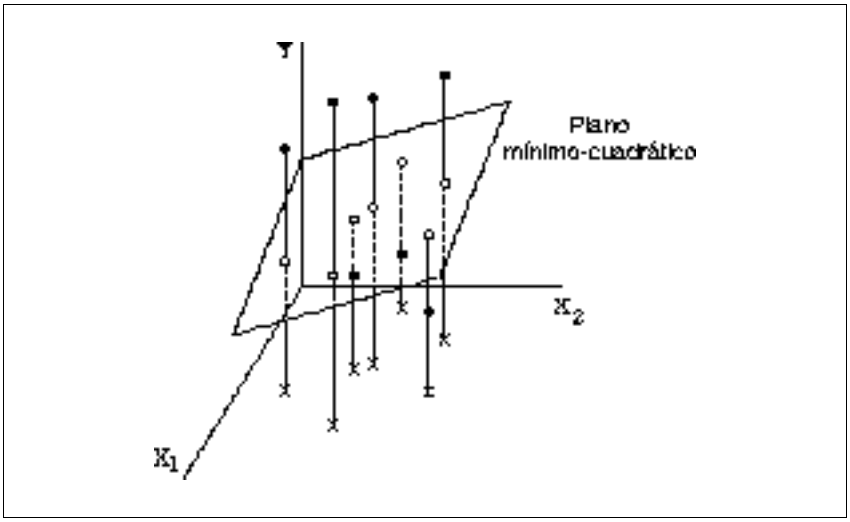


Figura 6.25: Correlación entre X_1 y X_2 despreciable: *Ausencia de colinealidad*

Para que los estimadores de los coeficientes de regresión estén definidos, la matriz **X'X** debe ser no singular, ya que si no $(\mathbf{X'X})^{-1}$ no estaría definida.

El rango de la matriz **X** es el mismo que el de **X'X**, siendo el número de variables regresoras (k); si tenemos 'n' observaciones debe cumplirse lo siguiente:

1.- Las variables X_j deben ser independientes; si esto no ocurre, y alguna es combinación lineal perfecta de otras, el determinante de **X'X** se anularía (es decir, tanto **X** como **X'X** serían singulares) y ello conlleva a que $(\mathbf{X'X})^{-1}$ no esté definida y no podamos hacer las estimaciones de los parámetros. Este es el caso de COLINEALIDAD PERFECTA (ver figura 6.26)

Pero cuando la colinealidad no es perfecta (ver figura 6.27), los estimadores de los parámetros de regresión se hacen inestables, de modo que nos podemos encontrar incluso anomalías tan graves como un signo contrario al que realmente debería tener.

Además cuanto mayor sea el grado de colinealidad más difícil resulta establecer el aporte particular de cada una de las variables regresoras, ya que si el coeficiente de correlación es distinto de cero, el tanto por ciento de variaciones explicado por una variable va a depender del resto de las variables que estén en el modelo, influyendo incluso el mayor o menor grado de asociación entre las regresoras y la variable dependiente.

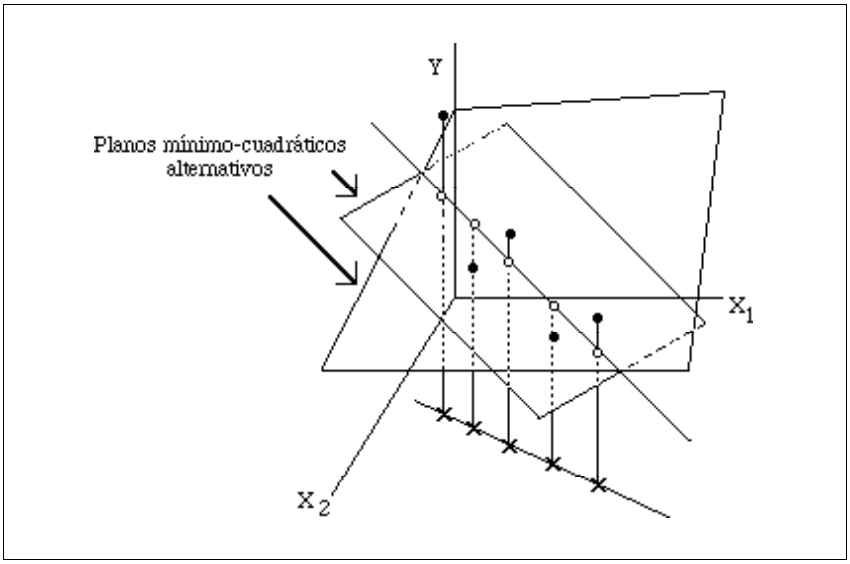


Figura 6.26: Colinealidad perfecta $X_1 = a + b\,X_2$

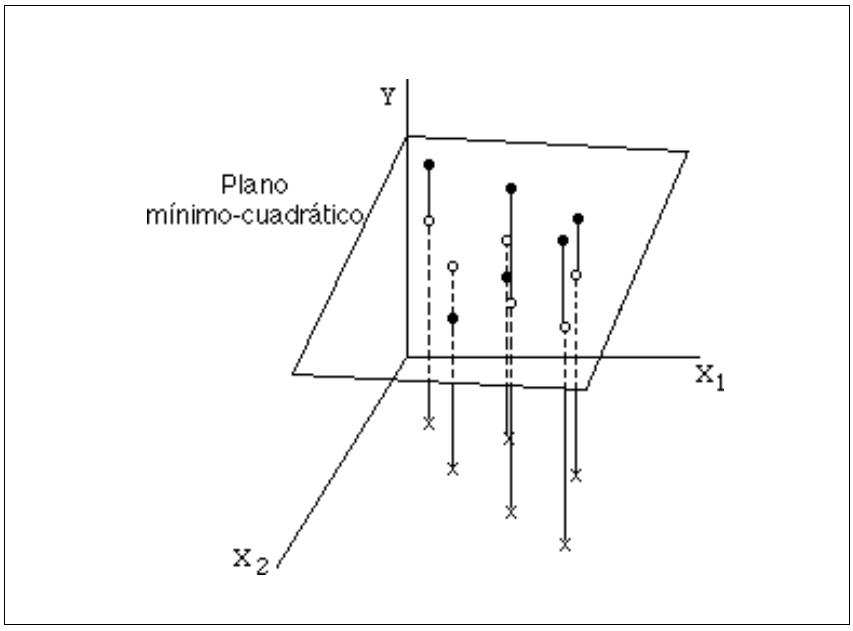


Figura 6.27: Colinealidad

Así la común interpretación de los coeficientes de regresión como medida del cambio en el valor esperado de la variable dependiente cuando la correspondiente variable independiente se incrementa en una unidad, cuando todas las demás variables regresoras permanecen constantes, no es totalmente aplicable cuando existe colinealidad.

2.- Debe verificarse que $n \geq k+1$, ya que sino el rango no sería k sino n, y por tanto podríamos detectar una falsa colinealidad.

Siguiendo a GALINDO (1987), dividiremos el análisis de esta problemática en tres grandes apartados. El primero de ellos consistirá en encontrar los indicios que nos puedan hacer sospechar sobre la existencia de colinealidad. Pasaremos en un segundo punto a estudiar cómo realizar el diagnóstico adecuado de la misma, y en último lugar analizaremos las técnicas existentes para tratar de paliar dicho problema.

6.3.3 Sintomatología

El primer paso para poder actuar frente a la colinealidad, es tomar conciencia de su posible existencia.

Hay una serie de síntomas o indicios que pueden presentarse cuando se da el problema de la colinealidad. Entre otros citaremos los siguientes:

1.- El valor absoluto de la correlación empírica entre dos variables regresoras varía entre 0 y 1 (en el caso de que no exista colinealidad o que ésta sea total, respectivamente). Por ello, si al analizar la matriz de correlaciones, se detecta que un subconjunto de dichas variables está altamente correlacionado, será un síntoma a tener en cuenta.

2.- Si las pruebas de nulidad de los coeficientes de regresión, conducen a eliminar del modelo variables que el investigador, basándose en su experiencia, considera relevantes.

3.- Si el signo de un coeficiente de regresión es opuesto al que cabría esperar.

4.- Si las varianzas de los estimadores de los coeficientes de regresión tienen valores anormalmente grandes, disminuyendo drásticamente al eliminar una o varias variables regresoras del modelo.

5.- Encontrar un coeficiente de correlación múltiple entre cada regresora y las demás muy elevado.

6.- Intervalos de confianza grandes para los coeficientes de regresión que representan a variables importantes en el modelo.

De todas formas, puede haber colinealidad sin que estos síntomas se hagan patentes.

6.3.4 Diagnóstico

Solamente la diagonalización de la matriz de correlaciones y el examen de los últimos valores propios proporcionará una información precisa.

Si tenemos k variables regresoras y llamamos $\lambda_1, \lambda_2, \ldots, \lambda_k$ a los k valores propios de su matriz de correlaciones en orden descendente, es decir $\lambda_1 > \lambda_2 > \ldots > \lambda_k$. Supondremos -sin pérdida de generalidad- que las variables están estandarizadas de forma que $\mathbf{X'X}$ sea proporcional a la matriz de correlaciones; entonces:

1.- El tamaño relativo de estos valores propios nos puede servir como indicador de la presencia de colinealidad, ya que como se verifica:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_k = k = \text{Traza}\left(\frac{1}{n}\mathbf{X'X}\right)$$

Si la razón λ_k / k es muy pequeña, entonces existe colinealidad.

2.- Hemos visto que los estimadores mínimo cuadráticos de regresión para variables estandarizadas son $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ con matriz de varianzas-covarianzas $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1}$.

El j-ésimo valor de la diagonal de $(\mathbf{X'X})^{-1}$ es precisamente $1/1-R_j^2$ siendo R_j^2 el cuadrado del coeficiente de correlación múltiple para la variable regresora X_j con el resto de las variables.

Al término $1/1-R_j^2$ se le denomina ***Factor de Inflación de la Varianza (VIF)*** y *es la cantidad que aumenta el error estándar del estimador j-ésimo por efecto de la correlación entre X_j y el resto de las variables regresoras.*

En condiciones óptimas (ausencia de colinealidad) $VIF_j = 1$ (ya que $R_j^2 = 0$). Conforme aumenta el problema de colinealidad el valor VIF se va haciendo cada vez

mayor de modo que el correspondiente estimador para la j-ésima variable se va haciendo cada vez más inestable. (THEIL, 1971). ***Por lo tanto, un VIF grande nos indica que el coeficiente de regresión asociado se encuentra afectado por el problema de colinealidad.***

Realizando la descomposición espectral de la matriz de correlaciones, tenemos:

$$\mathbf{X'X} = \mathbf{ALA'}$$

donde:

A es la matriz de vectores propios
L es la matriz diagonal de valores propios

Por lo tanto, podemos escribir: $(\mathbf{X'X})^{-1} = \mathbf{AL}^{-1}\mathbf{A'}$

Utilizando la anterior expresión, el VIF_j se pueden expresar en función de los valores propios de la matriz de correlaciones como sigue: $VIF_j = \sum_{r=1}^{k} A_{jr}^2 \Big/ \lambda_r$ donde A_{jr} es la j-ésima componente del l-ésimo vector propio, de modo que aquellos valores propios más pequeños son los que más contribuyen a que las varianzas sean más grandes, pero sólo para aquellas regresoras que tienen coeficientes grandes asociados a vectores propios con valores propios muy pequeños. ***Por lo tanto, regresoras con coeficientes grandes para componentes cortas, son la implicadas en la colinealidad.***

> Por ello, basta con realizar la descomposición espectral de la matriz de correlaciones entre las regresoras, analizar los valores propios, cuando uno de ellos sea próximo a cero, nos está indicando un posible problema de colinealidad, de modo que aquellas regresoras cuyos coeficientes del vector propio asociado al valor propio cercano a cero, sean muy grandes serán las que se encuentren implicadas en la colinealidad.

3.- Además, la relación entre los valores propios nos sirve como indicador del grado de colinealidad existente en nuestros datos. De este modo, la raíz cuadrada de la razón existente entre el primer autovalor y el último (mayor y menor respectivamente):

$$K = \sqrt{\lambda_1 \Big/ \lambda_k}$$

se denomina **"Condition number",** y es un índice de la inestabilidad global de los coeficientes de regresión mínimo cuadráticos (BELSLEY, KUH & WELSCH, 1980).

Los autores manifiestan que un "Condition *number*" grande, indica que, cambios relativamente pequeños en los datos, tienden a producir grandes cambios en la solución mínimo cuadrática; en este caso **X'X** será casi singular, de modo que valores de K >30 se consideran como "peligrosos".

4.- Estos mismos autores definen el **"condition index"** como:

$$K_j = \sqrt{\lambda_1 / \lambda_j}$$

5.- Analizando la contribución proporcional de cada componente principal al VIF para cada regresora, es posible ver ***qué variables están involucradas en la colinealidad:***

$$P_{jr} = \frac{\left(A_{jr}^2 / \lambda_r\right)}{VIF_j} = \frac{\left(A_{jr}^2 / \lambda_r\right)}{\sum_{r=1}^{k} A_{jr}^2 / \lambda_r}$$

Si P_{jr} es grande *(estudios de simulación llevan a pensar en valores próximos a 0.5)* ***y también K_r entonces la j-ésima regresora está implicada en la colinealidad.***

Cuando hay varias relaciones de colinealidad coexistentes, no siempre es fácil separar las variables involucradas en cada una.

Sin embargo, en la mayoría de las situaciones es suficiente determinar:

1- Si está presente una colinealidad importante.
2- Qué coeficientes de regresión están afectados por la colinealidad
3- Qué regresoras están involucradas en cada cuasi-dependencia

*El punto 1 se sigue del "**condition índices**"; el punto 2 del **VIF**; y el punto 3 de la **contribución de cada componente al factor de inflación**.*

6.3.5 Tratamiento

6.3.5.1 Análisis del origen de la colinealidad

En primer lugar hay que asegurarse de que lo que se detecta no es una **colinealidad aparente**, debida quizás a:

- Una muestra sesgada, dándose relaciones en ella que realmente no son ciertas en la población y que al elegir otra muestra quizás no las encontraríamos.

- Que tengamos en nuestro estudio menor nº de individuos que de variables, con lo que la inversa de la matriz **X'X** no estaría definida.

Supongamos que el examen de los valores propios, mediante las pruebas señaladas en el apartado 'Diagnostico', nos indican la **existencia de colinealidad**, entonces la actitud a tomar dependerá de cuál es su posible origen:

❖ Si observamos que se debe a una relación cuasi-funcional entre las variables regresoras, conviene mantenerlas a todas en el modelo, ya que posiblemente al eliminar una de ellas disminuya la suma de cuadrados de la regresión. *Así sólo se podrá interpretar la fórmula de modo global sin interpretar los coeficientes de cada una de las regresoras.*

❖ Si en realidad lo que queremos es estudiar precisamente las aportaciones o influencias de cada una de las variables independientes, deberemos analizar los primeros ejes principales normalizados en el espacio de las regresoras (vectores propios de su matriz de correlaciones), y calcular la correlación entre la variable dependiente y el primer eje principal, luego el coeficiente de correlación múltiple con los dos primeros ejes, y así sucesivamente.

a) Si al hacer esto la variable Y está suficientemente correlacionada con dichos ejes, entonces utilizaremos una regresión normal, ya que en este caso la colinealidad resulta beneficiosa, según los estudios de simulación de CARBONELL y cols. (1983)

b) Si por el contrario, la correlación no es alta, la solución consiste en la eliminación de las variables (aquellas que sean combinación lineal de otras), y que se

pueden identificar por distintos procedimientos como: Estudio del R^2, el factor de tolerancia, los métodos Biplot (GABRIEL, 1971; GALINDO, 1985, 1986), o con otros distintos como la REGRESIÓN RIDGE (HOERL Y KENNARD, 1970a, b), el método de MALLOWS (1964), o bien con los procedimientos PASO A PASO.

CARBONELL y cols (1983), propone el siguiente árbol de decisiones (figura 6.28) a la hora de analizar la problemática de la colinealidad:

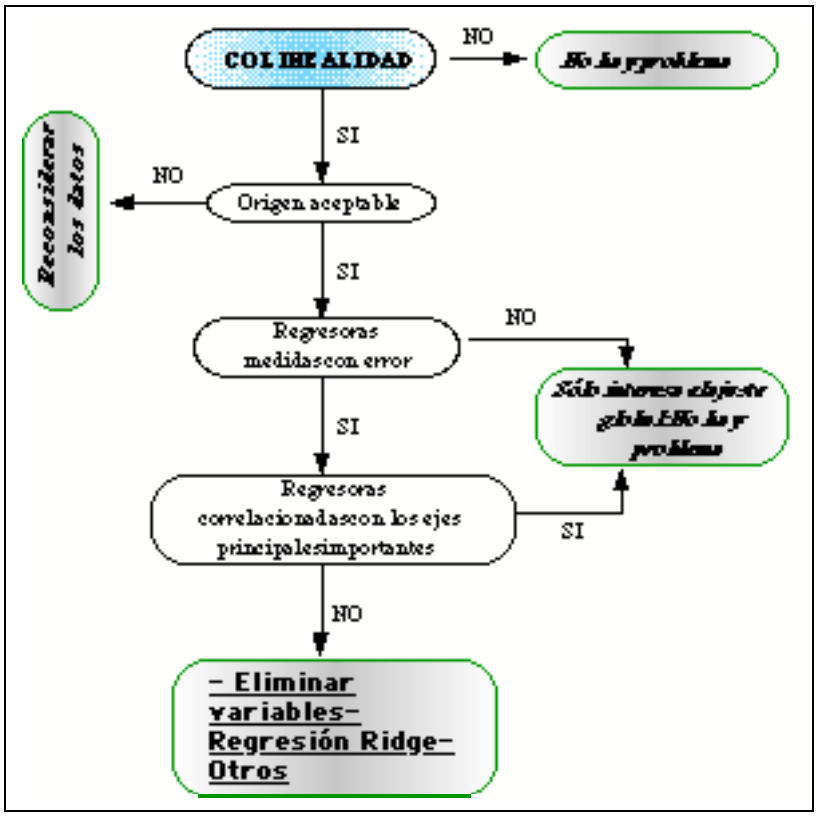


Figura 6.28: Árbol de decisiones en el análisis de la colinealidad (CARBONELL, op. cit.).

6.3.5.2 Selección de variables en regresión

INTRODUCCIÓN

En ese apartado haremos una síntesis de algunos de los métodos anteriormente citados para paliar la problemática de la colinealidad.

Pero hay algo que hay que tener muy en cuenta, y es que esta selección debe hacerse siempre después de un detallado estudio de la colinealidad. Este problema puede estudiarse en profundidad en NETER, WASSERMAN & KUTNER, 1985 y CARBONELL y cols. (1983).

Supongamos que 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, que sería el grupo total de variables entre las cuales estarán aquellas que formarán parte de la ecuación buscada.

Para que el modelo encontrado sea el más adecuado, deberemos incluir en él el mayor número de variables posible, cuyo efecto en la variable dependiente pueda ser interpretado, para así poder evitar un modelo con una gran varianza en las predicciones.

Obviamente, no existe un único procedimiento estadístico para llevar a cabo esta tarea, y es más, generalmente los diferentes métodos no conducen a la misma solución, por lo cual bajo nuestra experiencia, se deberá tener cierta cautela a la hora de utilizarlos, y sobre todo nunca debe menospreciarse el criterio del investigador a la hora de la selección del subconjunto de variables más adecuado, ya que su conocimiento sobre las variables en estudio puede ser vital a la hora de decidirse por la inclusión o exclusión de una de ellas en el modelo.

MÉTODO DE TODAS LAS REGRESIONES POSIBLES

Este método de selección consiste en calcular todas las posibles ecuaciones de regresión, combinando el número total de variables regresoras y luego hacer una selección de la ecuación óptima.

Como se puede intuir, se trata de un procedimiento laborioso y sólo es posible cuando se puede acceder a un ordenador de alta velocidad. Por ello hay otros más utilizados en la actualidad y que veremos con posterioridad.

El procedimiento consiste en lo siguiente:

El número de posibles ecuaciones de regresión es: $2^K - 1$, lo cual nos da ya una idea de la magnitud del mismo, de modo que cada variable regresora X_i (i = 1, 2, ... , K), puede estar o no incluida en la ecuación.

En primer lugar se separan las ecuaciones por grupos, de modo que tengamos un grupo con una variable regresora solamente, otro con dos, otro con tres, y así sucesivamente, hasta uno con K , que será $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$

Si denotamos con p al número de variables que hay en un modelo, entonces habrá p+1 parámetros en la función de regresión para ese grupo.

Por lo tanto se verifica:

$$1 \leq p \leq K$$

Hay distintos criterios que pueden ser utilizados para comparar los distintos modelos de regresión obtenidos:

Criterio R_p^2

Lo que se hace es examinar el coeficiente de determinación R_p^2, para seleccionar uno o varios subconjuntos de las variables regresoras, y donde p es el número de parámetros en el modelo. Así R_p^2 nos indica que hay p parámetros o p-1 variables en el mismo, y se va observando cómo varía R_p^2 al pasar de un modelo a otro. *Lo que se intenta es encontrar el modelo en el que añadiéndole más variables, no es ya útil, porque el incremento en* R_p^2 *es ínfimo*.

Criterio C_p de Mallows.

Nos permite seleccionar de entre todas las ecuaciones de regresión posibles cuál es la que tiene mejor bondad de ajuste.

Con C_p denotamos el "*error cuadrático medio total*" definido por MALLOWS (1964) y lo componen: la suma de las desviaciones al cuadrado respecto del modelo completo, y el cuadrado de los errores aleatorios en Y, para el conjunto total de n observaciones, es decir:

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

Como estimador generalmente se utiliza el cuadrado medio del error del modelo de regresión completo, bajo la hipótesis de que este modelo es verdadero. El C_p de aquellos modelos con poco sesgo tiende a ser cercano a p, de modo que podremos identificar los modelos que tengan un pequeño valor (Ver figura 6.29)

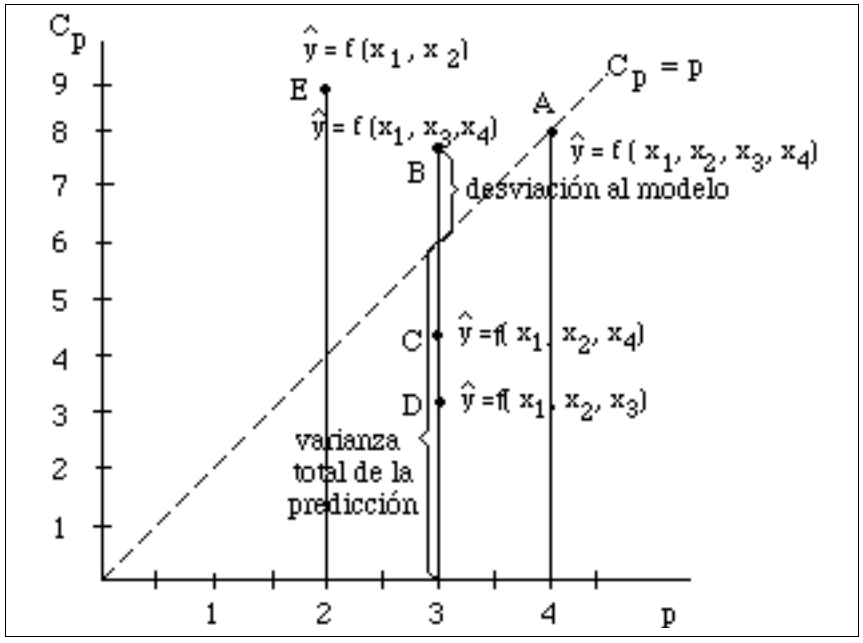


Figura 6.29: Criterio del Cp de Mallows. Tomada de CARBONELL y cols.(1983)

MÉTODOS PASO A PASO

Como hemos apuntado anteriormente, debido al alto grado de complejidad que posee el método de todas las regresiones posibles, se hacen necesarios otros que evalúen solamente un pequeño número de subconjuntos de variables, adicionando o eliminando éstas según determinados criterios.

Se han desarrollado algunas técnicas de estas características, que generalmente se denominan MÉTODOS PASO A PASO (Stepwise Methods), y que consisten en

** Siendo SCE_p la suma de cuadrados del error (para el modelo de p variables).

variaciones de dos ideas básicas: Eliminación descendente* y Selección Ascendente**. Se ha hecho una breve referencia a estos métodos en el apartado 6.13. Aquí comentaremos un poco más.

❖ Selección ascendente

Se comienza sin ninguna variable en el modelo y se va añadiendo una a una hasta que se obtenga una ecuación satisfactoria -según un determinado criterio- o bien hasta que se haya completado la inclusión de todas ellas.

Generalmente el criterio de entrada, consiste en introducir aquella variable que proporcione el máximo incremento en el coeficiente de correlación múltiple.

HOCKING, propone en 1976, incluir la variable i-ésima en la ecuación con p términos si:

$$F_i = \max_i \frac{SCR_p - SCR_{p+i}}{\sigma^2_{p+i}} > F_{input}$$

donde:

SCR_p es la suma de cuadrados de los residuales con un subconjunto p de variables

SCR_{p+i} es la suma de cuadrados de los residuales añadiendo la i-ésima variable a un subconjunto p de variables.

Se calcula, por lo tanto el término F_i añadiendo una a una las variables que no están en el modelo y se busca la variable para la cual ese valor es máximo, esa es precisamente la que entra en el modelo si $F_i > F_{input}$.

Si para todo i, $F_i < F_{input}$ el proceso termina.

❖ Eliminación descendente

Se parte del modelo contrario, es decir, con todas las variables regresoras incluidas en el mismo, y según un determinado criterio vamos eliminando variables del modelo hasta encontrar aquella ecuación más adecuada.

---

* Del término inglés: *Backward Elimination (BE)*

** Del término inglés: *Forward Selection (FS)*

La variable i-ésima será excluida del modelo con p términos si:

$$F_i = \min_i \frac{SCR_{p-i} - SCR_p}{\sigma_p^2} < F_{out}$$

donde:

SCR_{p-i} denota la suma de cuadrados de los residuales cuando la variable i es borrada de la ecuación en la que había p términos

Se calcula la expresión F_i, eliminando una a una las variables que forman parte del modelo, y se busca la variable para la cual es mínima esa expresión; esa variable es la que se elimina si $F_i < F_{out}$.

Si para todo i, $F_i > F_{out}$ el proceso termina.

El método de inclusión de variables en el modelo de regresión (selección ascendente), presenta la ventaja de que sólo se maneja el número de variables estrictamente necesario, pero en ningún caso se estudia el efecto que puede producir la inclusión de una variable en el papel que desempeñan las ya incluidas en modelos anteriores.

❖ "Regresión Stepwise" (EFROMYSON, 1960)

Para solventar el problema citado anteriormente, EFROMYSON propuso en 1960 el método de *Regresión Stepwise* que se corresponde más con lo que entendemos como métodos paso a paso.

Consiste en una selección ascendente (FS), pero en cada paso consideramos la posibilidad de eliminar una variable, de modo similar a como se hace en el método de eliminación descendente (BE).

Una variable que fue la mejor candidata 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.

El proceso Stepwise continua hasta que ninguna variable pueda ser introducida y ninguna eliminada. Es menos riguroso estadísticamente que los anteriores (CARBONELL y cols, 1983). Es el que se emplea normalmente al utilizar programas estándar.

Una crítica a los métodos FS y BE es que los investigadores, generalmente dan un grado de importancia a las variables, dependiendo del orden en el que entran (FS) o en el que salen (BE), lo cual no es correcto, ya que no es raro encontrarnos con que la primera que entra en uno es la primera que sale en el otro* , o que incluso en el método stepwise entra en un paso y sale en el siguiente.

También se critican porque no proporcionan resultados óptimos, ya que puede que no identifiquen aquellos subconjuntos de regresoras de determinado tamaño, de modo que maximicen R^2, incluso cuando es éste el criterio utilizado para la inclusión en el modelo.

Más detalladamente se pueden encontrar estos métodos en DRAPER y SMITH (1966) CHATTERJEE y PRICE (1977).

❖ Regresión Ridge

Hasta ahora el método de ajuste de los coeficientes de regresión utilizado ha sido el de los mínimos cuadrados, y según el teorema de Gauss-Markov, este método de ajuste nos proporciona estimadores eficientes, es decir, insesgados y de varianza mínima, bajo las condiciones del modelo de regresión.

En presencia de colinealidad, como hemos visto, se incrementa notablemente la varianza muestral de los estimadores, con lo que disminuye, por lo tanto, su eficiencia.

Para intentar paliar esto, utilizaremos un método mediante el cual podramos encontrar estimadores sesgados de modo que disminuya la varianza muestral, ya que el error cuadrático medio de un estimador es la suma de su varianza muestral y el cuadrado del sesgo.

---

* En el trabajo de investigación correspondiente a este capítulo veremos cómo, efectivamente, la primera de las variables que entra utilizando el método de selección ascendente, y a la que por lo tanto, el investigador no familiarizado con estas técnicas le daría la máxima importancia, es precisamente la que sale en primer lugar utilizando el método de eliminación descendente.

La Regresión Ridge fue originalmente propuesta por HOERL (1962)y posteriormente elaborada por HOERL y KENNARD (1970a,b) . *Consiste en un método de estimación sesgado que busca mejorar la acción de la estimación mínimo-cuadrática en presencia de colinealidad.*

Se propone como vector de estimadores de los coeficientes de regresión:

$$\hat{\boldsymbol{\beta}}^{*} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{W}\mathbf{X}'\mathbf{y} \qquad k>0$$

Dando valores a k se encuentra una familia de estimadores denominada *ESTIMADORES RIDGE.*

El módulo del estimador Ridge es menor que el del estimado por el método mínimo-cuadrático, ya que éstos son demasiado grandes cuando **X'X** es casi singular (HOERL & KENNARD, 1970 a).

El principal problema al aplicar la regresión Ridge está en encontrar aquel valor de k de modo que se compense el sesgo y la reducción de varianza.

Se han desarrollado muchos métodos para seleccionar el valor de k. Algunos son aproximativos y otros proporcionan fórmulas específicas.

HOERL & KENNARD (1970, a,b) sugieren el "TRAZADO RIDGE", en el que se representan valores de los estimadores dependiendo del valor de k. En el se pone de manifiesto la inestabilidad de los coeficientes de regresión y el incremento de la suma de los cuadrados. (Veremos en el trabajo de investigación este tipo de trazado gráficamente)

Llega un momento, cuando se continúa incrementando k, en que los coeficientes se estabilizan. Durante este proceso los VIF decrecen, al principio rápidamente y luego de modo más gradual.

La estimación de la varianza de los errores $S^2_{E^*} = \hat{\sigma}^2_{\varepsilon^*}$ aumenta suavemente cuando se incrementa k.

Entonces, para seleccionar k podremos tener en cuenta el trazado Ridge, los VIFs y la varianza del error.

HOERL & KENNARD, proponen en el mismo trabajo elegir k de modo que los coeficientes de regresión estén estabilizados y la varianza del error no se incremente desde su valor mínimo

MARQUARDT & SNEE (1975) sugieren elegir k de modo que el máximo VIF sea menor de 10, y preferiblemente no mucho mayor que 1.

La regresión Ridge también puede ser utilizada como método de selección de variables, eliminando aquellas regresoras cuyos coeficientes de regresión tiendan a 0 tan rápidamente como se incremente k (MARQUARDT & SNEE (1975); HOCKING, 1976). Pone de manifiesto aquellos coeficientes inestables que deben ser eliminados del modelo porque no son capaces de mantener su poder predictivo.

En la figura 6.30 puede verse un ejemplo de trazado Ridge (que será el del ejemplo que utilizaremos en el trabajo de investigación) que evidencia la inestabilidad de los coeficientes de regresión y el incremento en suma de cuadrados.

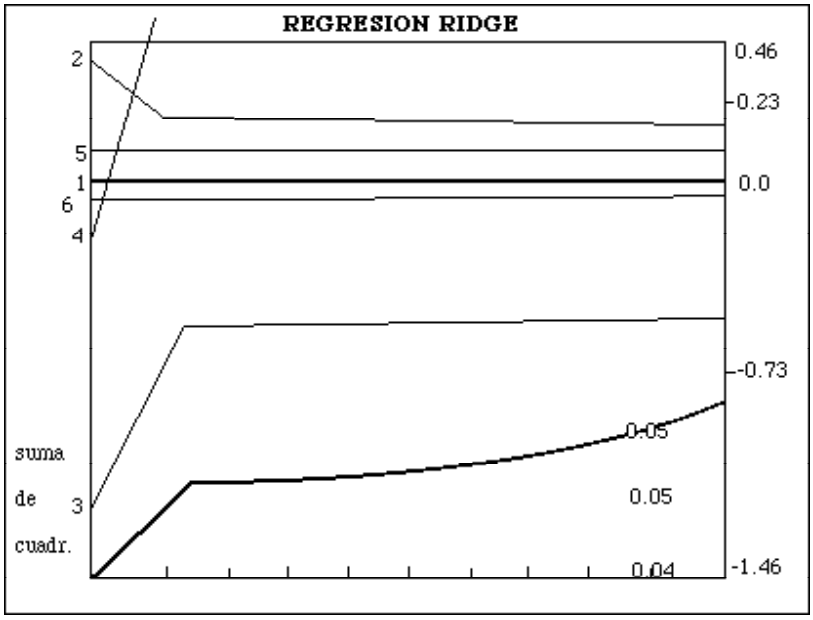


Figura 6.30: Trazado Ridge correspondiente al ejemplo del trabajo de investigación de este capítulo

BIBLIOGRAFIA CITADA

BELSLEY, D.A.; KUH, E. & WELSCH, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley. New York.

CARBONELL, E.; DENIS, J.B; CALVO, R; GONZALEZ, F. y PRUÑONOSA, V. (1983). *Regresión Lineal: Un enfoque conceptual y práctico*. I.N.I.A.

CHATTERJEE, S & PRICE, B (1977). *Regression Analysis by Example*. Wiley. New York.

DRAPER, N.R. & SMITH, H. (1966). *Applied Regression Analysis*. Wiley. New York.
EFROMYSON, M.A. (1960). 'Multiple regression analysis'. In A. Ralston & H.S. Wilf (eds.) *Mathematical Methods for Digital Computers*. Vol. **1**: 191-203.

FOX, J. (1984). *Linear Statistical Models and Related Methods*. New York. Wiley.
GABIEL, K.R. (1971). 'The biplot graphic display of matrices with applications to principal component analysis'. *Biometrika,* **58**: 543-467.

GALINDO, M.P. (1985). 'Contribuciones a la representación simultánea de datos multidimensionales'. Ser. Resum. Tesis Doct. T.D. 395/1985. pgs 1-38. Universidad de Salamanca.

GALINDO, M.P. (1986). 'Una alternativa de representación simultánea: HJ-Biplot'. *Questiio*. Vol.**10**, nº**1**: 13-23.

GALINDO, M.P. (1987). 'Diagnóstico y tratamiento de los problemas en los modelos lineales'. *Cuadernos de Bioestadística y su Aplicación Informática*. Vol. **5**, nº1**:** 116-128.

HOCKING, R.R. (1976). 'The analysis and selection fo variables in linear regression'. *BIOMETRICS*. **32**: 1-49.

HOERL, A.E. (1962). 'Application of Ridge Analysis to regression problems'. *Chemical Engineering Progress,* **58**: 54-59.

HOERL, A.E. & KENNARD, R.W. (1970a). 'Ridge Regression: Biased estimation for nonorthogonal problems'. *Technometrics,* **12**: 55-67.

HOERL, A.E. & KENNARD, R.W. (1970a). 'Ridge Regression applications to nonorthogonal problems'. *Technometrics,* **12**: 69-82.

MALLOWS, C.L. (1964). 'Choosing variables in a linear regression: a graphical aid'. Presented at the central Regional Meeting of the Inst. of Math. Statist. Manhattan, Kansas.

MARQUARDT, D.W; & SNEE, R.D. (1975)· 'Ridge regression in practice'. *The American Statistician*, **29**: 3-20

NETER, J.; WASSERMAN, W. & KUTNER, M.H. (1985). *Applied Linear Statistical Models*. (2nd. Ed.) Richard D. Irwin, INC

THEIL, H. (1971). *Principles of Econometrics*. New York. Wiley

"COLINEALIDAD"

6.4 Trabajo de investigación

En este apartado trabajaremos sobre un estudio de simulación que nos permita poner de manifiesto cómo en presencia de colinealidad, los estimadores clásicos de Gauss-Marcov proporcionan estimaciones sesgadas e inestables que no son interpretables. Asimismo, se pretende poner de manifiesto la cautela con la que debe trabajarse al utilizar los métodos de regresión paso a paso, tan profusamente utilizado por los investigadores en todos los ámbitos científicos.

6.4.1 Modelo establecido "a priori"

Sean X_1, X_2, X_3, X_4 variables cuyos valores son obtenidos con ayuda de un generador de números aleatorios.

Tomamos X_4 de manera que sea combinación lineal de otras tres; es decir:

$$X_4 = 1250 + 6.5X_2 - 20.7X_3 + \varepsilon$$

La variable dependiente se elige deliberadamente según el siguiente modelo:

$$Y = 1350 - 3X_1 + 12X_2 - 20X_3 + 15X_4 + 25X_5 - 13X_6 + \varepsilon$$

6.4.2 Sintomatología

6.4.2.1 Estimación de los coeficientes de regresión

La matriz de correlaciones **X'X** entre las variables independientes es la que aparece a continuación (ver tabla 6.2).

	1	**2**	**3**	**4**	**5**	**6**
1	1.000	0.057	0.130	-0.115	0.048	0.152
2	0.057	1.000	0.231	0.063	0.051	-0.264
3	0.130	0.231	1.000	-0.956	0.010	-0.238
4	-0.115	0.063	-0.956	1.000	0.004	0.165
5	0.048	0.051	0.010	0.004	1.000	-0.245
6	0.152	-0.264	-0.238	0.165	-0.245	1.000

Tabla 6.2. Matriz de correlaciones entre las variables

Vemos como el coeficiente de correlación entre las variables X_4 y X_3 es próximo a 1, lo cual es ya un primer indicio sobre la posible existencia de colinealidad.

Los estimadores mínimo-cuadráticos para el modelo de regresión son los que aparecen en la tabla 6.3:

Número	*Coeficiente*	*Error estándar*	*Estadístico t*
Corte	21789.6569		
1	-2.7580	1.8872	-1.4614
2	130.8591	32.0335	4.0851
3	-393.3484	102.6947	-3.8303
4	-3.0357	4.9657	-0.6113
5	23.4743	1.4740	15.9253
6	-15.1239	1.7799	-8.4971

Tabla 6.3: Parámetros del modelo de regresión

Los errores estándar para las variables 2. 3 y 4 son muy grandes lo cual es también un síntoma de una potencial colinealidad.

Resumen del análisis

Varianza residual:	69992.3431
% de variaciones no controladas:	0.0293
Coeficiente de determinación:	0.9991
% de variaciones controladas	***99.91%***
Coeficiente de correlación múltiple:	***0.9996***

Obsérvese cómo a pesar de que el porcentaje de variaciones explicadas es 99.91%. los valores de los estimadores de algunos de los coeficientes de regresión difieren sensiblemente de los verdaderos coeficientes (ver tabla 6.2). *siendo incluso en alguno de los casos de signo contrario al que debería (lo que ocurre con el de la variable 4). lo cual es también un síntoma del posible problema de colinealidad.*

Vemos asimismo cómo el coeficiente de correlación múltiple es muy alto.

6.4.3 Diagnóstico

Para hacer un efectivo diagnóstico del problema. deberemos conocer:

1.- Si está presente una colinealidad importante
2.- Qué coeficientes de regresión están afectados por la misma.
3.- Qué regresoras está involucradas en la cuasi-dependencia.

Para ello deberemos. respectivamente. conocer el "condition number". los factores de inflación de la varianza y la contribución de cada componente al factor de inflación.

Seguiremos los siguientes pasos:

6.4.3.1 Cálculo de los valores propios de la matriz de correlaciones

Comenzaremos estudiando los valores propios de la matriz de correlaciones de las regresoras (ver tabla 6.4):

	1	2	3	4	5	6
Valor propio	2.0922	1.3341	1.0419	0.9634	0.5682	***0.0001***

Tabla 6.4: Valores propios de la matriz de correlaciones entre las regresoras

Vemos como el último valor propio es muy próximo a cero. lo cual nos indica ya que deberemos estar alerta por un posible problema de colinealidad. pues nos está indicando que la matriz **X'X** es casi singular.

6.4.3.2 Estudio de los vectores propios de la matriz de correlaciones entre las regresoras

Analizaremos ahora la matriz de vectores propios de las regresoras. puesto que deberemos localizar cuáles son las variables con coeficientes grandes en componentes

cortas (ver tabla 6.5) (vimos en el paso anterior. cómo el último vector propio era próximo a cero).

	1	2	3	4	5	6
1	0.1198	-0.2757	0.8602	-0.0337	-0.4107	0.0024
2	0.1838	0.4696	0.3232	0.6869	0.3579	-0.2033
3	0.6745	-0.1377	-0.0567	0.0593	0.1697	***0.7004***
4	-0.6366	0.2813	0.1517	0.1427	-0.0644	***0.6842***
5	0.0854	0.4813	0.3082	-0.7084	0.4054	-0.0009
6	-0.2905	-0.6113	0.1856	0.0375	0.7113	-0.0009

Tabla 6.5: Matriz de vectores propios para las regresoras

La tabla anterior (tabla pone de manifiesto que las variables X_3 y X_4 son las que están implicadas en la colinealidad. *(Vemos como esta afirmación coincide con la construcción del modelo. además el siguiente coeficiente más grande se corresponde con la variable X_2).*

6.4.3.3 Cálculo del "Condition Index" y del "Condition number"

El valor para el "*condición number*" es 135.21 lo cual evidencia la inestabilidad global de los coeficientes mínimo-cuadráticos (recordemos que se considera peligroso para valores mayores de 30).

Los "*condition index*" para las distintas componentes principales aparecen en la tabla 6.6:

1	2	3	4	5	6	
		1	1.2523	1.4171	1.4737	1.9188	***135.2131***

Tabla 6.6: "Condition index"

El alto valor para el index correspondiente a la variable indica una vez más *que una colinealidad importante está presente*.

6.4.3.4 Factores de inflación de la varianza (VIF)

Los factores de inflación (V I F) para cada regresora son los que aparecen en la tabla 6.7:

1	2	3	4	5	6
1.1229	***362.0870***	***4287.0318***	***4090.7179***	1.0849	1.2528

Tabla 6.7: Factores de inflación de la varianza para cada regresora

Los V I F para las variables 2. 3 y 4 son muy grandes; valdrían 1 en el caso de ser ortogonales. Nos están indicando que. efectivamente. son los coeficientes para dichas variables los que se ven afectados por el problema de colinealidad. La misma información se obtiene estudiando el incremento en el error estándar de cada regresora.

6.4.3.5 Incremento en el error de cada regresora

Calcularemos. por tanto. el incremento relativo en el error estándar de cada coeficiente de regresión. debido a la colinealidad (no es más que la raíz cuadrada del VIF correspondiente).

Estos valores son los que aparecen en la tabla 6.8:

1	2	3	4	5	6
1.0597	***19.0286***	***65.4754***	***63.9587***	1.0416	1.1193

Tabla 6.8: Incremento relativo en el error estándar del coeficiente para cada regresora

Obsérvese cómo el error estándar para las variables 2. 3 y 4 se ha incrementado sensiblemente por efecto de la colinealidad (ver para la comparación la tabla 6.3) como cabría esperar. ya que la variable X_4 se había construido como combinación de X_2 y X_3.

6.4.3.6 Identificación de las variables involucradas en el problema de la colinealidad

Nos falta aún. identificar las variables ***involucradas*** en la relación de colinealidad. para lo cual calcularemos la contribución proporcional de los componentes a los factores de inflación de varianza. que se recogen en la siguiente tabla 6.9:

	Componente					
Var	1	2	3	4	5	6
1	0.0061	0.0507	0.6324	0.0010	0.2643	0.0455
2	0.0000	0.0005	0.0003	0.0014	0.0006	***0.9972***
3	0.0001	0.0000	0.0000	0.0000	0.0000	***0.9999***
4	0.0000	0.0000	0.0000	0.0000	0.0000	***0.9999***
5	0.0032	0.1600	0.0840	0.4801	0.2666	0.0061
6	0.0322	0.2236	0.0264	0.0012	0.7108	0.0058

Tabla 6.9: Contribuciones proporcionales de las componentes a los VIF
(los valores superiores a 0.5 se consideran peligrosos)

Como la contribución proporcional de los componentes a los V I F son muy grandes para las variables 2.3.4. es evidente que estas tres variables están implicadas en la cuasi-dependencia representada por la 6ª componente.

6.4.4 Conclusiones

Según hemos podido comprobar los estimadores mínimo cuadráticos son inestables y pierden. por tanto. su poder predictivo. poniendo de manifiesto la importancia de llevar a cabo un estudio sobre la posible colinealidad a la hora de llevar a cabo un análisis de regresión múltiple. pues dicho problema puede llevarnos a conclusiones totalmente erróneas.

6.4.5 Tratamiento de la colinealidad

Una vez finalizada la fase de sintomatología y diagnóstico analizaremos distintas opciones para su tratamiento. bajo el supuesto de querer interpretar las influencias respectivas de cada regresora regresoras.

6.4.5.1 Método de selección ascendente

Tratamos de eliminar aquellas variables que sean correlación lineal de otras. Utilizando el método de selección ascendente (Forward Selection).

La primera variable que interviene en el modelo es la variable 4. La prueba de bondad de ajuste global y los parámetros del modelo de regresión aparecen en la tabla 6.10:

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	-1453.43			
X_4	15.92	0.2487	64.00	p<0.0001

Tabla 6.10

En el paso número 2 entra la variable 5 y la tabla 6.11 recoge la prueba de significación y los parámetros del modelo:

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	-2847.09			
X_4	15.92	0.14	107.41	p=0.0001
X_5	27.51	2.93	9.39	p=0.0000

Tabla 6.11

En el paso número 3 la variable introducida es la 5. En el paso número 4, la variable 2 y en el paso número 5 la variable 3.

La prueba de significación y los parámetros para el modelo de regresión aparecen en la tabla siguiente (tabla 6.12):

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	22648.48			
X_2	135.43	32.28	4.19	p=0.0001
X_3	-408.95	103.44	-3.95	p=0.003
X_4	15.92	0.14	107.41	p=0.0000
X_5	27.51	2.93	9.39	p=0.0000
X_6	-15.70	1.75	-8.93	p=0.0000

Tabla 6.12

El tanto por ciento de variaciones controladas es del 99.09% y el coeficiente de correlación múltiple es altamente significativo. Sin embargo. los estimadores están muy alejados de los valores reales. que recordemos. son los siguientes:

$$Y = 1350 - 3X_1 + 12X_2 - 20X_3 + 15X_4 + 25X_5 - 13X_6 + \varepsilon$$

El tratamiento de la colinealidad mediante el método de selección ascendente no ha resultado fructífero.

6.4.5.2 Método de eliminación descendente

Utilizando el método de eliminación descendente se obtuvieron los resultados siguientes:

En primer lugar comenzamos con el modelo completo, y analizamos cual es la primera variable que debe ser eliminada del mismo (ver tabla 6.13),

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	21789.6569			
X_1	-2.7580	1.8872	-1.4614	p=0.1510
X_2	130.8591	32.0335	4.0851	p=0.0002
X_3	-393.3484	102.6947	-3.8303	p=0.0004
X_4	-3.0357	4.9657	-0.6113	p=0.5441
X_5	23.4743	1.4740	15.9253	p=0.0000
X_6	-15.1239	1.7799	-8.4971	p=0.0000

Tabla 6.13: Parámetros del modelo de regresión con todas las variables.

El porcentaje de variaciones controladas fue del 99.91%.

En el primer paso la variable eliminada es la 4; *conviene destacar que era la primera que entraba en la selección ascendente; lo cual evidencia que el orden de entrada de las variables en ningún caso implica su grado de importancia en el modelo.*

El modelo con todas las variables excepto la cuarta, es el que se muestra a continuación:

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	17942.8551			
X_1	-2.8750	1.8641	-1.5424	p=0.1300
X_2	111.3014	1.6131	68.9997	p=0.0000
X_3	-330.5757	1.5204	-217.4230	p=0.0000
X_5	23.5596	1.4569	16.1707	p=0.0000
X_6	-15.0869	1.7662	-8.5422	p=0.0000

Tabla 6.14: Parámetros del modelo de regresión.

En el paso número 2 se elimina la variable 1. La prueba de significación y los parámetros para el modelo de regresión resultante son:

Var	Coeficiente	Error Est.	Estadístico t	Significac.
Corte	17863,4888			
X_2	111,1179	1,6331	68,0394	p=0.0000
X_3	-330,9589	1,5228	-217,3356	p=0.0000
X_5	23,3354	1,4717	15,8560	p=0.0000
X_6	-15,6846	1,7493	-8,9660	p=0.0000

Tabla 6.15: Parámetros del modelo de regresión.

Ya nos salen más variables, por lo que le modelo final es:

$$Y= 17863,4888 + 111,1179\ X_2 - 330,9589\ X_3 + 23,3354\ X_5 - 15,6846\ X_5$$

Como puede observarse las variables implicadas en la colinealidad no desaparecen del modelo y los estimadores siguen siendo muy diferentes. aunque si tienen el mismo signo que los verdaderos coeficientes del modelo.

El tanto por ciento de variaciones controladas también en este caso supera el 99%. pero el tratamiento de la colinealidad no es bueno.