Regresión lineal múltiple

Josep Gibergans Bàguena

P08/75057/02312

Universitat Oberta de Catalunya
www.uoc.edu

Índice

El modelo de regresión múltiple

1. Introducción

La regresión lineal simple nos proporciona un modelo para explicar la relación entre dos variables: la variable *Y*, que llamamos *variable dependiente* o *explicada*, y la variable *X*, que recibe el nombre de *variable independiente* o *explicativa*.

El peso no sólo depende de la altura

Sabemos que el peso (*Y*) está relacionado linealmente con la altura (X_1). Pero también sabemos que puede estar relacionado con la edad (X_2), el número semanal de horas de deporte (X_3), la cantidad de calorías totales de las comidas (X_4), etc.

En este módulo tendremos en cuenta que en la realidad casi siempre son más de uno los factores o variables que influyen en los valores de otra variable y definiremos un nuevo modelo.

Ejemplos de variables afectadas por más de una variable

El precio de un ordenador depende de la velocidad del procesador, de la capacidad del disco duro, de la cantidad de memoria RAM, etc.

El sueldo de un titulado por la UOC depende de la edad, de los años que hace que acabó los estudios, de los años de experiencia en la empresa, etc.

El precio de alquiler de un piso depende de los metros cuadrados de superficie, de la edad de la finca, de la proximidad al centro de la ciudad, etc.

El precio de un coche depende de la potencia del motor, del número de puertas y de multitud de accesorios que puede llevar: *air bag*, ordenador de viaje, equipo de alta fidelidad, volante deportivo, llantas especiales, etc.

El planteamiento del problema es idéntico al realizado para la regresión simple, con la diferencia de que en este caso no tendremos una variable explicativa, sino varias. En este caso, será muy útil la notación matricial.

2. El modelo de regresión lineal múltiple

El **modelo de regresión lineal múltiple** es una generalización del modelo de regresión lineal simple, en el que relacionamos la variable que queremos explicar, *Y*, con las *k* variables explicativas X_1, X_2, ..., X_k. Lo encontraremos a partir de los valores (x_i, y_i) que toman estas variables sobre los elementos de una muestra y mediante la expresión siguiente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

Notación

La variable *Y* se denomina *variable dependiente* o *explicada*. Las variables X_i reciben el nombre de *variables independientes* o *explicativas*.

Expresando esta ecuación para cada observación de la muestra, obtenemos el sistema de ecuaciones siguiente:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1\, x_{11} + \beta_2\, x_{21} + \ldots + \beta_k\, x_{k1} + e_1 \\
y_2 &= \beta_0 + \beta_1\, x_{12} + \beta_2\, x_{22} + \ldots + \beta_k\, x_{k2} + e_2 \\
&\ldots \quad \ldots \quad \ldots \quad \ldots \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \ldots \\
y_n &= \beta_0 + \beta_1\, x_{1n} + \beta_2\, x_{2n} + \ldots + \beta_k\, x_{kn} + e_n
\end{aligned}$$

Podemos representar este sistema de forma matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

De manera que podemos escribir el modelo de la forma siguiente:

$$y = X\beta + e$$

donde:

- y: es el vector $(n \times 1)$ de observaciones de la variable Y.
- X: es la matriz $n \times (k + 1)$ de observaciones. A partir de la segunda columna, cada columna x_i tiene las observaciones correspondientes a cada una de las variables que consideremos.
- β : es el vector $(k + 1) \times 1$ de los coeficientes de la regresión.
- e: es el vector $(n \times 1)$ de los residuos o errores.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Supongamos que estamos interesados en explicar los gastos (en decenas de euros/año) de los ordenadores de un departamento comercial a partir de su edad (en años) y del número de horas diarias que trabajan (horas/día).

Hemos tomado una muestra de cinco ordenadores y hemos obtenido los resultados siguientes:

Gastos (Y) (decenas de euros/año)	Antigüedad (X_1) (años)	Horas de trabajo (X_2) (horas/día)
24,6	1	11
33,0	3	13
36,6	4	13
39,8	4	14
28,6	2	12

Queremos encontrar un modelo de regresión de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Si desarrollamos esta ecuación en todas las observaciones de la muestra, obtenemos el sistema de ecuaciones siguiente:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 + 11\ \beta_2 + \dots + e_1 \\ y_2 &= \beta_0 + 3\ \beta_1 + 13\ \beta_2 + \dots + e_2 \\ y_3 &= \beta_0 + 4\ \beta_1 + 13\ \beta_2 + \dots + e_3 \\ y_4 &= \beta_0 + 4\ \beta_1 + 14\ \beta_2 + \dots + e_4 \\ y_5 &= \beta_0 + 2\ \beta_1 + 12\ \beta_2 + \dots + e_5 \end{aligned}$$

Que podemos escribir matricialmente como $y = X\,\beta + e$, donde:

$$y = \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}60 \\ 39{,}80 \\ 28{,}60 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

En el modelo de regresión lineal múltiple, que hemos expresado matricialmente como:

$$y = X\beta + e$$

- $X\,\beta$ es la parte correspondiente a la variación de y que queda explicada por las variables X_i;
- e es un término que llamamos de los *residuos* o *errores* y que de alguna manera recoge el efecto de todas aquellas variables que también afectan a y y que no se encuentran incluidas en el modelo porque son desconocidas o porque no se tienen datos suyos. Sobre este término haremos dos suposiciones importantes:

1. Los errores se distribuyen según una distribución normal de media cero y una varianza σ^2.

2. Los errores son independientes.

Con estas dos suposiciones tenemos dos consecuencias importantes:

1. Fijando unos valores $x_1, x_2, ..., x_k$ de las variables $X_1, X_2, ..., X_k$ y tomando valores esperados sobre la ecuación del modelo, tenemos que:

$$E(Y|x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Recordemos que...

... en el modelo de regresión lineal simple la recta de regresión pasa por $(x_i, E(y))$.

2. Del mismo modo, la varianza de la distribución de Y es constante:

$$Var(Y|x_1, x_2, ..., x_k) = \sigma^2$$

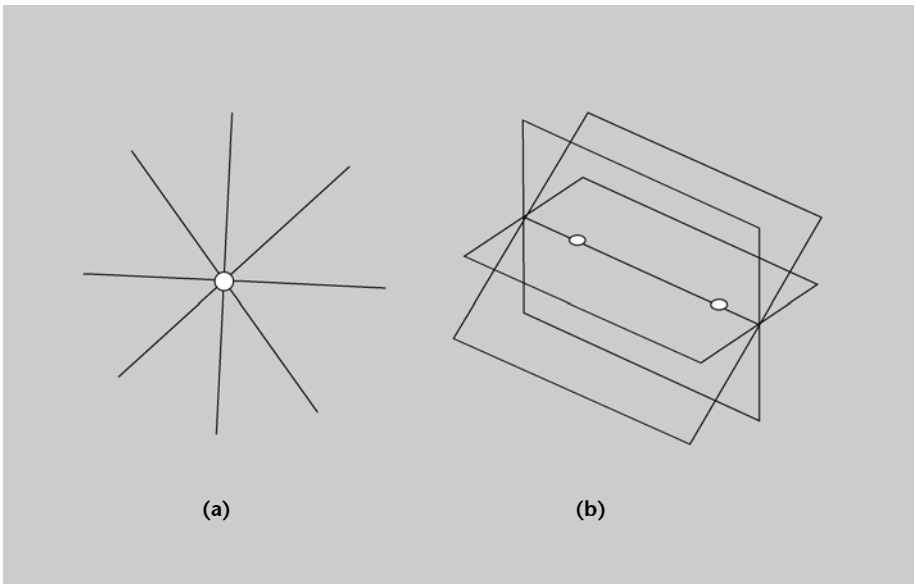
Añadiremos un par de suposiciones adicionales sobre el modelo:

1. No podemos tener más parámetros por estimar $(k + 1)$ que datos disponibles (n) y, por tanto, $n > k + 1$.

En el caso del modelo lineal simple resulta claro que si tenemos más parámetros que datos, tenemos un único dato. Es imposible encontrar cuál es la recta que mejor se ajusta a un único punto, ya que tenemos infinitas rectas que pasan por este punto.

Se podría aplicar este mismo razonamiento si tuviéramos más variables explicativas, aunque sería difícil de visualizar.

En el caso del modelo lineal múltiple, en el que tenemos dos variables explicativas, el número de parámetros que hay que estimar es tres. Si resulta que tenemos dos o menos datos, es decir, como mucho dos puntos, tampoco tiene sentido buscar un modelo de regresión, ya que tenemos un número infinito de planos que pasan por dos puntos fijados.



Leyenda

a) Modelo de regresión lineal simple con una observación.

b) Modelo de regresión múltiple con dos variables explicativas y dos observaciones.

2. Ninguna de las variables explicativas puede ser combinación lineal de las otras, ya que en este caso no tendríamos un modelo de *k* variables, sino de *k* – 1 variables (queremos que las variables X_i sean independientes):

Por ejemplo, si: $X_2 = a + b\ X_1$, entonces:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e = \\ &= \beta_0 + \beta_1 x_1 + \beta_2 (a + b x_1) + \ldots + \beta_k x_k + e = \\ &= (\beta_0 + a) + (\beta_1 + b) x_1 + \beta_3 x_3 + \ldots + \beta_k x_k + e = \\ &= \beta'_0 + \beta'_1 x_1 + \beta_3 x_3 + \ldots + \beta_k x_k + e \end{aligned}$$

Tenemos sólo *k* – 1 variables.

3. Ajuste del modelo: método de los mínimos cuadrados

Para determinar los parámetros de la recta de regresión en el modelo lineal simple, utilizamos el método de los mínimos cuadrados. Este método consiste en encontrar la recta que hace mínima la suma de los residuos al cuadrado.

Residuo en el modelo de regresión lineal simple

En el modelo de regresión lineal simple el residuo es la diferencia entre el valor observado de la variable *Y* y el valor estimado sobre una recta.

En el caso que ahora nos ocupa, procederemos de una forma muy similar. Buscaremos la suma de los residuos al cuadrado y después determinaremos los parámetros del modelo que hacen que esta suma tenga un valor mínimo.

Definiremos los residuos como la diferencia entre los valores observados en la muestra (y_i) y los valores estimados por el modelo ($\hat{y}_i$):

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

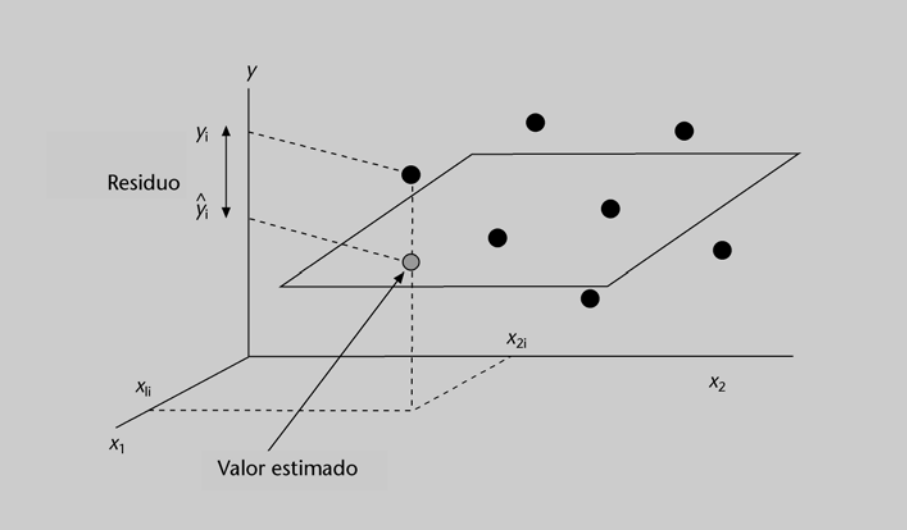
donde x_{1i} y x_{2i} son dos observaciones de las variables X_1 y X_2, respectivamente.

Si consideramos un modelo de regresión lineal múltiple con dos variables explicativas X_1 y X_2, los residuos vendrán dados por:

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$

Geométricamente, podemos interpretarlo como la diferencia entre el valor observado y el valor estimado sobre un plano. Los parámetros del modelo se determinan encontrando el plano que hace mínima la suma de los residuos al cuadrado. Este plano se conoce como *plano de regresión por mínimos cuadrados*.

Representamos el residuo para un modelo de regresión múltiple con dos variables explicativas.



En un modelo de regresión múltiple con *k* variables explicativas tenemos la siguiente expresión para los residuos:

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) \text{ para } i = 1, 2, \ldots, n$$

que matricialmente podemos escribir:

$$\begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \ldots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix}$$

$$e = y - \hat{y} = y - X\beta$$

donde e es el vector de los residuos, $\hat{y}$ es el vector de las estimaciones de y y β es el vector de los parámetros de la regresión.

Para calcular la suma de los cuadrados de los elementos de un vector, hay que hacer el producto escalar del vector por sí mismo, o lo que es lo mismo, el producto matricial del vector transpuesto por el mismo vector.

Si lo hacemos con el vector de los residuos e:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - \hat{y}_i)^2 = (y - X\beta)^t (y - X\beta)$$

Haciendo ahora los productos y utilizando algunas propiedades del cálculo matricial, obtenemos la suma de los cuadrados de los residuos:

$$\sum_{i=1}^{n} e_1^2 = (y - X\beta)^t (y - X\beta) = y^t y - 2\beta X^t y + \beta^t X^t X\beta$$

Para encontrar los valores de los parámetros que hacen mínima esta suma, debemos derivar parcialmente con respecto a los parámetros:

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} e_1^2 \right) = -2X^t y + 2X^t X\beta$$

Y encontrar aquellos valores que hacen nulas estas derivadas parciales:

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} {e_1}^2 \right) = 0 \Rightarrow -2\, X^t y + 2X^t X\beta = 0$$

Notación

Los estimadores de los parámetros de la regresión que buscamos son las soluciones de esta ecuación matricial, así que ponemos el "sombrero", que nos indica que se trata de estimadores.

Simplificando un poco, tenemos $X^t X\hat{\beta} = X^t y$

Podemos aislar el vector de parámetros incógnita:

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

El vector $\hat{\beta}$ es el vector de los estimadores mínimos cuadráticos de los parámetros.

Finalmente, sólo queda por comentar que, si en la ecuación $X^tX\hat{\beta} = X^ty$ efectuamos la multiplicación matricial, obtenemos el sistema de ecuaciones siguiente, llamado *sistema de ecuaciones normales de la regresión*:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & .... & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}x_{i1} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & ... & \sum_{i=1}^{n} x_{ik}x_{i2} \\ ... & .... & .... & .... & .... \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ ... \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Tenemos:

$$y = \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}60 \\ 39{,}80 \\ 28{,}60 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix}$$

La matriz transpuesta de la matriz X es:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix}$$

De manera que:

$$X^tX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix} \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} = \begin{bmatrix} 5 & 14 & 63 \\ 14 & 46 & 182 \\ 63 & 182 & 799 \end{bmatrix}$$

Si calculamos la inversa de esta matriz:

$$(X^tX)^{-1} = \begin{bmatrix} 4 & 14 & 63 \\ 14 & 46 & 182 \\ 63 & 182 & 799 \end{bmatrix}^{-1} = \begin{bmatrix} 181{,}5 & 14 & -17{,}5 \\ 14 & 1{,}3 & -1{,}4 \\ -17{,}5 & -1{,}4 & 1{,}7 \end{bmatrix}$$

Por otro lado, tenemos:

$$X^ty = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix} \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}6 \\ 39{,}8 \\ 28{,}6 \end{bmatrix} = \begin{bmatrix} 162{,}60 \\ 486{,}40 \\ 2075{,}80 \end{bmatrix}$$

Y el vector de los parámetros estimados de la regresión es:

$$\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1}(X^t y) = \begin{bmatrix} 181,5 & 14 & -17,5 \\ 14 & 13 & -1,4 \\ -17,5 & -1,4 & 1,7 \end{bmatrix} \begin{bmatrix} 162,60 \\ 486,40 \\ 2075,80 \end{bmatrix} = \begin{bmatrix} -5 \\ 2,6 \\ 2,4 \end{bmatrix}$$

La ecuación de regresión es, pues:

$$\hat{y} = -5,0 + 2,6x_1 + 2,4x_2$$

4. Interpretación de los parámetros

De la misma manera que en la regresión lineal, una vez obtenido el modelo de regresión lineal múltiple, es muy importante hacer una buena interpretación de los resultados obtenidos. De momento, sólo hemos obtenido los parámetros estimados del modelo de regresión:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Para interpretarlos correctamente, debemos tener presente el fenómeno que estudiamos.

1. Interpretación de $\hat{\beta}_0$:

Este parámetro representa la estimación del valor de Y cuando todas las X_j toman valor cero. No siempre tiene una interpretación vinculada al contexto (geométrica, física, económica, etc.). Para que sea posible interpretarlo, necesitamos lo siguiente:

a. Que sea realmente posible que las $X_j = 0$.
b. Que se tengan suficientes observaciones cerca de los valores $X_j = 0$.

2. Interpretación de $\hat{\beta}_j$:

Representa la estimación del incremento que experimenta la variable Y cuando X_j aumenta su valor en una unidad y las demás variables se mantienen constantes.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Continuando con el ejemplo de los ordenadores y a partir de los resultados obtenidos en el ajuste:

$$\hat{y} = -5,0 + 2,6x_1 + 2,4x_2$$

1. $\hat{\beta}_0 = -5$ (por decenas de euros)

Nos indica los gastos en decenas de euros de un ordenador con cero años de antigüedad y cero horas semanales de trabajo. Es evidente que este ejemplo no tiene nigún sentido.

2. $\hat{\beta}_1 = 2,6$ (por decenas de euros/año de antigüedad)

Nos indica el incremento de los gastos en decenas de euros por cada año de antigüedad del ordenador, sin tener en cuenta el número de horas diarias de uso. Así pues, por cada año que pase, tendremos $2,6 \cdot 10 = 26$ euros más en los gastos de mantenimiento de un ordenador.

3. $\hat{\beta}_2 = 2,4$ (en decenas de euros/horas diarias de trabajo)

Nos indica el incremento en los gastos en decenas de euros por cada hora diaria de uso sin tener en cuenta la antigüedad del ordenador. Tenemos que por cada hora de más de trabajo, tendremos $2,4 \cdot 10 = 24$ euros más en los gastos anuales de mantenimiento de un ordenador.

5. Resumen

En esta sesión se ha presentado el modelo de regresión lineal múltiple como una generalización del modelo de regresión lineal simple en aquellos casos en los que se tiene más de una variable explicativa. Hemos visto la manera de buscar los parámetros del modelo por el método de los mínimos cuadrados, así como la comodidad que puede suponer el uso de la notación matricial a la hora de expresar y realizar los cálculos.

Ejercicios

1. Los datos siguientes se han obtenido experimentalmente para determinar la relación entre la ganancia de corriente (y), el tiempo de difusión (x_1) y la resistencia (x_2) en la fabricación de un determinado tipo de transistor:

Y	5,3	7,8	7,4	9,8	10,8	9,1	8,1	7,2	6,5	12,6
X_1 (horas)	1,5	2,5	0,5	1,2	2,6	0,3	2,4	2,0	0,7	1,6
X_2 (ohmios-cm)	66	87	69	141	93	105	111	78	66	123

Os pedimos lo siguiente:

a) Especificad un modelo lineal múltiple para expresar la ganancia de corriente en términos del tiempo de difusión y de la resistencia.

b) Estimad los parámetros del modelo de regresión lineal múltiple.

2. Se realiza un experimento para ver si es posible determinar el peso de un animal después de un periodo de tiempo determinado a partir de su peso inicial y de la cantidad de alimento que se le suministra. A partir los resultados obtenidos para una muestra de $n = 10$:

$$\sum_{i=1}^{10} x_{i1} = 379\,,\ \sum_{i=1}^{10} x_{i2} = 2.417\,,\ \sum_{i=1}^{10} x_{i1}^2 = 14.533\,,\ \sum_{i=1}^{10} x_{i2}^2 = 601.365$$

$$\sum_{i=1}^{10} x_{i1}x_{i2} = 92.628\,,\ \sum_{i=1}^{10} y_i = 825\,,\ \sum_{i=1}^{10} x_{i1}y_i = 31.726\,,\ \sum_{i=1}^{10} x_{i2}y_i = 204.569$$

Encontrad la ecuación del modelo de regresión lineal múltiple correspondiente.

Solucionario

1.

a) Ahora tenemos:

Número de observaciones: $n = 10$

Número de variables independientes: 2

Número de parámetros: $k = 2 + 1 = 3$

El modelo lineal múltiple : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

$$\begin{bmatrix} 5,3 \\ 7,8 \\ 7,4 \\ 9,8 \\ 10,8 \\ 9,1 \\ 8,1 \\ 7,2 \\ 6,5 \\ 12,6 \end{bmatrix} = \begin{bmatrix} 1 & 1,5 & 66 \\ 1 & 2,5 & 87 \\ 1 & 0,5 & 69 \\ 1 & 1,2 & 141 \\ 1 & 2,6 & 93 \\ 1 & 0,3 & 105 \\ 1 & 2,4 & 111 \\ 1 & 2,0 & 78 \\ 1 & 0,7 & 66 \\ 1 & 1,6 & 123 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix}$$

$$Y = \quad X \quad \beta + e$$

b) Estimaremos los parámetros mediante el método de los mínimos cuadrados:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y$$

donde $(X^t X)^{-1}$ es la matriz inversa de la matriz $(X^t X)$:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,5 & 2,5 & 0,5 & 1,2 & 2,6 & 0,3 & 2,4 & 2,0 & 0,7 & 1,6 \\ 66 & 87 & 69 & 141 & 93 & 105 & 111 & 78 & 66 & 123 \end{bmatrix}$$

Atención

Según el número de cifras decimales que cojáis a partir de aquí, los resultados pueden ser un poco diferentes, sin que esto signifique que sean incorrectos.

$$X^t Y = \begin{bmatrix} 84,6 \\ 132,27 \\ 8.320,2 \end{bmatrix}; \qquad X^t X = \begin{bmatrix} 10 & 15,3 & 939 \\ 15,3 & 29,85 & 1.458,9 \\ 939 & 1.458,9 & 94.131 \end{bmatrix}$$

$$(X^t X)^{-1} \begin{bmatrix} 1,7985570396 & -1,855438825 & -0,01506576037 \\ -0,1855438825 & 0,1572804381 & -0,0005867432141 \\ -0,01506576037 & -0,0005867432141 & 0,0001700050851 \end{bmatrix}$$

Ya podemos calcular los coeficientes:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y = \begin{bmatrix} 2,2680510 \\ 0,2224947452 \\ 0,062325502 \end{bmatrix}$$

Obtenemos:

$$\hat{\beta}_0 = 2,2680510\,,\ \hat{\beta}_1 = 0,224947452\,,\ \hat{\beta}_2 = 0,062325502$$

El modelo de regresión lineal múltiple obtenido es:

$$\hat{y} = 2{,}2680510 + 0{,}224947452x_1 + 0{,}062325502x_2$$

2. A partir de las ecuaciones normales de la regresión múltiple:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & .... & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}x_{i1} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & ... & \sum_{i=1}^{n} x_{ik}x_{i2} \\ ... & .... & .... & .... & .... \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ ... \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

$$\begin{bmatrix} 10 & 379 & 2417 \\ 379 & 14.533 & 92.628 \\ 2.417 & 92.628 & 601.365 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 825 \\ 31.726 \\ 204.569 \end{bmatrix}$$

$$X^t X\hat{\beta} = X^t y$$

Aislando el vector de parámetros estimados: $\hat{\beta} = (X^t X)^{-1} X^t y$

Primero debemos calcular la matriz inversa:

$$(X^t X)^{-1} = \begin{bmatrix} 8{,}6176 & -0{,}21777 & -0{,}0010927 \\ -0{,}21777 & 0{,}0092689 & -0{,}00055243 \\ -0{,}0010927 & -0{,}00055243 & 0{,}000091145 \end{bmatrix}$$

Finalmente, tenemos que:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 8{,}6176 & -0{,}21777 & -0{,}0010927 \\ -0{,}21777 & 0{,}0092689 & -0{,}00055243 \\ -0{,}0010927 & -0{,}00055243 & 0{,}000091145 \end{bmatrix} \begin{bmatrix} 825 \\ 31.726 \\ 204.569 \end{bmatrix} =$$

$$= \begin{bmatrix} -22{,}984 \\ 1{,}395 \\ 0{,}218 \end{bmatrix}$$

El modelo de regresión lineal múltiple que obtenemos es:

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

La calidad del ajuste

1. Introducción

Una vez encontrado el modelo de regresión lineal múltiple a partir de los datos de una muestra, queremos utilizarlo para hacer inferencias a toda la población. Sin embargo, antes es necesario llevar a cabo una comprobación de la idoneidad del modelo obtenido.

En esta sesión estudiaremos el coeficiente de determinación para la regresión múltiple como indicador de la calidad del ajuste. También utilizaremos los gráficos de los residuos como una importante herramienta de diagnóstico del modelo.

2. Calidad del ajuste. El coeficiente de determinación R^2

De la misma manera que en la regresión lineal simple, también podemos definir ahora el **coeficiente de determinación R^2** como la proporción de variabilidad explicada por el modelo con respecto a la variabilidad total, es decir:

$$= \frac{\text{Variabilidad explicada por el?modelo}}{\text{Variabilidad total de la muestra}}$$

Terminología

R también se conoce como *coeficiente de correlación múltiple*.

Si consideramos que la varianza total observada en la variable *Y* se descompone en dos términos, la varianza explicada por el modelo de regresión lineal más la varianza que no queda explicada por el modelo, es decir, la varianza de los residuos:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

podemos expresar el coeficiente de determinación así:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Y a partir de las fórmulas de las varianzas:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}e_1^2$$

$$s_y^2 = \frac{SQT}{n-1} \qquad s_{\hat{y}}^2 = \frac{SQR}{n-1} \qquad s_e^2 = \frac{SQE}{n-1}$$

donde:

$SCT = \sum_{i=1}^{n} (y_i - \bar{y}_2)$ **S**uma de **C**uadrados **T**otales

$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_2)^2$ **S**uma de **C**uadrados de la **R**egresión

$SCE = \sum_{i=1}^{n} e_1^2$ **S**uma de **C**uadrados de los **E**rrores

Se puede demostrar que: $SCT = SCR + SCE$.

Y teniendo en cuenta que hemos definido el coeficiente de determinación como $R^2 = s_{\hat{y}}^2 / s_y^2$, finalmente podemos escribirlo como:

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

Para calcular las sumas de cuadrados, podemos utilizar el cálculo matricial.

- **Suma de los cuadrados totales**

Siendo D el vector de desviaciones de las y_i con respecto a la media $\bar{y}$:

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix}$$

Podemos escribir la suma de los cuadrados totales de la forma siguiente:

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

- **Suma de los cuadrados de la regresión:**

A partir de los valores estimados:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix}$$

podemos calcular el vector de las desviaciones de los valores estimados $\hat{y}_i$ con respecto a la media $\bar{y}$:

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \dots \\ \hat{y}_n - \bar{y} \end{bmatrix}$$

y, por tanto, $SCR = \sum_{i=1}^{n} (\hat{y}_1 - \bar{y})^2 = (w)^t w$.

- **Suma de los cuadrados de los errores**

A partir de los residuos:

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \dots \\ y_n - \hat{y}_n \end{bmatrix}$$

es fácil calcular la suma de sus cuadrados:

$$SCE = \sum_{i=1}^{n} e_1^2 = e^t e$$

De la misma manera que en la regresión lineal simple, tenemos que el valor del coeficiente de determinación está siempre entre 0 y 1: $0 \leq R^2 \leq 1$.

1. $R^2 = 1$ se tiene cuando $SCT = SCR$, es decir, cuando toda la variabilidad de Y se explica por el modelo de regresión. En este caso tenemos que los valores estimados por el modelo son exactamente iguales a los observados.

2. $R^2 = 0$ se tiene cuando $SCR = 0$, es decir, cuando el modelo no explica absolutamente nada de Y.

3. Cuanto mayor sea R^2, mayor será la proporción de variabilidad de Y explicada por el modelo y, por tanto, mayor será la bondad del ajuste.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Consideremos de nuevo el ejemplo de los gastos anuales en el mantenimiento de un ordenador. Teníamos que $\overline{y} = 32{,}52$, de manera que la suma de cuadrados totales vale:

$$SCT = \sum_{i=1}^{n} (y_1 - \bar{y})^2 = D^t D = (y_1 - \bar{y}\ y_2 - \bar{y}\ \dots y_n - \bar{y}) \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix} =$$

$$= (-7{,}92 \;\; 0{,}48 \;\; 4{,}08 \;\; 7{,}28 \;\; -3{,}92) \begin{bmatrix} -7{,}92 \\ 0{,}48 \\ 4{,}08 \\ 7{,}28 \\ -3{,}98 \end{bmatrix} = 147{,}97$$

Los valores estimados por el modelo de regresión múltiple son:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = X\beta = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} -5 \\ 2{,}6 \\ 2{,}4 \end{bmatrix} = \begin{bmatrix} 24 \\ 34 \\ 36{,}6 \\ 39 \\ 29 \end{bmatrix}$$

De manera que la suma de cuadrados de la regresión es:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (w)^t w = (\hat{y}_1 - \bar{y} \;\; \hat{y}_2 - \bar{y} \;\; \ldots \hat{y}_n - \bar{y}) \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \ldots \\ \hat{y}_n - \bar{y} \end{bmatrix} =$$

$$= (-8{,}52 \;\; 1{,}48 \;\; 4{,}08 \;\; 6{,}48 \;\; -3{,}52) \begin{bmatrix} -8{,}52 \\ 1{,}48 \\ 4{,}08 \\ 6{,}48 \\ -3{,}52 \end{bmatrix} = 145{,}81$$

La diferencia entre los valores observados y los valores estimados nos permite obtener los residuos:

$$e = \begin{bmatrix} 24{,}6 \\ 33 \\ 36{,}6 \\ 39{,}8 \\ 28{,}6 \end{bmatrix} - \begin{bmatrix} 24 \\ 34 \\ 36{,}6 \\ 39 \\ 29 \end{bmatrix} = \begin{bmatrix} 0{,}6 \\ -1 \\ 0 \\ 0{,}8 \\ -0{,}4 \end{bmatrix}$$

Así, la suma de los cuadrados de los residuos es:

$$SCE = \sum_{i=1}^{n} e_1^2 = e^t e = (e_1 \;\; e_1 \;\; \ldots \;\; e_n) \begin{bmatrix} e_1 \\ e_1 \\ \ldots \\ e_n \end{bmatrix} = (0{,}6 \;\; -1 \;\; 0 \;\; 0{,}8 \;\; -0{,}4) \begin{bmatrix} 0{,}6 \\ -1 \\ 0 \\ 0{,}8 \\ -0{,}4 \end{bmatrix} = 2{,}16$$

El coeficiente de determinación es:

$$R^2 = \frac{SCR}{SCT} = \frac{145{,}81}{147{,}97} = 0{,}985$$

También se puede calcular haciendo: $R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{2.16}{147{,}97} = 1 - 0{,}0015 = 0{,}985$

Este resultado nos dice que el modelo de regresión múltiple obtenido explica el 98,5% de la variabilidad de los gastos de los ordenadores. Dado que está muy cerca del 100%, en principio es un buen modelo.

3. El análisis de los residuos

De la misma manera que en la regresión lineal simple, los residuos del modelo de regresión lineal múltiple tienen un papel importante a la hora de determinar la adecuación del modelo.

En el caso de regresión lineal múltiple es habitual construir dos tipos de gráficos:

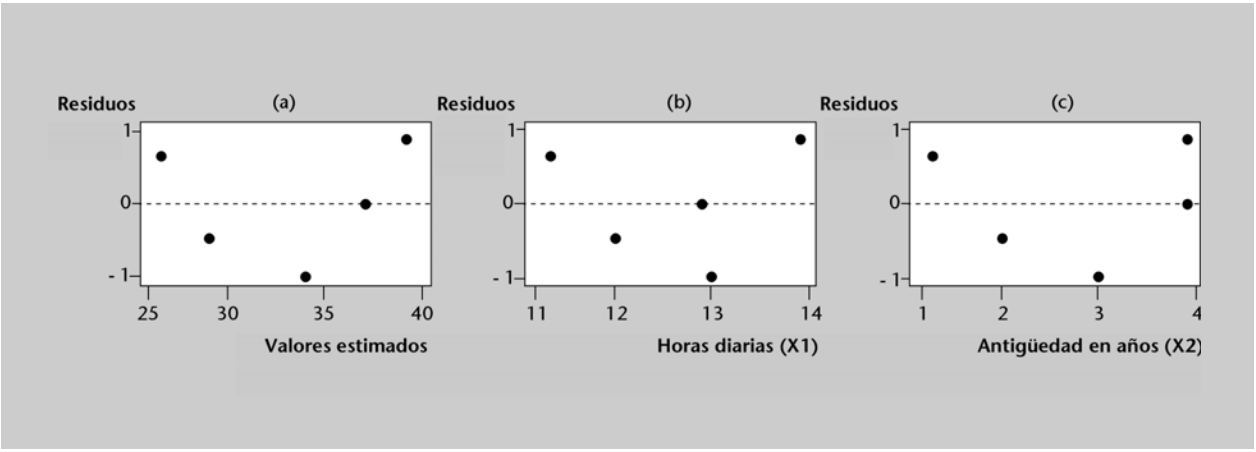
1. Gráfico de *residuos frente a valores estimados*: representamos en el eje de ordenadas los valores de los residuos y en el eje de abscisas, los valores estimados, de manera que la nube de puntos $(\hat{y}_i, e_i)$ no debe tener ningún tipo de estructura y es cercano al eje de abscisas.

2. Gráfico de *residuos frente a variables explicativas*: representamos sobre el eje de ordenadas los valores de los residuos y sobre el eje de abscisas, los valores observados de la variable explicativa. Tenemos un gráfico de este tipo para cada una de las variables explicativas.

Siempre que el modelo sea correcto, ningún gráfico de residuos debe mostrar ningún tipo de estructura. Los residuos siempre deben estar distribuidos al azar alrededor del cero.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

En el caso de los ordenadores y sus gastos en mantenimiento, tenemos los gráficos de representación de los residuos siguientes:



Los tres gráficos representan:

a) residuos frente a valores estimados por el modelo;

b) residuos frente a valores de la variable X_1: horas diarias de trabajo;

c) residuos frente a valores de la variable X_2: antigüedad de los ordenadores en años.

No observamos ningún tipo de estructura organizada de los residuos que nos haga pensar en una falta de linealidad del modelo. Tampoco observamos ningún dato atípico.

4. Aplicaciones a la predicción

La aplicación básica de un modelo de regresión lineal múltiple es predecir (estimar) el valor de la variable Y a partir de un conjuntos de valores de las variables independientes X_j .

Sólo hay que sustituir estos valores x_i en la ecuación de regresión obtenida:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

Ejemplo de los gastos de los ordenadores según su antigüeded y las horas diarias de trabajo

Considerando una vez más el problema de los ordenadores, si queremos calcular el gasto correspondiente a un ordenador que tiene dos años de antigüedad y trabaja catorce horas diarias, utilizaremos la ecuación encontrada:

$$\hat{y} = -5{,}0 + 2{,}6x_1 + 2{,}4x_2$$

con $x_1 = 2$ y $x_2 = 14$:

$$\hat{y} = -5{,}0 + 2{,}6 \cdot 2 + 2{,}4 \cdot 14 = -5{,}0 + 5{,}4 + 33{,}6 = 34$$

Por tanto, podemos esperar un gasto de mantenimiento de 340 euros anuales para este ordenador.

A la hora de aplicar la ecuación de regresión encontrada, siempre debemos mirar si los valores de las variables X_i para los que queremos estimar el valor de la variable Y se encuentran dentro del conjunto de valores que hemos utilizado para construir el modelo. Si no es así, debemos ir con mucha cautela, ya que puede ser que el resultado que nos dé el modelo no tenga ningún sentido. El peligro de la extrapolación también está presente en la regresión lineal múltiple.

Ejemplo de resultado irreal

Si queremos utilizar nuestro modelo para calcular el gasto de mantenimiento de nuestro ordenador cuando tenga una antigüedad de cincuenta años, es evidente que no tiene ningún sentido utilizar la ecuación encontrada: ni el ordenador existirá de aquí a cincuenta años (y si existe estará en un museo), ni los precios de mantenimiento tendrán nada que ver con los de ahora, etc.

5. Resumen

En esta sesió hemos estudiado el coeficiente de determinación como una medida de la bondad del ajuste del modelo a los datos de la muestra. A continuación se ha comentado la importancia de efectuar un análisis de los residuos para tener un diagnóstico del modelo lineal obtenido. Hemos acabado la sesión con la aplicación de la regresión a la predicción, que pone de manifiesto el peligro de la extrapolación.

Ejercicios

1. Los datos siguientes se han obtenido de forma experimental para determinar la relación entre la ganancia de corriente (Y), el tiempo de difusión (X_1) y la resistencia (X_2) en la fabricación de un determinado tipo de transistor:

Y	5,3	7,8	7,4	9,8	10,8	9,1	8,1	7,2	6,5	12,6
X_1 (horas)	1,5	2,5	0,5	1,2	2,6	0,3	2,4	2,0	0,7	1,6
X_2 (ohmios-cm)	66	87	69	141	93	105	111	78	66	123

Si el modelo de regresión obtenido a partir de estos datos es:

$$\hat{y} = 2{,}268 + 0{,}225x_1 + 0{,}062x_2$$

haced un análisis de los residuos y comentad los resultados obtenidos.

2. Se lleva a cabo un experimento para ver si es posible determinar el peso de un animal después de un periodo de tiempo determinado a partir de su peso inicial y de la cantidad de alimento que se le suministra. A partir de los resultados obtenidos para una muestra de $n = 10$:

Peso final (kg)	95	77	80	100	97	70	50	80	92	84
Peso inicial (kg)	42	33	33	45	39	36	32	41	40	38
Alimento (kg)	272	226	259	292	311	183	173	236	230	235

Se ha obtenido el modelo de regresión lineal:

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

Calculad el coeficiente de determinación e interpretadlo.

Solucionario

1. Para llevar a cabo un análisis de residuos, debemos construir dos tipos de gráficos:

- Gráfico de *residuos frente a valores estimados*: representaremos en el plano la nube de puntos: $(\hat{y}_1,\ e_i)$.

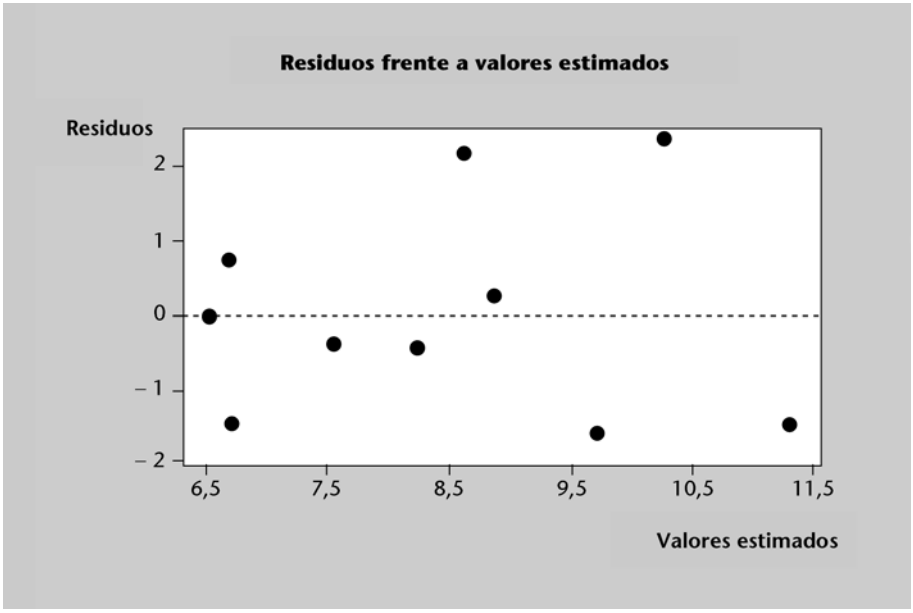
Antes deberemos calcular los valores estimados: $\hat{y} = X\beta$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \\ \hat{y}_6 \\ \hat{y}_7 \\ \hat{y}_8 \\ \hat{y}_9 \\ \hat{y}_{10} \end{bmatrix} = \begin{bmatrix} 1 & 1{,}5 & 66 \\ 1 & 2{,}5 & 87 \\ 1 & 0{,}5 & 69 \\ 1 & 1{,}2 & 141 \\ 1 & 2{,}6 & 93 \\ 1 & 0{,}3 & 105 \\ 1 & 2{,}4 & 111 \\ 1 & 2{,}0 & 78 \\ 1 & 0{,}7 & 66 \\ 1 & 1{,}6 & 123 \end{bmatrix} \begin{bmatrix} 2{,}268 \\ 0{,}225 \\ 0{,}062 \end{bmatrix} = \begin{bmatrix} 6{,}71 \\ 8{,}25 \\ 6{,}68 \\ 11{,}32 \\ 8{,}64 \\ 8{,}87 \\ 9{,}72 \\ 7{,}57 \\ 6{,}53 \\ 10{,}29 \end{bmatrix}$$

Y los residuos: $e = \hat{y} - y$

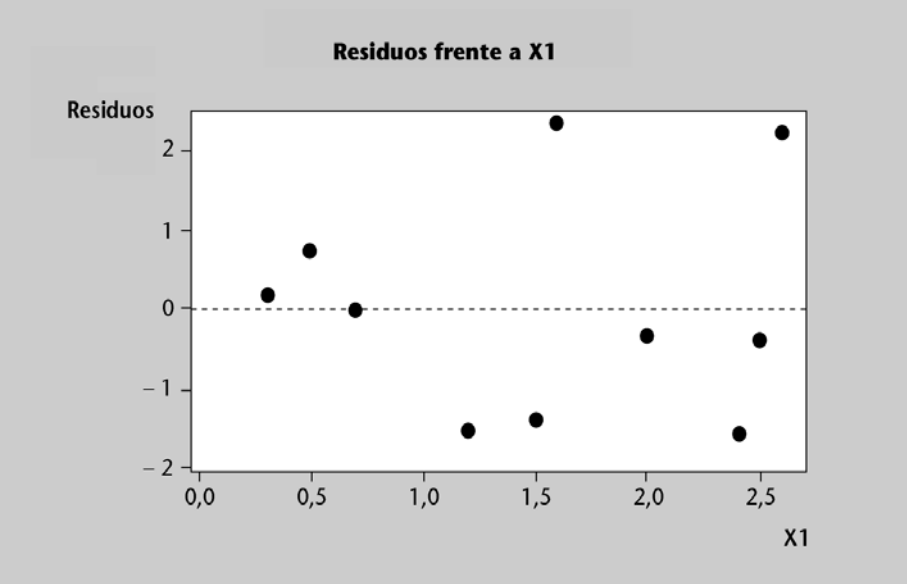
$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \\ \hat{y}_6 \\ \hat{y}_7 \\ \hat{y}_8 \\ \hat{y}_9 \\ \hat{y}_{10} \end{bmatrix} = \begin{bmatrix} 5{,}3 \\ 7{,}8 \\ 7{,}4 \\ 9{,}8 \\ 10{,}8 \\ 9{,}1 \\ 8{,}1 \\ 7{,}2 \\ 6{,}5 \\ 12{,}6 \end{bmatrix} - \begin{bmatrix} 6{,}71 \\ 8{,}25 \\ 6{,}68 \\ 11{,}32 \\ 8{,}64 \\ 8{,}87 \\ 9{,}72 \\ 7{,}57 \\ 6{,}53 \\ 10{,}29 \end{bmatrix} = \begin{bmatrix} -1{,}41 \\ -0{,}45 \\ 0{,}72 \\ -1{,}52 \\ 2{,}16 \\ 0{,}22 \\ -1{,}62 \\ -0{,}37 \\ -0{,}03 \\ 2{,}31 \end{bmatrix}$$

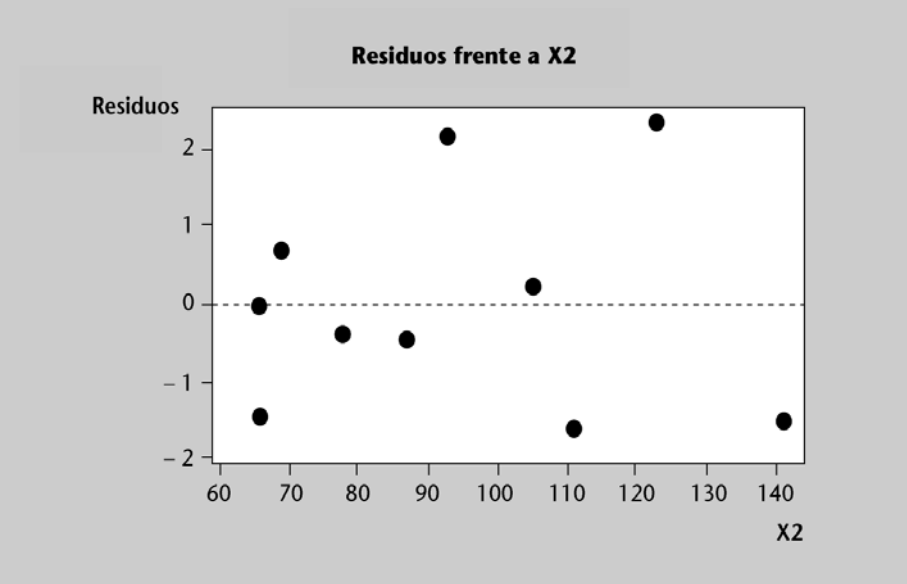
El gráfico resultante es:



No observamos ningún tipo de estructura en la nube de puntos.

- Gráficos de residuos frente a variables explicativas: ahora, por cada variable explicativa tenemos un gráfico. En este gráfico representamos (x_{ii}, e_i).





En ninguna de estas dos representaciones podemos ver ningún tipo de estructura en las nubes de puntos.

2. Podemos calcularlo a partir de cualquiera de las expresiones:

$$R^2 = \frac{SCR}{SCT} \qquad R^2 = 1 - \frac{SCE}{SCT} \qquad 0 \leq R^2 \leq 1$$

Deberemos tener en cuenta que, si lo calculamos de las dos formas, los resultados serán ligeramente diferentes a causa del error de redondeo asociado a los cálculos.

Para calcular la suma de cuadrados de la regresión (SCR), tenemos que conocer la media de y:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i = 82{,}5$$

Y los valores estimados de y_i, $\hat{y}_i$:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \\ \hat{y}_6 \\ \hat{y}_7 \\ \hat{y}_8 \\ \hat{y}_9 \\ \hat{y}_{10} \end{bmatrix} = \begin{bmatrix} 1 & 42 & 272 \\ 1 & 33 & 226 \\ 1 & 33 & 259 \\ 1 & 45 & 292 \\ 1 & 39 & 311 \\ 1 & 36 & 183 \\ 1 & 32 & 173 \\ 1 & 41 & 236 \\ 1 & 40 & 230 \\ 1 & 38 & 235 \end{bmatrix} \begin{bmatrix} -22{,}984 \\ 1{,}395 \\ 0{,}218 \end{bmatrix} = \begin{bmatrix} 94{,}778 \\ 72{,}216 \\ 79{,}396 \\ 103{,}314 \\ 99{,}079 \\ 67{,}0.45 \\ 59{,}290 \\ 85{,}550 \\ 82{,}850 \\ 81{,}148 \end{bmatrix}$$

Y para calcular la suma de cuadrados de los errores (SCE), necesitamos el vector de errores:

$$e = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \\ \hat{y}_6 \\ \hat{y}_7 \\ \hat{y}_8 \\ \hat{y}_9 \\ \hat{y}_{10} \end{bmatrix} = \begin{bmatrix} 0{,}222 \\ 4{,}784 \\ 0{,}604 \\ -3{,}314 \\ -2{,}079 \\ 2{,}955 \\ -9{,}290 \\ -5{,}550 \\ 9{,}150 \\ 2{,}852 \end{bmatrix}$$

Las sumas de cuadrados son:

- $SCT = \sum(y_i - \bar{y})^2 = 2.020{,}50$
- $SCR = \sum(\hat{y}_i - \bar{y})^2 = 1{,}762{,}99$
- $SCE = \sum(y_i - \hat{y})^2 = \sum e_i^2 = 256{,}30$

Por tanto, el coeficiente de determinación es $R^2 = \frac{SCR}{SCT} \approx 0{,}873$

Puesto que el coeficiente de determinación es la relación entre la varianza explicada y la varianza total, tenemos que es bastante cercano a 1; por tanto, significa que tenemos bondad en el ajuste. El modelo de regresión explica el 87,3% de la variabilidad del peso de los animales a partir de su peso inicial y la cantidad de alimento.

Inferencia en la regresión lineal múltiple

1. Introducción

Una vez estimado el modelo de regresión, estamos interesados en poder aplicarlo a la población de la que hemos sacado la muestra. Ahora determinaremos intervalos de confianza para los parámetros del modelo y haremos contrastes de hipótesis para así poder detectar cuáles son las variables realmente significativas. Finalmente comentaremos cómo podemos detectar y evitar el problema de la duplicación de información que surge cuando se utilizan variables correlacionadas, conocido con el nombre de *multicolinealidad*.

2. Estimación de la varianza de los errores

Dada una muestra de observaciones, el modelo estará totalmente determinado una vez que se especifiquen los valores estimados de los coeficientes β_0, β_1, ... β_k y se estime la varianza de los errores σ^2. Todavía nos falta determinar esta última.

Considerando los residuos como estimaciones de los valores del término de error, entonces podemos estimar la varianza de este término a partir de la varianza de los residuos:

$$s^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki})^2$$

Si tenemos en cuenta que este sumatorio es la suma de los cuadrados de los errores, podemos escribirlo de esta manera:

$$s^2 = \frac{SCE}{n-k-1}$$

Residuos no independientes

Se divide por:

$n - (k + 1) = n - k - 1$

porque los n residuos no son independientes (están relacionados por las $(k + 1)$ ecuaciones normales de la regresión).

3. Distribuciones probabilísticas de los parámetros de la regresión

En primer lugar, debe quedar muy claro que cada muestra determina una regresión lineal múltiple y, por tanto, un conjunto de coeficientes:

De manera que tendríamos para cada coeficiente de la regresión una colección de valores estimados de los parámetros:

$\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\beta}_{03}, \ldots, \hat{\beta}_{0m} \longrightarrow \beta_0$

$\hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \ldots, \hat{\beta}_{1m} \longrightarrow \beta_1$

....

$\hat{\beta}_{k1}, \hat{\beta}_{k2}, \hat{\beta}_{k3}, \ldots, \hat{\beta}_{km} \longrightarrow \beta_k$

Notación

El primer subíndice nos indica el parámetro y el segundo, que se trata de una observación de éste obtenida a partir de la muestra.

Así, $\beta_0, \beta_1, \ldots \beta_k$ son unas variables aleatorias que habrá que estudiar para poder inferir nuestros resultados a la población de la que hemos extraído las muestras. Primero las caracterizaremos calculando sus valores esperados y las desviaciones estándar:

a) Valor esperado de $\hat{\beta}_j$: $E(\beta_j) = \beta_j$; para $j = 1, \ldots, k$. Observamos que los valores esperados de estos parámetros son iguales a los valores poblacionales de éstos. Aunque estos valores sean desconocidos, este resultado nos será de gran utilidad a la hora de hacer inferencia estadística.

Estos cálculos se muestran de forma detallada en los anexos 3.1 y 3.2.

b) Varianza de $\hat{\beta}_j$. Las varianzas de las $\hat{\beta}_j$ son los elementos de la diagonal de la matriz $\sigma^2(X^tX)^{-1}$, es decir:

$$\begin{bmatrix} \mathrm{var}(\hat{\beta}_0) & & & \\ & \mathrm{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \mathrm{var}(\hat{\beta}_k) \end{bmatrix} = \sigma^2 \; diag \; (X^tX)^{-1}$$

Ya hemos calculado la media y la varianza de los estimadores $\hat{\beta}_j$. Puesto que la variable Y se distribuye normalmente y las $\hat{\beta}_j$ son combinación lineal de las observaciones y_j, se puede asegurar que las $\hat{\beta}_j$ se distribuirán normalmente:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{jj})$$

donde q_{jj} es el elemento de la fila j y columna j de la matriz $(X^tX)^{-1}$. Dado que la varianza σ^2 es desconocida, deberemos utilizar el valor estimado a partir de los datos de la muestra, algo que ya hemos hecho en el apartado 1 de esta sesión:

$$s^2 = \frac{SCE}{n-k-1}$$

De manera que:

$$\begin{bmatrix} \mathrm{var}(\hat{\beta}_0) & & & \\ & \mathrm{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \mathrm{var}(\hat{\beta}_k) \end{bmatrix} = s^2 \; diag \; (X^tX)^{-1}$$

Y las desviaciones estándar de los estimadores serán:

$$s_{\hat{\beta}_j} = \sqrt{\text{var}(\hat{\beta}_j)} \text{ , per } j = 1, 2, ..., k$$

Una vez conocidas las estimaciones de los parámetros, $\hat{\beta}_j$, y de sus desviaciones estándar, $s_{\hat{\beta}_j}$, escribiremos el resultado de la regresión de la forma siguiente:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$$

$$s_{\hat{\beta}_0} \quad s_{\hat{\beta}_1} \quad s_{\hat{\beta}_2} \quad s_{\hat{\beta}_k}$$

$$s^2 \quad R^2$$

Es decir:

1) Escribimos el modelo de regresión lineal obtenido.

2) Bajo cada uno de los parámetros estimados escribimos su desviación típica.

3) Por último, en la línea siguiente escribimos la estimación de la varianza de los residuos y el coeficiente de determinación.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Continuamos con el caso en el que queríamos explicar los gastos (en decenas de euros/año) de los ordenadores de un departamento comercial a partir de su edad (en años) y del número de horas diarias que trabajan (horas/día). Con esta finalidad se había tomado una muestra de cinco ordenadores y se habían obtenido los resultados siguientes:

Gastos (Y) (decenas de euros/año)	Antigüedad (X_1) (años)	Horas de trabajo (X_2) (horas/día)
24,6	1	11
33,0	3	13
36,6	4	13
39,8	4	14
28,6	2	12

El modelo de regresión obtenido era el siguiente: $\hat{y} = -5,0 + 2,6x_1 + 2,4x_2$. Habíamos encontrado:

$$(X^tX)^{-1} = \begin{bmatrix} 181,4 & 14 & -17,5 \\ 14 & 1,3 & -1,4 \\ -17,5 & -1,4 & 1,7 \end{bmatrix}, \text{ y también } s^2 = \frac{SQE}{n-k-1} = \frac{2,16}{5-2-1} = \frac{2,16}{2} = 1,08$$

De manera que:

- $\text{var}(\hat{\beta}_0) = 1{,}08 \cdot 181{,}4 = 195{,}91 \Rightarrow s_{\hat{\beta}_0} = 14{,}0$
- $\text{var}(\hat{\beta}_1) = 1{,}08 \cdot 1{,}3 = 1{,}404 \Rightarrow s_{\hat{\beta}_1} = 1{,}18$
- $\text{var}(\hat{\beta}_2) = 1{,}08 \cdot 1{,}7 = 1{,}836 \Rightarrow s_{\hat{\beta}_2} = 1{,}35$

Podemos escribir los resultados de la manera siguiente:

$$\begin{array}{ccccc} y = -5{,}0 & + & 2{,}6x_1 & + & 2{,}4x_2 \\ (14{,}0) & & (1{,}18) & & (1{,}35) \end{array}$$

$$S^2 = 1{,}08 \qquad R^2 = 0{,}985$$

4. Intervalos de confianza de los parámetros del modelo

En los modelos de regresión lineal múltiple resulta útil construir estimaciones de intervalos de confianza para los coeficientes de la regresión $\hat{\beta}_j$. Como hemos visto en el apartado anterior, los estimadores $\hat{\beta}_j$ siguen distribuciones $N(\beta_j, s^2_{\hat{\beta}_j})$. Por tanto, se puede demostrar que la variable tipificada:

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}}$$

sigue una **distribución *t* de Student con $n - k - 1$ grados de libertad**. Puesto que:

$$P\left(-t_{\alpha/2,n-k-1} \leq \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \leq t_{\alpha/2,n-k-1}\right) = 1 - \alpha$$

Un **intervalo de confianza** con un nivel de confianza de $100(1 - \alpha)\%$ para el coeficiente $\hat{\beta}_j$ de la regresión viene dado por:

$$[\hat{\beta}_j - t_{\alpha/2,n-k-1}s_{\hat{\beta}_j}, \hat{\beta}_j + t_{\alpha/2,n-k-1}s_{\hat{\beta}_j}]$$

donde $\hat{\beta}_j$ es el valor estimado del parámetro a partir de la muestra.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Calculemos ahora los intervalos de confianza para los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ de nuestro ejemplo:

- Intervalo de confianza para $\hat{\beta}_1$ con un nivel de confianza del 95%. Observando la tabla de la distribución *t* de Student con $n - k - 1 = 5 - 2 - 1 = 2$ grados de libertad, el valor crítico correspondiente para $\alpha/2 = 0{,}025$ es: $t_{0{,}025;2} = 4{,}3027$. El intervalo de confianza será:

$$[2{,}6 - 4{,}3027 \cdot 1{,}18;\ 2{,}6 + 4{,}3027 \cdot 1{,}18] = [-2{,}50;\ 7{,}70]$$

- Intervalo de confianza para $\hat{\beta}_2$ con un nivel de confianza del 95%. Ahora el intervalo de confianza será:

$$[2{,}4 - 4{,}3027 \cdot 1{,}35;\ 2{,}4 + 4{,}3027 \cdot 1{,}35] = [-3{,}43;\ 8{,}23]$$

5. Contraste de hipótesis sobre los parámetros del modelo

Muchas veces es interesante hacer tests de hipótesis sobre los coeficientes de la regresión. Casi siempre nos interesará saber si un coeficiente β_i es igual a cero, ya que esto querría decir que la variable X_i correspondiente no figura en el modelo de regresión y, por tanto, no es una variable explicativa del comportamiento de la variable Y.

Para hacer este contraste de hipótesis, seguimos el procedimiento que exponemos a continuación:

1) Establecemos las hipótesis. Para cada β_j:

- Hipótesis nula: H_0: $\beta_j = 0$ (la variable X_j no es explicativa).
- Hipótesis alternativa: H_1: $\beta_j \neq 0$.

En caso de que no rechacemos la hipótesis nula, esto querrá decir que la variable X_j no es una variable explicativa y que, por tanto, podemos eliminarla del modelo.

2) Calculamos el estadístico de contraste: si la hipótesis nula es cierta ($\beta_j = 0$), entonces obtenemos el estadístico de contraste:

$$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

que es una observación de una distribución t de Student con $n - k - 1$ grados de libertad.

3) Finalmente, a partir de un nivel de significación (α) estableceremos un criterio de decisión. Para hacerlo, tenemos dos opciones:

a) A partir del p-valor. El p-valor es la probabilidad del resultado observado o de otro más alejado si la hipótesis nula es cierta. Es decir:

$$p = 2P(t_{n-k-1} > |t|)$$

- Si $p \leq \alpha$, se rechaza la hipótesis nula H_0.
- Si $p > \alpha$, no se rechaza la hipótesis nula H_0.

b) A partir de los valores críticos $\pm t_{\alpha/2;n-k-1}$, de manera que:

- Si $|t| > t_{\alpha/2;n-k-1}$, se rechaza la hipótesis nula H_0; por tanto, la variable X_j es una variable explicativa de la variable Y y, por tanto, no podemos eliminarla del modelo.

- Si $|t| \leq t_{\alpha/2,n-k-1}$, no se rechaza la hipótesis nula H_0; por tanto, la variable X_j no es una variable explicativa de la variable Y y, por tanto, podemos eliminarla del modelo.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Volvemos a nuestro ejemplo para hacer un contraste de hipótesis sobre los parámetros de la regresión y enterarnos de si las variables son explicativas de los gastos anuales de mantenimiento de los ordenadores o no. Utilizaremos un nivel de significación $\alpha = 0{,}05$.

- Contraste por β_1

1. Establecemos las hipótesis nula y alternativa:

– Hipótesis nula: H_0: $\beta_1 = 0$.
– Hipótesis alternativa: H_1: $\beta_1 \neq 0$.

2. Calculamos el estadístico de contraste: $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{2{,}6}{1{,}18} = 2{,}20$

3. Calculamos el *p*-valor correspondiente a este estadístico de contraste:

$$p = 2\,P(t_{n-k-1} > |t|) = 2\,P(t_2 > 2{,}20) = 2 \cdot 0{,}0794 = 0{,}1588.$$

Dado que 0,1588 > 0,05, no rechazamos H_0. Por tanto, la variable X_1 no es una variable explicativa y, por tanto, podemos eliminarla del modelo.

- Contraste por β_2

1. Establecemos las hipótesis:

– Hipótesis nula: H_0: $\beta_2 = 0$
– Hipótesis alternativa: H_1: $\beta_2 \neq 0$

2. Calculamos el estadístico de contraste: $t = \dfrac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \dfrac{2{,}4}{1{,}35} = 1{,}77$.

3. Calculamos el *p*-valor correspondiente a este estadístico de contraste:

$$p = 2P(t_{n-k-1} > |t|) = 2P(t_2 > 1{,}77\ ) = 2 \cdot 0{,}1094 = 0{,}2188$$

Dado que 0,2188 > 0,05, no rechazamos H_0. Por tanto, la variable X_2 tampoco es una variable explicativa y, por tanto, podemos eliminarla del modelo.

En este modelo de regresión lineal múltiple ninguna de las dos variables nos explica la variable "gasto en mantenimiento".

6. Contrastación conjunta del modelo

Hemos visto cómo hay que hacer el contraste de hipótesis para ver si cada una de las variables X_i, individualmente, contribuye a explicar la variable Y.

Ahora queremos contrastar el modelo de forma global, teniendo en cuenta todas las variables X_i que hemos utilizado para encontrarlo.

1) Establecemos las hipótesis:

- Hipótesis nula: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. Nos indica que no existe relación lineal entre la variable Y y ninguna de las variables X_i.

- Hipótesis alternativa: H_1: al menos una $\beta_0 \neq 0$.

Otras formas de expresar las hipótesis

Otra forma de expresar estas hipótesis es la siguiente:

Hipótesis nula:

H_0: $R^2 = 0$

Nos indica que la parte de la variación explicada por el modelo es cero, es decir, que no existe ninguna relación lineal entre la variable Y y cualquiera de las variables X_i.

Hipótesis alternativa:

H_1: $R^2 > 0$

2) Calculamos el estadístico de contraste.

Esta prueba se basa en un estadístico de contraste que es una observación de una distribución F cuando H_0 es cierta.

Buscaremos una relación entre la variación explicada por el modelo de regresión múltiple y la no explicada por el mismo modelo. Si la proporción de variación explicada en relación con la no explicada es grande, entonces se confirmará la utilidad del modelo y no rechazaremos la hipótesis nula H_0.

Recordemos que...

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

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

$$SCE = \sum_{i=1}^{n} e_i^2$$

A partir de la descomposición de la suma de cuadrados totales según la suma de cuadrados de la regresión más la suma de los cuadrados de los errores:

$$SCT = SCR + SCE$$

- *SCT*: es la suma de cuadrados que, dividida por $(n - 1)$, nos da la varianza muestral de la variable Y. Esta suma tiene $n - 1$ grados de libertad.
- *SCE*: es la suma de los cuadrados de los errores, que como ya hemos comentado en más de una ocasión, tiene $(n - k + 1)$ grados de libertad.
- *SCR*: es la suma de los cuadrados de la regresión. Esta cantidad tiene k grados de libertad.

Bajo la hipótesis nula, H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$:

- *SCR* tiene una distribución χ^2 con k grados de libertad.
- *SCE* tiene una distribución χ^2 con $n - k - 1$ grados de libertad.
- *SCR* y *SCE* son independientes.

El cociente de dos variables χ^2 divididas por sus grados de libertad da una variable F de Snedecor con los grados de libertad correspondientes al numerador y denominador del cociente.

Así pues, podemos definir el **estadístico de contraste**:

$$f = \frac{SCR/k}{SCE/(n-k-1)}$$

Es una observación de una distribución F de Snedecor con k y $(n - k - 1)$ grados de libertad.

Si la hipótesis nula es cierta y, por tanto, no existe ningún tipo de relación lineal entre Y y las variables X_i, el estadístico tendrá un valor cercano a uno. Pero cuando existe cierta relación, la suma de los cuadrados de la regresión (numerador) aumenta y la suma de los cuadrados de los errores (denominador) disminuye, de manera que el valor del estadístico de contraste aumenta. Si este valor supera un valor crítico de la distribución F, entonces rechazamos la hipótesis nula.

3) Establecemos un criterio de decisión a partir de un nivel de significación α:

A partir de este valor crítico de la distribución F de Snedecor:

- Si $f > F_{\alpha;k;n-k-1}$, rechazamos H_0; por tanto, el modelo explica significativamente la variable Y. Es decir, el modelo sí que contribuye con información a explicar la variable Y.

- Si $f < F_{\alpha;k;n-k-1}$, no rechazamos H_0; por tanto, el modelo no explica de forma significativa la variable Y.

También podemos hacerlo a partir del p-valor: $p = P(F_{\alpha;k;n-k-1} > f)$.

- Si $p \leq \alpha$, se rechaza la hipótesis nula H_0.
- Si $p > \alpha$, no se rechaza la hipótesis nula H_0.

Los cálculos necesarios se pueden resumir en la tabla siguiente, conocida como **tabla de análisis de la varianza**:

Tabla de análisis de la varianza

En la primera columna se pone la fuente **de la variación**, es decir, los elementos del modelo responsables de variación.

En la segunda columna ponemos las **sumas de cuadrados** correspondientes.

En la tercera columna ponemos los **grados de libertad** correspondientes a las sumas de cuadrados.

En la cuarta columna y bajo el nombre de **media de cuadrados** se ponen las sumas de cuadrados divididas por los grados de libertad correspondientes. Sólo para *SCR* y *SCE*.

Fuente de la variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
$x_1, x_2, \ldots, x_k$	*SCR*	k	SCR/k
e	*SCE*	$n-k-1$	$SCE/(n-k-1)$
y	*SCT*	$n-1$	

Es muy importante tener presente el hecho siguiente: que el modelo lineal explique de forma significativa la variable Y no implica que todas las variables sean explicativas; para saberlo, deberemos contrastarlas de una en una, tal como se ha explicado en el apartado anterior.

Ejemplo de los gastos de los ordenadores según su antigüedad y las horas diarias de trabajo

Haremos un contraste conjunto del modelo obtenido anteriormente para los ordenadores. Tomaremos $\alpha = 0{,}05$.

1. Establecemos las hipótesis nula y alternativa:

- Hipótesis nula: H_0: $\beta_1 = \beta_2 = 0$
- Hipótesis alternativa: H_1: al menos una $\beta_i \neq 0$, $i = 1, 2$

2. Calculamos el estadístico de contraste:

Fuente de la variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
x_1, x_2	*SCR*	2	145,81/2 = 72,9
E	*SCE*	5 – 2 – 1 = 2	2,16/2 = 1,08
y	*SCT*	5 – 1 = 4	

Tenemos que: $f = \dfrac{SCR/k}{SCE/(n-k-1)} = \dfrac{72{,}9}{1{,}08} = 67{,}5$.

3. Establecemos un criterio de decisión a partir de un nivel de significación $\alpha = 0{,}05$. Mirando las tablas de la distribución F de Snedecor, tenemos que el valor crítico para $\alpha = 0{,}05$ y 2 grados de libertad en el numerador y 2 en el denominador es $F_{0{,}05;2;2} = 19{,}0$.

Puesto que $67{,}5 > 19{,}0$, entonces rechazamos la hipótesis nula, de manera que el modelo en conjunto es bueno para explicar la variable Y.

Con el p-valor tenemos que: $p = P(F_{2;2} > 67{,}5) = 0{,}0146 < 0{,}05$; por tanto, rechazamos la hipótesis nula.

Llegados a este punto, nos hacemos la pregunta siguiente: ¿cómo puede ser que el modelo en conjunto sea bueno para explicar la variable Y y, en cambio, el contraste por separado para cada una de las variables X_1 y X_2 nos haya dado que ninguna de las dos era explicativa de la variable Y? A primera vista parece que sean resultados contradictorios. Esto se debe a la presencia de multicolinealidad en nuestro problema. Lo trataremos en el apartado siguiente.

7. El problema de la multicolinealidad

En los problemes de regresión lineal múltiple esperamos encontrar dependencia entre la variable Y y las variables explicativas $X_1, X_2, ..., X_k$. Pero en algunos problemas de regresión podemos tener también algún tipo de dependencia entre algunas de las variables X_j. En este caso tenemos información redundante en el modelo.

Ejemplo de modelo que puede presentar multicolinealidad

Si queremos construir un modelo para predecir el precio (Y) de un ordenador según la velocidad del procesador (X_1), la capacidad del disco duro (X_2) y la cantidad de memoria RAM (X_3), es posible que las variables X_1 y X_3 estén relacionadas: sería el caso de que el procesador necesitase un mínimo de memoria RAM para funcionar de manera óptima.

En caso de que haya algún tipo de dependencia entre las variables, diremos que existe **multicolinealidad**. La multicolinealidad puede tener efectos muy importantes en las estimaciones de los coeficientes de la regresión y, por tanto, sobre las posteriores aplicaciones del modelo estimado.

Variables explicativas independientes

En las hipótesis estructurales básicas del modelo de regresión lineal múltiple ya hemos pedido que las variables $X_1, X_2, ..., X_k$ sean independientes.

Como ya se ha comentado antes, un efecto de la multicolinealidad lo hemos sufrido durante esta sesión en nuestro ejemplo de los ordenadores.

Hemos hecho contraste sobre los parámetros de la regresión y sobre el modelo conjunto y hemos obtenido resultados aparentemente contradictorios, pero que realmente no lo son.

Los contrastes individuales sobre los parámetros indican que la contribución de una variable, como por ejemplo antigüedad de los ordenadores, no tiene significación después de haber descontado el efecto de la variable "número de horas de funcionamiento".

Por otra parte, el contraste conjunto indica que al menos una de las dos variables contribuye a la predicción de Y (es decir, uno de los parámetros o los dos son diferentes de cero). De hecho, es muy probable que las dos variables contribuyan a ello, pero la contribución de la una encubre la de la otra.

Así pues, en estos casos en los que tenemos variables independientes muy correlacionadas en un modelo de regresión, los resultados pueden ser confusos. Habitualmente, lo que se hace es incluir sólo una de estas variables en el modelo.

8. Resumen

En esta última sesión, dedicada a la regresión lineal múltiple, hemos visto cómo debemos hacer inferencia sobre los coeficientes de la regresión obtenidos a partir de la muestra, en particular cómo debemos calcular un intervalo de confianza y cómo debemos hacer un contraste de hipótesis para cada uno de los coeficientes obtenidos para decidir si las variables X_j nos explican realmente el comportamiento de la variable Y o podemos prescindir de algunas de ellas. También hemos visto cómo debemos hacer un contraste conjunto del modelo. Finalmente, hemos presentado los posibles problemas de multicolinealidad que podemos tener y que son debidos a la relación entre algunas de las variables explicativas que supuestamente son independientes.

Ejercicios

1. Se realiza un experimento para ver si es posible determinar el peso de un animal después de un periodo de tiempo determinado a partir de su peso inicial y de la cantidad de alimento que se le suministra. A partir de los resultados obtenidos para una muestra de $n = 10$:

Peso final (kg)	95	77	80	100	97	70	50	80	92	84
Peso inicial (kg)	42	33	33	45	39	36	32	41	40	38
Alimento (kg)	272	226	259	292	311	183	173	236	230	235

se ha obtenido el modelo de regresión lineal:

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

y las sumas de cuadrados siguientes:

$$SCR = 1.762{,}99 \qquad SCE = 256{,}30 \qquad SCT = 2.020{,}50$$

a) ¿Podéis afirmar que las variables "peso inicial" y "cantidad de alimento suministrado" son explicativas del "peso final" del animal?

b) ¿Creéis que este modelo lineal múltiple explica de forma significativa el peso final de los animales?

2. Consideremos una muestra aleatoria de cinco familias con las características siguientes:

Familia	**Ahorros (euros) Y**	**Ingresos (euros) X_1**	**Capital (euros) X_2**
A	600	8.000	12.000
B	1.200	11.000	6.000
C	1.000	9.000	6.000
D	700	6.000	3.000
E	300	6.000	18.000

a) Especificad un modelo lineal múltiple para expresar el ahorro de acuerdo con los ingresos y los capitales.

b) Estimad los parámetros del modelo de regresión lineal múltiple.

c) ¿Podéis afirmar que las variables x_1 y x_2 son explicativas?

d) ¿Creéis que este modelo lineal múltiple explica de manera significativa los ahorros?

Solucionario

1.

a) Para saber si las variables del modelo de regresión son explicativas, deberemos hacer un contraste de hipótesis sobre los parámetros obtenidos.

- **Variable X_1**:

1) Establecemos las hipótesis nula y alternativa:

- Hipótesis nula: $\beta_1 = 0$. Si este coeficiente es nulo, entonces la variable X_1 no participaría en el modelo y, por tanto, no sería explicativa del peso final de los animales.

- Hipótesis alternativa: $\beta_1 \neq 0$. En este caso la variable X_1 aporta información al modelo; por tanto, sí es explicativa del peso final.

2) Determinamos un nivel significativo $\alpha = 0{,}05$.

3) Calculamos el estadístico de contraste: $t = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = 2{,}3943$.

4) El estadístico de contraste calculado es una observación de una distribución t de Student con $10 - 2 - 1 = 7$ grados de libertad. Buscando en las tablas, encontramos el valor crítico correspondiente:

$$t_{0{,}025;7} = 2{,}3646$$

Dado que $2{,}3943 > 2{,}3646$, rechazamos H_0. La variable X_1 es significativa, aunque por muy poco.

- **Variable X_2**:

1) Establecemos las hipótesis:

- Hipótesis nula: $\beta_2 = 0$.
- Hipótesis alternativa: $\beta_2 \neq 0$.

2) Determinamos un nivel de significación: $\alpha = 0{,}05$.

3) Calculamos el estadístico de contraste: $t = \dfrac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = 3{,}7663$.

4) Dado que $3{,}7663 > 2{,}3646$, rechazamos H_0. La variable X_2 (cantidad de alimento) es significativa del peso final de los animales.

b) Haremos una contrastación conjunta del modelo:

1) Establecemos las hipótesis:

- Hipótesis nula: H_0: $\beta_1 = \beta_2 = 0$
- Hipótesis alternativa: H_1: hay un $\beta_j \neq 0$

2) Fijamos el nivel de significación: $\alpha = 0{,}05$.

3) Calculamos el estadístico de contraste. Sin embargo, primero construimos la tabla de análisis de la varianza:

Fuente de la variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
X_1, X_2	$SCR = 1.762{,}99$	$k = 2$	$SCR/k = 881{,}50$
e	$SCE = 256{,}30$	$n - k - 1 = 7$	$SCE / (n - k - 1) = 36{,}61$
Y	$SCT = 2.020{,}50$	$n - 1 = 9$	–

Estadístico de contraste: $f = \dfrac{SCR/k}{SCE/(n-k-1)} = 24{,}07$

Es una observación de una distribución F de Snedecor con $k = 2$ y $n - k - 1 = 7$ grados de libertad.

4) De las tablas tenemos un valor crítico de $F_{0{,}05;2;7} = 4{,}74$. Puesto que $24{,}07 > 4{,}74$, rechazamos H_0 con una confianza del 95%. Entonces el modelo explica de forma significativa el peso final de los animales.

2.

a) En este problema tenemos que el número de observaciones es $n = 5$ y que el número de variables independientes es $k = 2$.

Modelo lineal múltiple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

Matricialmente:

$$\begin{bmatrix} 600 \\ 1.200 \\ 1.000 \\ 700 \\ 300 \end{bmatrix} = \begin{bmatrix} 1 & 8.000 & 12.000 \\ 1 & 11.000 & 6.000 \\ 1 & 9.000 & 6.000 \\ 1 & 6.000 & 3.000 \\ 1 & 6.000 & 18.000 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

b) Los valores estimados del modelo de regresión vienen dados por:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{bmatrix} = (X^t X)^{-1} X^t Y$$

donde $(X^t X)^{-1}$ es la matriz inversa de la matriz $(X^t X)$.

Ahora tenemos:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 8.000 & 11.000 & 9.000 & 6.000 & 6.000 \\ 12.000 & 6.000 & 6.000 & 3.000 & 18.000 \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} 3.800 \\ 33.000.000 \\ 27.900.000 \end{bmatrix}$$

$$X^t X = \begin{bmatrix} 5 & 4.000 & 45.000 \\ 40.000 & 338.000.000 & 342.000.000 \\ 45.000 & 342.000.000 & 549.000.000 \end{bmatrix}$$

Atención

Según el número de cifras decimales que cojáis a partir de aquí, los resultados pueden ser un poco diferentes, sin que ello signifique que sean incorrectos.

$$(X^t X)^{-1} = \begin{bmatrix} 6{,}0492063 & -0{,}00057936 & -0{,}00013492 \\ -0{,}00057936 & 0{,}6349206 \cdot 10^{-7} & 0{,}79365079 \cdot 10^{-8} \\ -0{,}00013492 & 0{,}79365079 \cdot 10^{-8} & 0{,}79365079 \cdot 10^{-8} \end{bmatrix}$$

Ya podemos calcular los parámetros:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y = \begin{bmatrix} 103{,}6507937 \\ 0{,}1150793651 \\ -0{,}02936507937 \end{bmatrix}$$

Tenemos:

$\hat{\beta}_0 = 103{,}6507937$ $\quad \hat{\beta}_1 = 0{,}1150793651$ $\quad \hat{\beta}_2 = -0{,}02936507937$

El modelo de regresión obtenido es:

$$\hat{y} = 103{,}6507937 + 0{,}1150793651 x_1 - 0{,}02936507937 x_2$$

c) Para determinar si las variables son explicativas, debemos hacer inferencia estadística sobre los parámetros del modelo.

Sin embargo, antes debemos hacer algunos cálculos más. Primero calcularemos las varianzas de los parámetros estimados. Vienen dadas por los términos de la diagonal de la matriz:

$$s^2 diag(X^t X)^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_0) & & & \\ & \text{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

donde s^2 es la varianza de los errores:

$$s^2 = \frac{SCE}{n-(k+1)} = 3.896{,}825$$

En este caso tenemos las varianzas y desviaciones típicas de los estimadores siguientes:

$$s_{\hat{\beta}_0} = \sqrt{\text{var}(\hat{\beta}_0)} = 125{,}3600174$$

$$s_{\hat{\beta}_1} = \sqrt{\text{var}(\hat{\beta}_1)} = 0{,}012843081$$

$$s_{\hat{\beta}_2} = \sqrt{\text{var}(\hat{\beta}_2)} = 0{,}00454071$$

Ahora ya estamos en condiciones de hacer contrastes de hipótesis sobre los parámetros del modelo.

- **Variable X_1**:

1) Establecemos las hipótesis:

- Hipótesis nula: $\beta_1 = 0$. Si el coeficiente β_1 que vincula la relación entre X_1 e Y puede ser cero, esto significa que X_1 puede no tener ningún efecto sobre Y; entonces diremos que x_1 no es una variable explicativa.

- Hipótesis alternativa: $\beta_1 \neq 0$. En este caso diremos que X_1 es una variable explicativa.

2) Determinamos un nivel de significación: $\alpha = 0{,}05$.

3) Calculamos el estadístico de contraste:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 8{,}96041$$

Es una observación de una distribución t de Student con $n - k - 1 = 2$ grados de libertad.

4) Si miramos las tablas, tenemos para un valor crítico: $t_{0,025;2} = 4,3027$. Dado que $8,96041 > 4,3027$, rechazamos H_0. La variable X_1 (ingresos) es explicativa de los ahorros.

- **Variable X_2**: haremos lo mismo para la variable X_2 (capital).

1) Establecemos las hipótesis:

- Hipótesis nula: $\beta_2 = 0$
- Hipótesis alternativa: $\beta_2 \neq 0$

2) Determinamos un nivel de significación: $\alpha = 0,05$.

3) Calculamos el estadístico de contraste:

$$t = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = -6,46705$$

que es una observación de una distribución t de Student con $n - k - 1 = 2$ grados de libertad.

4) De las tablas teníamos un valor crítico: $t_{0,025;2} = 4,3027$. Puesto que $6,46705 > > 4,3027$, rechazamos H_0. La variable X_2 (capital) también es explicativa de los ahorros.

d) Para determinar si este modelo lineal múltiple explica de forma significativa los ahorros de las familias, deberemos hacer una contrastación conjunta del modelo.

1) Establecemos las hipótesis nula y alternativa:

- Hipótesis nula: H_0: $\beta_1 = \beta_2 = 0$.
- Hipótesis alternativa: H_1: hay al menos un $\beta_j \neq \ddot{0}$.

2) Determinamos un nivel significativo, por ejemplo $\alpha = 0,05$.

3) Calcularemos el estadístico de contraste. Sin embargo, antes deberemos calcular las sumas de cuadrados y construir la tabla del análisis de la varianza. Para calcular la suma de cuadrados de la regresión (*SCR*) necesitamos conocer:

– la media de las $y_i = \bar{y} = 760,0$.

– y los valores estimados de y_i y :

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \dots\\ \hat{y}_n\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{21} & \dots & x_{k1}\\ 1 & x_{12} & x_{22} & \dots & x_{k2}\\ \dots & \dots & \dots & \dots & \dots\\ 1 & x_{1n} & x_{2n} & \dots & x_{kn}\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \dots\\ \hat{\beta}_k\end{bmatrix}$$

, ahora tenemos:

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\end{bmatrix} = \begin{bmatrix}671{,}9047619\\ 1.193{,}33333\\ 963{,}1746032\\ 706{,}0317460\\ 265{,}5555556\end{bmatrix}$$

Para *SCE*, antes debemos calcular el vector de los errores:

$$e = \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\end{bmatrix} - \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\end{bmatrix} = \begin{bmatrix}-71{,}90476190\\ 6{,}666666667\\ 36{,}82539683\\ -6{,}031746032\\ 34{,}44444444\end{bmatrix}$$

Así pues, las sumas de cuadrados son:

$$SCT = \sum(y_i - \bar{y})^2 = 492.000$$

$$SCR = \sum(\hat{y}_i - y)^2 = 484.206{,}34$$

$$SCE = \sum(y_i - \hat{y})^2 = \sum e_i^2 = 7.793{,}65$$

Podemos construir la tabla de análisis de la varianza:

Fuente de la variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
X_1, X_2	SCR = 484.206,34	$k = 2$	SCR/k = 242.103,17
e	SCE = 7.793,65	$n - k - 1 = 2$	$SCE/(n - k - 1)$ = 3.896,825
Y	SCT = 492.000	$n - 1 = 4$	–

Estadístico de contraste: $f = \dfrac{SCR/k}{SCE/(n-k-1)} = 62{,}12$

Es una observación de una distribución *F* de Snedecor con $k = 2$ y $n - k - 1 = 2$ grados de libertad.

4) De las tablas tenemos un valor crítico de $F_{0,05;2;2} = 19{,}0$. Dado que 62,12 > > 19,0, rechazamos H_0. Así pues, este modelo de regresión múltiple explica de forma significativa los ahorros de las familias a partir de los ingresos y del capital.

Anexos

Anexo 1

Valor esperado de $\hat{\beta}_j$:

Para buscar los valores esperados de $\hat{\beta}_j$, utilizaremos la notación matricial que ya hemos introducido en el módulo anterior y que nos permitirá cierta comodidad a la hora de escribir todas las ecuaciones. A partir de la ecuación matricial que nos permitía encontrar los estimadores de los coeficientes de la regresión:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Para simplificar todavía más los cálculos, llamaremos $C = (X^t X)^{-1} X^t$ y así podremos escribir la última ecuación de la forma: $\hat{\beta} = CY$. Por otra parte, el modelo de regresión lineal múltiple:

Obserad que...

... $CX((X^t X)^{-1} X^t X) = I$
es la matriz identidad.

$$y_i = \beta_0 = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$

escrito matricialmente: $Y = X\beta + e$. De manera que:

$$\hat{\beta} = CY = C(X\beta + e) = CX\beta + Ce = \beta + Ce$$

Si ahora calculamos el valor esperado:

$$E(\hat{\beta}) = E(\beta) + E(Ce) = \beta + CE(e) = \beta$$

donde hemos considerado que $E(e) = 0$, tal como supusimos en la sesión anterior en las hipótesis estructurales básicas del modelo de regresión lineal múltiple.

Linealidad

Hemos utilizado la propiedad de linealidad de la esperanza matemática:

$E(aX) = aE(X)$

En resumen, hemos obtenido que: $E(\hat{\beta}) = \beta$, es decir:

$$E(\hat{\beta}_1) = \beta_1$$
$$E(\hat{\beta}_2) = \beta_2$$
$$.....$$
$$E(\hat{\beta}_k) = \beta_k$$

Anexo 2

Varianza de $\hat{\beta}_j$:

Para una $\hat{\beta}_j$, su varianza vendrá dada como siempre por:

Aquí ya hemos utilizado el resultado anterior:

$E(\hat{\beta}_j) = \beta_j$

$$Var(\hat{\beta}_j) = E[(\hat{\beta}_j - \beta_j)^2]$$

Para calcular esta varianza, utilizaremos una vez más la notación y el cálculo matricial.

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t] = E\left[\begin{bmatrix}(\hat{\beta}_0-\beta_0)\\(\hat{\beta}_1-\beta_1)\\\dots\\(\hat{\beta}_k-\beta_k)\end{bmatrix}((\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\dots(\hat{\beta}_k-\beta_k))\right] =$$

$$= \begin{bmatrix} E[(\hat{\beta}_0-\beta_0)^2] & E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)] & \dots & E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_k-\beta_k)] \\ E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)] & E[(\hat{\beta}_1-\beta_1)^2] & \dots & E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_k-\beta_k)] \\ \dots & \dots & \dots & \dots \\ E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_k-\beta_k)] & E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_k-\beta_k)] & \dots & E[(\hat{\beta}_k-\beta_k)^2] \end{bmatrix}$$

La matriz anterior recibe el nombre de **matriz de varianzas-covarianzas**, ya que sus elementos de la diagonal son las varianzas de las $\hat{\beta}_j$ y los elementos de fuera de la diagonal son las covarianzas de los pares de variables $\hat{\beta}_j$ y $\hat{\beta}_m$. A nosotros nos interesan las varianzas de las $\hat{\beta}_j$, es decir, los valores esperados de los elementos de la diagonal de la matriz:

La esperanza de una matriz

Hemos dicho que la esperanza de una matriz es la matriz de las esperanzas de sus elementos.

$$(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t$$

Por otra parte, hemos visto antes que $\hat{\beta} = \beta + Ce$, de manera que podemos escribir: $\hat{\beta} - \beta = Ce$ y, por tanto:

$$(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t = (Ce)(Ce)^t$$

Combinando estos resultados, tenemos que las varianzas de las $\hat{\beta}_j$ son los valores esperados de los elementos de la diagonal de la matriz $(Ce)(Ce)^t$, es decir:

Producto de matrices

Recordemos la importante propiedad del producto de matrices:

$(AB)^t = B^tA^t$

$$E[Cee^tC^t] = CE[ee^t]C^t = \sigma^2CC^t = \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}$$

donde hemos tenido en cuenta que $E[ee^t] = \sigma^2 I$ para las hipótesis estructurales básicas del modelo de regresión lineal múltiple que supusimos en la sesión anterior.

Finalmente tenemos que las varianzas de las $\hat{\beta}_j$ son los elementos de la diagonal de la matriz: $\sigma^2(X^tX)^{-1}$, es decir:

$$\begin{bmatrix}\text{var}\,(\hat{\beta}_0) & & & \\ & \text{var}\,(\hat{\beta}_1) & & \\ & & \dots & \\ & & & \text{var}\,(\hat{\beta}_k)\end{bmatrix} = \sigma^2 diag(X^tX)^{-1}$$