


## 1. El modelo de regresión múltiple

### 1.1. Introducción

Cuando estudiamos la relación entre dos variables (modelos simples), distinguíamos entre el análisis de la correlación en que las dos variables eran aleatorias y buscábamos una medida de la dependencia, que representábamos mediante el coeficiente de correlación de Pearson  $r_{X,Y}$ , y el análisis de la regresión, en la cual sólo una de las variables (Y) era aleatoria (la que queremos explicar), mientras que la otra (X) era fija y controlable por el investigador.

En un intento de ampliar el modelo de regresión, tendremos en cuenta la posibilidad de explicar una variable Y (variable endógena) a partir de diferentes variables explicativas  $X_2, \dots, X_K$  (variables exógenas). Ciertamente, este enfoque nos acerca más a la realidad económica que se quiere modelizar, ya que gran parte de las relaciones económicas son multivariantes. ¿De qué factores puede depender la inflación de un país? Tenemos presente que son muchas las variables exógenas: salarios, precios de las materias primas, presión fiscal, etc. Nuestro objetivo es traducir estas relaciones económicas en modelos estadísticos precisos y completos. 

La precisión implica establecer el tipo de ecuación matemática que relaciona las variables, y su cumplimiento exige que en el modelo se encuentren todas las variables que contribuyen a explicarlo. En cambio, el modelo económico es impreciso por lo que respecta al tipo de relación e incompleto por lo que respecta a las variables que incluye:

inflación = f (salarios, precios de las materias primas, presión fiscal...)

El modelo de regresión explicita la función y utiliza una variable no observable o perturbación (u) que tiene, además de otras finalidades, la de presentar el efecto sobre la variable endógena entre todas las que, a pesar de ser explicativas, han sido excluidas del modelo:

inflación =  $\beta_1 + \beta_2$  salarios +  $\beta_3$  precios materias primas +  $\beta_4$  presión fiscal + u

Antes de continuar, es necesario hacer algunas consideraciones con relación al modelo de regresión. En primer lugar, tenemos que decir que siempre haremos referencia a modelos lineales:


$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_K X_K + u$$

o a modelos que sean cómodamente transformables en ecuaciones lineales. De hecho, en este punto la estadística no ha avanzado lo suficiente como para trabajar con éxito con otros tipos de funciones, de manera análoga a lo que hicimos en la recta de regresión.

Así, por ejemplo:

$$\ln Q = \ln A + \alpha \ln L + \beta \ln K + u$$

es la ecuación del modelo de regresión de la función de Cobb-Douglas estudiada en economía, donde Q es el volumen de producción, L el factor trabajo empleado y K el factor capital en las relaciones técnicas de producción. Si utilizamos  $\ln Q$  como variable endógena y  $\ln L$  y  $\ln K$  como variables exógenas, resultará fácil la estimación del modelo.

Por otro lado, hay que entender la perturbación  $u$  –como ya hemos dicho– como una especie de cajón de sastre donde va a parar todo aquello que puede contribuir a explicar el modelo, y que hemos excluido, y no sólo variables, sino también los errores de medida u otros factores aleatorios que se puedan producir y que no hemos controlado. Por tanto,  $u$  es la parte no determinista del modelo que confirma el carácter aleatorio de la variable endógena. 

#### La perturbación...

... es el componente errático del modelo y, por tanto, la desviación entre las observaciones reales de los fenómenos económicos y las observaciones esperadas de acuerdo con este modelo. En la perturbación se incluyen cuestiones tan diferentes como los errores de medida, la omisión de variables relevantes, la especificación incorrecta de la relación funcional, el comportamiento imprevisible de los individuos y de los agentes económicos, etc.

En este apartado del modelo de regresión múltiple aprenderéis:

- Cómo se especifica el modelo de regresión múltiple y cómo se expresa en notación matricial.
- Cómo se estiman los parámetros del modelo mediante los mínimos cuadrados ordinarios.
- Cómo se efectúan contrastes de hipótesis sobre los parámetros del modelo.
- Cómo se evalúa la bondad del modelo estimado.

## 1.2. Especificación del modelo

Una generalización del modelo de regresión simple, que ya estudiamos, es el modelo de regresión lineal múltiple, en el cual relacionamos la variable que queremos explicar  $Y$  (variable endógena) con las  $K-1$  variables explicativas  $X_2, X_3, \dots, X_K$  (variables exógenas) por medio de la expresión:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_K X_K + u$$

donde  $u$  es la perturbación del modelo.

Si desarrollamos la ecuación para cada observación muestral, obtendremos el sistema siguiente:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_K X_{K1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_K X_{K2} + u_2 \\ &\dots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots + \beta_K X_{Kn} + u_n \end{aligned}$$

Sería el caso, por ejemplo, de querer explicar la renta ( $Y_i$ ) de los diferentes municipios catalanes a partir de la cantidad de teléfonos instalados ( $X_{2i}$ ), de la cantidad de licencias comerciales ( $X_{3i}$ ), ..., de la recaudación del impuesto de circulación ( $X_{Ki}$ ) y de muchas otras variables que no se incorporan al modelo ( $u_i$ ). Si partimos de una muestra representativa de  $n$  municipios, se generará un sistema de ecuaciones como el que hemos descrito con anterioridad, a partir del cual estimaremos la relación lineal que presenta.

Este sistema de ecuaciones se puede expresar en términos matriciales, hecho que debe facilitar el tratamiento operativo del modelo. Así, podemos escribir:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & X_{31} & \dots & X_{K1} \\ 1 & X_{22} & X_{32} & \dots & X_{K2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{2n} & X_{3n} & \dots & X_{Kn} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}$$

lo cual permite escribir el modelo de manera general:

$$Y = X\beta + u$$

donde  $Y$  es un vector ( $n \times 1$ ) de observaciones de la variable endógena,  $X$  es una matriz ( $n \times K$ ) de observaciones de las variables exógenas,  $\beta$  es el vector ( $K \times 1$ ) de los coeficientes del modelo y  $u$  es el vector ( $n \times 1$ ) de los términos de perturbación.

Ejemplo:

Supongamos que estamos interesados en explicar los ingresos de los ingenieros titulados recientemente que trabajan por cuenta ajena a partir de los años que hace que acabaron los estudios y del número medio de horas trabajadas.

Hemos tomado una muestra de cinco de estos profesionales y hemos obtenido los resultados que vemos a continuación:

Ingresos/mes (miles de u.m.)	Antigüedad (años)	Horas/semana
245,7	1	35
329,8	3	38
365,7	4	38
398,2	4	42
286,1	2	40

El modelo que queremos estimar presenta la relación siguiente:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

y, en expresión matricial:

$$Y = X\beta + u$$

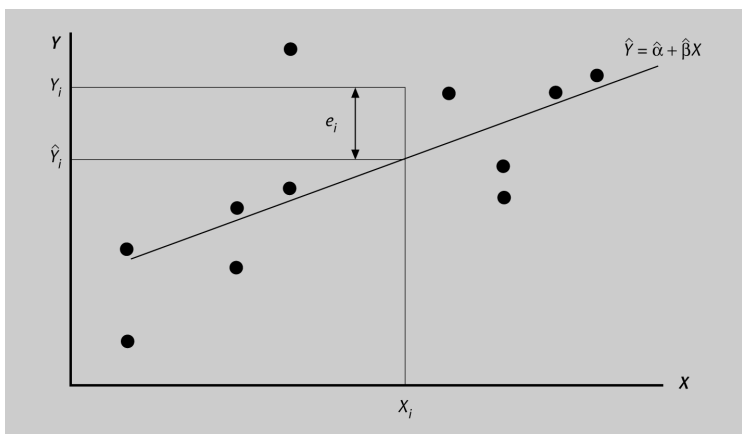
$$Y = \begin{pmatrix} 245,7 \\ 329,8 \\ 365,7 \\ 398,2 \\ 286,1 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 1 & 35 \\ 1 & 3 & 38 \\ 1 & 4 & 38 \\ 1 & 4 & 42 \\ 1 & 2 & 40 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}$$

### 1.3. Estimación mínima-cuadrática de los parámetros

Como hicimos en el modelo de regresión simple, estimaremos los parámetros del modelo mediante el método de los mínimos cuadrados.

Se trata de buscar, de entre todas las ecuaciones que se pueden trazar, la que minimice la suma de los cuadrados de los errores.

Si identificamos  $e_i$  como el error del ajuste, es decir, como la diferencia entre la observación de la variable endógena  $Y_i$  y su estimación mediante el modelo  $\hat{Y}_i$ , tendremos  $e_i = Y_i - \hat{Y}_i$  con  $i = 1, 2, \dots, n$ , que para el caso –ya estudiado– de la recta era:



y para el modelo general será:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}) \quad i = 1, 2, \dots, n$$

que tiene la siguiente traducción matricial:

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


donde  $e$  es el vector de los errores,  $\hat{Y}_i$  es el vector de las estimaciones de  $Y$  y  $\hat{\beta}$  es el vector de los estimadores mínimos-cuadráticos de los parámetros.

La suma de los cuadrados de los errores  $\sum e_i^2$  resulta de la expresión  $e'e$ , donde  $e'$  es el vector  $e$  transpuesto:

$$e'e = (e_1 e_2 \dots e_n) \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = \sum e_i^2$$

Esta operación admite una nueva formulación, cuya demostración es:

$$e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Un procedimiento idéntico al de la estimación de los parámetros de la recta de regresión es el que tenemos que desarrollar para minimizar  $\sum e_i^2 = e'e$ ; sin embargo, ahora lo haremos trabajando con notación matricial. El objetivo es encontrar la ecuación que tenga una mejor adherencia a la nube de puntos, es decir, coeficientes contenidos en el vector de estimadores que optimicen  $e'e$ . 

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{pmatrix}$$

Se puede demostrar que sería suficiente con la condición necesaria dada para la anulación de las primeras derivadas. Esto nos llevaría, con cálculo matricial, a:

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

Resulta la fórmula directa para el cálculo de los coeficientes estimados:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (X'X)^{-1} X'Y$$

donde  $X'$  es la matriz  $X$  transpuesta y  $(X'X)^{-1}$  la matriz inversa  $X'X$ . Observad que las operaciones  $X'X$  y  $X'Y$  nos las podemos ahorrar si tenemos en cuenta que:

$$\begin{aligned} X'X &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{pmatrix} \times \begin{pmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{2n} & \dots & X_{kn} \end{pmatrix} = \\ &= \begin{pmatrix} N & \sum X_{2i} & \dots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \dots & \sum X_{2i} X_{ki} \\ \dots & \dots & \dots & \dots \\ \sum X_{ki} & \sum X_{ki} X_{2i} & \dots & \sum X_{ki}^2 \end{pmatrix} \\ X'Y &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{pmatrix} \times \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_{2i} Y_i \\ \dots \\ \sum X_{ki} Y_i \end{pmatrix} \end{aligned}$$

Además, a partir de la derivada que hemos igualado a cero, obtendremos fácilmente:

$$X'(Y - X\hat{\beta}) = X'e = 0$$

propiedad importante que demuestra que las estimaciones mínimas-cuadráticas implican la suma nula de los errores para los valores de cualquier variable exógena.

$$X'e = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{pmatrix} \times \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = \begin{pmatrix} \sum e_i \\ \sum X_{2i} e_i \\ \dots \\ \sum X_{ki} e_i \end{pmatrix} = 0$$

Ejemplo

Si retomamos el modelo que hemos especificado con anterioridad:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

con

$Y_i$  = ingresos/mes, expresados en miles de u.m.,  
 $X_{2i}$  = antigüedad en el oficio, expresada en años,  
 $X_{3i}$  = horas/semana trabajadas.

A partir de la información que aportan los cinco ingenieros de la muestra tenemos:

$$\begin{array}{lll} \sum Y_i = 1.625,5 & \sum Y_i X_{2i} = 4.862,9 & \sum X_{2i}^2 = 46 \\ \sum X_{2i} = 14 & \sum Y_i X_{3i} = 63.196,9 & \sum X_{3i}^2 = 7.477 \\ \sum X_{3i} = 193 & \sum X_{2i} X_{3i} = 549 & n = 5 \end{array}$$

de donde calculamos :

$$X'X = \begin{pmatrix} 5 & 14 & 193 \\ 14 & 46 & 549 \\ 193 & 549 & 7.477 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{|X'X|} [Adj (X'X)]' = \begin{pmatrix} 76,650 & 2,3045 & -2,1477 \\ 2,3045 & 0,2450 & -0,0775 \\ -2,1477 & -0,0775 & 0,0613 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 1.625,5 \\ 4.862,9 \\ 63.196,9 \end{pmatrix}$$

La estimación de los parámetros del modelo es:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1} X'Y = \begin{pmatrix} 70,8782 \\ 41,2649 \\ 3,5928 \end{pmatrix}$$

lo cual nos lleva a la regresión siguiente:

$$\hat{Y}_i = 70,8782 + 41,2649 X_{2i} + 3,5928 X_{3i}$$

que nos permite efectuar estimaciones de la variable endógena para los diferentes valores muestrales de  $X_{2i}$  y  $X_{3i}$ .

$Y_i$	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
245,7	237,891	7,809
329,8	331,199	- 1,399
365,7	372,464	- 6,765
398,2	386,835	11,365
286,1	297,120	- 11,020

Podemos comprobar el cumplimiento de las identidades, salvando el hecho de redondear los resultados:

$$\begin{aligned}\sum Y_i &= \sum \hat{Y}_i \\ \sum e_i &= \sum X_{2i}e_i = \sum X_{3i}e_i = 0\end{aligned}$$

## Actividad

1.1. Podéis comprobar cómo la regresión lineal simple que ya estudiamos:

$$Y_i = \alpha + \beta_i X_i + u_i$$

es un caso particular del modelo general y que las expresiones que resultan de los estimadores a partir de la fórmula general:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1} X'Y$$

coinciden con las que se han obtenido para la recta de regresión:

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

## 1.4. Suma de cuadrados

Estimar un modelo es algo más que estimar sus parámetros; hay que evaluar la calidad del ajuste del mismo, la precisión de los estimadores, etc. Estas medidas complementan los cálculos cumplidos en el subapartado anterior, permiten una lectura más fácil de los resultados y –lo que es más importante– proporcionan los instrumentos necesarios para realizar una inferencia estadística que valide o no el modelo, o bien para discutir la capacidad explicativa de cada una de las variables exógenas.

Hemos podido comprobar anteriormente la siguiente relación:

$$Y_i = \hat{Y}_i + e_i$$

donde  $\hat{Y}_i$  es la estimación del modelo a partir del comportamiento de  $X_{2i}$ ,  $X_{3i}$ , ...,  $X_{ki}$  y  $e_i$  son los errores o residuos que se corresponden con la parte no explicada. También podemos ver, al asumir  $\sum e_i = 0$ , que las medias entre los valores estimados y los valores reales de  $Y_i$  coinciden:

$$\bar{Y} = \bar{\hat{Y}}$$

Recordad...

... que –como ya hemos visto en el subapartado 1.1– el comportamiento de la variable endógena es el resultado de la agregación de un componente determinista, obtenido por medio de las variables exógenas, y de otro componente de tipo aleatorio, es decir, la parte no explicada por las variables exógenas.



No obstante, es posible demostrar la relación entre varianzas:


$$S_Y^2 = S_Y^2 + S_e^2$$

donde es necesario interpretar que la dispersión que presenta la variable endógena  $S_Y^2$  se puede descomponer en la dispersión explicada por el modelo  $S_Y^2$  y en la dispersión no explicada por la regresión  $S_e^2$ .

#### Actividad

1.2. Con los resultados obtenidos al final del ejemplo anterior, calculad  $\bar{Y}$ ,  $\hat{Y}$ ,  $S_Y^2$ ,  $S_Y^2$ ,  $S_e^2$  y demostrad el cumplimiento de las relaciones siguientes:

$$\bar{Y} = \hat{Y} \quad S_Y^2 = S_Y^2 + S_e^2$$

Más cómodo que trabajar con varianzas puede ser utilizar expresiones equivalentes, como las sumas de cuadrados que se expresan en el numerador de las varianzas: 

$$S_Y^2 = \frac{SCT}{n-1} \quad S_Y^2 = \frac{SCR}{n-1} \quad S_e^2 = \frac{SCE}{n-1}$$

Así, identificamos con SCT la suma total de cuadrados de las desviaciones de Y, que es la medida de la variación de Y:

$$SCT = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - N\bar{Y}^2$$

y, en expresión matricial:

$$\sum Y_i^2 = Y'Y$$

SCR es la suma de los cuadrados de la regresión y mide la variación de la parte estimada del modelo:

$$SCR = \sum (\hat{Y}_i - \bar{Y})^2 = \sum \hat{Y}_i^2 - N\bar{Y}^2$$

Si utilizamos la notación matricial, podemos escribir:

$$\sum \hat{Y}_i^2 = \hat{Y}'\hat{Y} = \hat{\beta}'X'Y$$

SCE es la suma de los cuadrados de los errores y mide la variación residual o variación de la parte no estimada:

$$SCE = \sum e_i^2$$

que, como ya hemos visto, resulta de:

$$\sum e_i^2 = e'e$$

Por otro lado, si se llega a la relación entre las varianzas  $S_Y^2 = S_Y^2 + S_e^2$ , se llegará sin dificultad a una relación idéntica con la suma de los cuadrados:

$$SCT = SCR + SCE$$

que admite el cálculo de cualquier variación a partir del conocimiento de otros.

### 1.5. Coeficiente de determinación

La bondad del ajuste se deriva de la relación entre la variación explicada y la variación total a partir del llamado coeficiente de determinación  $R^2$ .


$$R^2 = \frac{SCR}{SCT}$$

o bien con el uso de la identidad básica  $SCT = SCR + SCE$ :

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

que tenemos que interpretar como el complemento de la fracción de la variación no explicada.

Sin tener que realizar grandes cálculos, tendríamos que llegar, de manera intuitiva, a la conclusión de que  $0 \leq R^2 \leq 1$ . Observad que, cuando  $R^2 = 1$  ( $SCT = SCR$ ), toda la dispersión de  $Y$  se explica por el modelo y que, en el otro extremo, cuando  $R^2 = 0$  ( $SCT = SCE$  y  $SCR = 0$ ), hay que entender que el modelo no explica absolutamente nada de  $Y$ .

Así pues, cualquier resultado entre 0 y 1 es técnicamente posible y, cuanto más grande sea  $R^2$ , más grande será la bondad del ajuste; en este caso, se podrá llevar a cabo una interpretación porcentual del resultado multiplicando simplemente el coeficiente por 100 ( $R^2 \times 100$ ). 

En el ejemplo que ya conocemos de los ingresos (Y) según la antigüedad ( $X_2$ ) y las horas trabajadas ( $X_3$ ), obtendremos estos resultados:

$$SCT = 14.839,4$$

$$SCR = 14.480,1$$

$$SCE = 359,3$$

Los tenemos que interpretar como medida de desigualdad entre los ingresos de los ingenieros igual a 14.839,4, de los cuales 14.480,1 son imputables a la antigüedad en el oficio y a la cantidad de horas trabajadas, mientras que 359,3 representa la cantidad de la variación de ingresos no explicada por las variables exógenas y que, por tanto, se puede asignar a otras causas. Expresado en términos relativos, obtendríamos el siguiente coeficiente de determinación:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} = 0,976$$

que significa que el 97,6% del comportamiento de Y viene explicado por  $X_2$  y  $X_3$ , al menos por lo que respecta a la muestra utilizada. Es muy importante no extrapolar esta interpretación más allá del simple ajuste muestral; si hubiésemos alterado la muestra cambiando a los ingenieros encuestados o aumentando el número de los mismos,  $R^2$  podría también haber cambiado.

La  $R^2$  mide únicamente la adherencia de la nube de puntos en la ecuación estimada, y la calidad de esta adherencia se ve condicionada por el tamaño de la muestra. Lógicamente, es más fácil conseguir buenos ajustes con pocos datos que con muchos; incluso puede haber relaciones casuales (que no tenemos que confundir con las causales) que inflen la  $R^2$ .

En definitiva, vemos que es impropio otorgar al coeficiente de determinación una relevancia que vaya más allá de la simple medida de bondad del ajuste. Otra cosa distinta es que lo tengamos que utilizar como un elemento más para la construcción de estadísticos que permitan contrastar la validez del modelo, como veremos más adelante.

### Actividad

1.3. Suponed que habéis observado en tres individuos unas variables que en apariencia están tan poco relacionadas como las siguientes:

- Y = talla de pie,
- $X_2$  = número de cigarros fumados al día,
- $X_3$  = número de hermanos.

Se obtienen los siguientes resultados:

Y	X <sub>2</sub>	X <sub>3</sub>
41	0	2
37	5	1
42	10	2

Comprobad que, cuando estimamos el modelo de regresión,

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

se llega a un coeficiente de determinación  $R^2 = 1$  y razonad que no es posible afirmar que el comportamiento de Y esté determinado en un 100% por las variables  $X_2$  y  $X_3$  para todos los individuos de una población.

### 1.6. Cálculo de las varianzas de los estimadores

Una medida de la eficiencia de los estimadores  $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_k$  se deriva de su varianza  $\text{var}(\hat{\beta}_1), \text{var}(\hat{\beta}_2), \text{var}(\hat{\beta}_3), \dots, \text{var}(\hat{\beta}_k)$  o de su desviación estándar  $S_{\hat{\beta}_1}, S_{\hat{\beta}_2}, \dots, S_{\hat{\beta}_k}$ .

Este aspecto es importante porque los parámetros poblacionales  $\beta_i$  son fijos, pero los estimadores  $\hat{\beta}_i$  toman diferentes valores de acuerdo con la muestra con la que trabajamos.

Se trata de determinar la dispersión que pueden presentar (varianza o desviación estándar) con el fin de controlar la desviación que presente con respecto a los coeficientes que se pretenden estimar. No es lo mismo un estimador que proporciona resultados muy dispersos (varianza grande) con muestras diferentes que otro que da resultados muy similares (varianza pequeña). Este último, se dice, es un estimador más eficiente que el anterior (tiene menos varianza que el primero).

Más adelante veremos cómo se puede calcular esta precisión de los estimadores de forma análoga a como lo hicimos en el caso de la recta de regresión. Llegados a este punto nos falta, todavía, otro elemento: la varianza del término de perturbación  $\sigma_u^2 = \text{var}(u)$ , que es básica en el momento de controlar los resultados obtenidos, ya que mide la dispersión de la parte aleatoria del modelo. La cuestión es que es imposible su cálculo si tenemos en cuenta que  $u_i$  no es una variable observable como lo son  $X_{2i}, X_{3i}, \dots, X_{ki}$ .

La perturbación  $u_i$  es una variable latente que está presente, que está en el modelo, pero que no podemos observar. En todo caso, nos podríamos acercar a partir de los errores o de los residuos del ajuste  $e_i$ ; podríamos demostrar que:

$$\hat{\sigma}^2 = \frac{SCE}{n - k}$$

es el estimador más adecuado de  $\sigma_u^2$ .

Ahora ya podemos calcular las varianzas de los estimadores de los parámetros o, por lo menos, estimar las de la relación:

$$\begin{pmatrix} \text{var}(\hat{\beta}_1) \\ \text{var}(\hat{\beta}_2) \\ \dots \\ \text{var}(\hat{\beta}_k) \end{pmatrix} = \hat{\sigma}^2 \text{diag} (X'X)^{-1}$$

donde  $\text{diag} (X'X)^{-1}$  contiene los elementos de la diagonal principal de la matriz  $(X'X)^{-1}$ .

A partir de aquel cálculo se obtienen las desviaciones estándar de los estimadores de la siguiente manera:

$$S_{\hat{\beta}_j} = \sqrt{\text{var}(\hat{\beta}_j)} \quad j = 1, 2, \dots, k$$

Como ya hemos dicho, la estimación del modelo va más allá de la simple estimación de los parámetros  $\beta_j$ ; es necesario obtener medidas que faciliten la lectura de los resultados y el análisis posterior. Con los ingredientes encontrados hasta ahora, la forma más habitual de presentar el modelo en la literatura estadística sería:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_{Ki} X_{Ki}$$

$$\begin{pmatrix} S_{\hat{\beta}_1} \\ S_{\hat{\beta}_2} \\ S_{\hat{\beta}_3} \\ S_{\hat{\beta}_k} \end{pmatrix} \quad \hat{\sigma}^2 \quad R^2$$

Si profundizásemos en el análisis de la regresión,...

... veríamos que hay más estadísticos que deberíamos incluir en la presentación de los resultados, pero para el objetivo de este curso es suficiente el detalle al cual se ha llegado.

Continuando con el modelo que permite explicar los ingresos de los ingenieros, podemos calcular la estimación de la varianza del término de perturbación:

$$\hat{\sigma}^2 = \frac{SCE}{n-K} = \frac{359,3}{5-3} = 179,65$$

o la estimación de la desviación estándar de las perturbaciones  $\hat{\sigma} = 13,4$ , que también se utiliza mucho.

Por lo que respecta a la estimación de las varianzas de  $\hat{\beta}$ , tendremos:

$$\begin{pmatrix} \text{var}(\hat{\beta}_1) \\ \text{var}(\hat{\beta}_2) \\ \text{var}(\hat{\beta}_3) \end{pmatrix} = \hat{\sigma}^2 \text{diag} (X'X)^{-1} =$$

$$= 179,65 \begin{pmatrix} 76,6505 & & \\ & 0,2450 & \\ & & 0,0613 \end{pmatrix} = \begin{pmatrix} 13.770,2623 & & \\ & 44,0143 & \\ & & 11,0125 \end{pmatrix}$$

y las desviaciones estándar:

$$S_{\hat{\beta}_1} = \sqrt{13.770,26} = 117,346$$

$$S_{\hat{\beta}_2} = \sqrt{44,0143} = 6,635$$

$$S_{\hat{\beta}_3} = \sqrt{11,0125} = 3,318$$

lo cual nos permite presentar la estimación del modelo de la manera habitual:

$$\hat{Y}_i = 70,8782 + 41,2649 X_{2i} + 3,5928 X_{3i}$$

$$\begin{matrix} (117,346) & (6,635) & (3,318) \\ \sigma = 179,65 & & R^2 = 0,976 \end{matrix}$$

### Actividad

1.4. Suponed que durante ocho sábados de la temporada de invierno hemos observado las variaciones que vemos a continuación en una estación de esquí:

$Y$  = cantidad de forfaits vendidos,

$X_2$  = grosor de nieve en la cota de 1.800 metros, medido en centímetros,

$X_3$  = porcentaje de remontes en funcionamiento ,

y que se han obtenido los siguientes resultados :

Sáb.	1.º	2.º	3.º	4.º	5.º	6.º	7.º	8.º
$Y$	3.030	4.770	4.630	5.910	5.860	2.090	5.670	6.090
$X_2$	110	180	150	200	200	100	190	220
$X_3$	50	90	100	100	100	30	100	100

Estamos interesados en estimar un modelo de regresión lineal que permita explicar el número de forfaits vendidos según la cantidad de nieve que hay en las pistas (cota 1.800) y las instalaciones que están abiertas; por eso, especificamos la relación:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + u_i$$

Comprobad que los cálculos matriciales necesarios para la solución del problema conducen a:

$$X'X = \begin{pmatrix} 8 & 1.350 & 670 \\ 1.350 & 241.500 & 120.700 \\ 670 & 120.700 & 61.500 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 2,30175 & -0,01749 & 0,00925 \\ -0,01749 & 0,00035 & -0,00050 \\ 0,00925 & -0,00050 & 0,00089 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 38,050 \\ 6.866,500 \\ 3.459,500 \end{pmatrix} \quad Y'Y = (19,624 \quad 3,504)$$

y que la estimación del modelo es:

$$\hat{Y}_i = -513,3 + 20,562X_{2i} + 21,488X_{3i}$$

$$(335,2) \quad (4,132) \quad (6,586)$$

$$\hat{\sigma} = 48,829 \quad R^2 = 0,984$$

### 1.7. Inferencia sobre los parámetros del modelo

Los cálculos efectuados hasta ahora responden a una estimación puntual del modelo y, en este momento, podría ser interesante ir un poco más allá con el fin de buscar intervalos de confianza para el comportamiento de los parámetros o bien para probar la validez de determinados valores hipotéticos.

La relación que hemos encontrado entre la variable endógena Y y las variables exógenas  $X_2, X_3, \dots, X_K$  es el resultado de haber trabajado con la información parcial que proporciona la muestra; lógicamente, los coeficientes obtenidos no coinciden necesariamente con los verdaderos coeficientes de la relación lineal.

Si la muestra ha sido bien elegida y el método de estimación es correcto,  $\hat{\beta}$  y  $\beta$  se parecerían más o menos, pero difícilmente coincidirían.

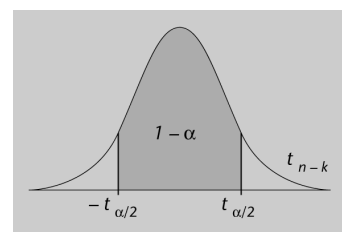
#### Una situación concreta

Un modelo para determinar el ahorro de las familias catalanas según los ingresos, el tipo de interés, etc., que se intenta estimar a partir de una muestra de 200 familias probablemente no coincidirá con el resultado obtenido después de haber analizado todas las familias del país; pero en cambio sí que podremos buscar, con un nivel de confianza prefijado, intervalos para los parámetros poblacionales. Tal como lo hicimos cuando queríamos estimar una media, una proporción o la pendiente de la recta de regresión, tendríamos que partir de un estadístico de prueba del cual derivarán la estimación del intervalo o la contrastación de hipótesis.

Supongamos para el caso de los estimadores de los parámetros de un modelo de regresión lineal  $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_K)$  el estadístico de distribución t de Student con n (tamaño de la muestra) menos K (número de parámetros de la ecuación del modelo) grados de libertad:

$$\frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{n-K} \quad \text{con } j = 1, 2, \dots, K$$

En la distribución t de Student es posible encontrar dos puntos centrados en el valor cero y, por tanto, opuestos en el signo,  $\pm t_{\alpha/2}$ , que acoten un área predeterminada o un nivel de confianza  $1 - \alpha$  (gráfico al margen).



Puesto que el estadístico mencionado presenta aquella distribución, existe una probabilidad de  $1 - \alpha$  de que tome valores entre aquellos dos puntos:

$$P\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \leq t_{\alpha/2}\right] = 1 - \alpha$$

donde es posible aislar el parámetro poblacional  $\beta_j$  de la manera:

$$P[\hat{\beta}_j - t_{\alpha/2} S_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} S_{\hat{\beta}_j}] = 1 - \alpha$$

lo cual nos permite afirmar que  $\beta_j$  (coeficiente de regresión entre  $Y_i$  y  $X_j$  en el modelo) es desconocido, pero que tenemos un grado  $1 - \alpha$  de confianza de que pertenece al intervalo que tenemos aquí:

$$\beta_j \in \hat{\beta}_j \pm t_{\alpha/2} S_{\hat{\beta}_j} \quad j = 1, 2, \dots, k$$

El margen de error o la precisión de la estimación cambia  $\pm t_{\alpha/2} S_{\hat{\beta}_j}$  de acuerdo con el nivel de confianza y con la desviación estándar de  $\hat{\beta}_j$ . No es lo mismo hacer una estimación que tenga una fiabilidad (confianza) del 80% que otra del 99%; cuanto más alto sea el nivel de confianza, más recorrido tendremos que contemplar para los posibles valores de  $\beta_j$ :

$$\Delta(1 - \alpha) \rightarrow \Delta t_{\alpha/2} \rightarrow \Delta \text{margen de error}$$

De la misma forma, tenemos que entender que, cuanto menos eficiente sea la estimación, es decir, cuanto mayor sea la desviación estándar  $\hat{\beta}_j$ , mayor será también el margen de error de la estimación:

$$\Delta S_{\hat{\beta}_j} \rightarrow \Delta \text{margen de error}$$

Ejemplo

Si partimos de la estimación efectuada con los datos que se proporcionaban en la actividad de la estación de esquí:

$$\hat{Y}_i = -513,3 + 20,562X_{2i} + 21,488 X_{3i}$$

Intuitivamente,...

... también podríamos llegar a la conclusión de que, en igualdad de condiciones, las muestras más grandes proporcionan intervalos de precisión mayores (es decir, intervalos menores).



donde recordamos que:

$Y_i$  = cantidad de forfaits vendidos,

$X_{2i}$  = grosor de nieve en la cota de 1.800 metros, medido en centímetros,

$X_{3i}$  = porcentaje de remontes en funcionamiento,

queremos estimar, al 95% de confianza, un intervalo para  $\beta_2$ , es decir, para el incremento de esquiadores a que da lugar un centímetro más de nieve en las pistas.

Con una muestra igual a 8 ( $n = 8$ ) y tres parámetros por estimar ( $K = 3$ ), buscamos los puntos críticos en la ley t de Student de  $(n - K) = 5$  grados de libertad y nivel de confianza de  $1 - \alpha = 0,95$  (nivel de significación  $\alpha = 0,05$ ). Estos valores son:

$$\pm t_{\alpha/2} = \pm 2,571$$

El intervalo será, por lo tanto:

$$\beta_2 \in 20,562 \pm 2,571 \times 4,132$$

lo cual implica, finalmente, el intervalo:

$$9,939 \leq \beta_2 \leq 31,185$$

resultado que hay que interpretar como pronóstico del aumento de forfaits vendidos por cada centímetro más de nieve y que tiene un 95% de probabilidad de ser acertado.

### Actividad

1.5. También podéis estimar el intervalo de confianza al 95% por el parámetro  $\beta_3$ , coeficiente que relaciona Y con  $X_3$ . Comprobad, para finalizar, que éste es:

$$4,555 \leq \beta_3 \leq 31,185$$

A partir del estadístico:

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

también podemos efectuar contrastes sobre los parámetros:

1) Ya sea en caso de hipótesis simples con solución bilateral: 

$$\begin{aligned} H_0 : \beta_j &= \beta_j^* \\ H_1 : \beta_j &\neq \beta_j^* \\ \frac{\hat{\beta}_j - \beta_j^*}{S_{\hat{\beta}_j}} &\notin (-t_{\alpha/2} ; t_{\alpha/2}) \rightarrow \text{rechazo de } H_0 \end{aligned}$$

2) O bien en caso de hipótesis compuestas con pruebas unilaterales:

$$\begin{array}{l}
 H_0 : \beta_j = \beta_j^* \\
 H_1 : \beta_j < \beta_j^* \\
 \frac{\hat{\beta}_j - \beta_j^*}{S_{\hat{\beta}_j}} < -t_\alpha \rightarrow \text{rechazo de } H_0 \\
 \\
 H_0 : \beta_j = \beta_j^* \\
 H_1 : \beta_j > \beta_j^* \\
 \frac{\hat{\beta}_j - \beta_j^*}{S_{\hat{\beta}_j}} > t_\alpha \rightarrow \text{rechazo de } H_0
 \end{array}$$

En este caso,  $t_\alpha$  es el punto crítico que determina una sola cola con área  $\alpha$  para el rechazo de la hipótesis nula.

Ejemplo

Imaginaos que se ha diseñado un modelo de demanda para un producto  $q_Y$  de acuerdo con el precio del producto  $P_Y$ , con el precio de otro producto  $P_X$  y con un nivel de renta disponible  $RD$ :

$$q_Y = f(P_Y, P_X, RD)$$

y que el tratamiento estadístico del modelo aconseja tomar las variables en logaritmos de manera que :

$$\ln q_Y = \beta_1 + \beta_2 \ln P_Y + \beta_3 \ln P_X + \beta_4 \ln RD + u$$

donde ahora  $\ln q_Y$  es la variable endógena e  $\ln P_Y$ ,  $\ln P_X$  y  $\ln RD$ , las variables exógenas; por lo que respecta a los coeficientes, tienen la lectura económica que encontramos a continuación:

$\beta_2$  = elasticidad-precio directa de la demanda,

$\beta_3$  = elasticidad-precio cruzada de la demanda,

$\beta_4$  = elasticidad-renta de la demanda.

Suponed que después de observar veinticuatro zonas de mercado hemos llegado a la estimación:

$$\begin{array}{cccc}
 \hat{\ln q}_Y = 2,243 + 0,767 \ln P_Y - 0,426 \ln P_X + 0,964 \ln RD \\
 (0,032) \quad (0,060) \quad (0,105) \quad (0,118) \\
 \hat{\sigma}^2 = 21,431 \quad R^2 = 0,886
 \end{array}$$

Investigación  
econométrica

Existe una cierta unanimidad a la hora de fijar los inicios de la investigación econométrica en el análisis de la demanda de bienes de consumo.

H.L. Moore, M. Ezaquiel, E. Working y M. Schultz, entre otros, son los que contribuyeron más, en los años treinta, a estos tipos de estudios.

Si operamos con una significación del 5% (95% de confianza), nos planteamos cuestiones relativas a la función de la demanda a partir de sus elasticidades.

En primer lugar, queremos saber si se trata de un producto de demanda absolutamente rígida (elasticidad-precio directa de la demanda igual a cero), hecho que tenemos que rechazar porque:

$$\begin{aligned} H_0 : \beta_2 &= 0 \\ H_1 : \beta_2 &\neq 0 \\ \frac{\hat{\beta}_2 - \beta_2^*}{S_{\hat{\beta}_2}} &= \frac{0,767 - 0}{0,060} = 12,783 \notin \pm 2,086 \end{aligned}$$

donde 2,086 es el punto crítico  $t_{\alpha/2}$  encontrado en una t de Student de 20 grados de libertad y una significación del 5%.

También pensamos que, a pesar del resultado obtenido ( $\hat{\beta}_3 < 0$ ), los dos productos no son complementarios, es decir, que incrementos de  $P_x$  no implican disminuciones en la demanda  $q_y$ ; por tanto, la elasticidad-precio cruzada es positiva o nula ( $\beta_3 \geq 0$ ):

$$\begin{aligned} H_0 : \beta_3 &\geq 0 \\ H_1 : \beta_3 &< 0 \\ \frac{\hat{\beta}_3 - \beta_3^*}{S_{\hat{\beta}_3}} &= \frac{-0,426 - 0}{0,105} = -4,057 < -1,725 \end{aligned}$$

en que  $-t_{\alpha} = -1,725$  en la prueba en una sola cola. Por tanto, tenemos que rechazar la convicción de que son productos sustantivos.

Para finalizar, queremos saber si la demanda se puede comportar como un bien de lujo atendiendo a la dicotomía de elasticidades-venta de la demanda:

elasticidad<sub>y</sub> ≤ 1 → producto de primera necesidad  
elasticidad<sub>y</sub> ≥ 1 → producto de lujo

$$\begin{aligned} H_0 : \beta_4 &\geq 1 \\ H_1 : \beta_4 &< 1 \\ \frac{\hat{\beta}_4 - \beta_4^*}{S_{\hat{\beta}_4}} &= \frac{0,964 - 1}{0,118} = -0,305 > -1,725 \end{aligned}$$

Este resultado no nos permite rechazar una elasticidad mayor o igual a la unidad, por lo cual no se puede descartar que se trate de un producto de lujo.


## 1.8. Contrastes sobre la capacidad explicativa de las variables

Construir un modelo puede ser más o menos laborioso, pero siempre nos podemos inspirar en las leyes económicas, en las financieras, etc., para especi-

car relaciones funcionales entre variables. Incluso, podemos ir a buscar modelos analizados por otros investigadores en ámbitos similares y, en el peor de los casos, siempre queda el recurso de improvisar un modelo con un mínimo de coherencia.

Sin embargo, cualquiera que sea la fuente de especificación empleada, es inexcusable validar de forma estadística el modelo y comprobar la capacidad explicativa de todas las variables que hemos incorporado.

Dejemos el análisis conjunto del modelo para más adelante, y centrémonos en la discusión individual de las variables pretendidamente explicativas. Un contraste de hipótesis sencillo puede resolver este problema.

Si el coeficiente  $\beta_j$  que vincula la relación entre  $X_j$  e  $Y$  puede ser cero, quiere decir que  $X_j$  puede no tener ningún efecto sobre  $Y$ ; entonces, diremos que  $X_j$  no es una variable explicativa en el modelo. 

Basta con presentar las hipótesis alternativas para todos los parámetros funcionales  $\beta_2, \beta_3, \dots, \beta_K$ , ( $\beta_1$  es el término independiente y no se encuentra asociado a ninguna otra variable):

$$\begin{aligned} H_0 : \beta_j &= 0 \text{ (} X_j \text{ no es explicativa)} \\ H_1 : \beta_j &\neq 0 \text{ (} X_j \text{ es explicativa)} \end{aligned}$$

Recordad que...

rechazar  $H_0$  quiere decir que la variable  $X_j$  es explicativa en el modelo.

La hipótesis simple tiene respuesta a partir del estadístico del contraste en una prueba de dos colas:

$$\frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \notin \pm t_{\alpha/2} \rightarrow \text{rechazamos } H_0$$

o, más cómodamente todavía, con el punto crítico de signo positivo de la  $t$  de Student:

$$\left| \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \right| > t_{\alpha/2} \rightarrow \text{rechazamos } H_0$$

## Actividad


1.6. Si recuperáis el ejemplo donde pretendíamos explicar los ingresos de los ingenieros según los años que hace que acabaron la carrera y según la cantidad de horas trabajadas por semana, veréis que la estimación obtenida a partir de una muestra de cinco observaciones era:

$$\begin{aligned} \hat{Y}_1 &= 70,8782 + 41,2649X_{2i} + 3,5928X_{3i} \\ &\quad (117,346) \quad (6,635) \quad (3,318) \\ \hat{\sigma}^2 &= 179,65 \quad R^2 = 0,976 \end{aligned}$$

¿Podemos afirmar que  $X_2$  es una variable explicativa en este modelo? ¿Y  $X_3$ ?  
Comprobad que, con una significación del 5%, la respuesta es positiva en el primer caso y negativa en el segundo.

Recordemos también que la distribución t de Student es asintóticamente normal; esto quiere decir que con muestras grandes (muchos grados de libertad) se puede utilizar perfectamente la ley  $N(0,1)$  para generar los puntos críticos. Así, con muestras de más de 50 observaciones, como pueden ser las que se utilizan en la mayoría de los estudios de mercado, en los controles de calidad, en las encuestas de opinión, etc., y con un nivel de confianza del 95%, que es el habitual en estos casos, el punto crítico sería, de manera aproximada, 2 ( $t_{\alpha/2} = 1,96$ ). Esto permite analizar la significación de las variables sin tener que emplear las tablas según si el estimador  $j$  toma, para la muestra con la cual se trabaja, un valor de más del doble de su desviación estándar  $\hat{\beta}_j$  o no lo hace.

### 1.9. Contrastación conjunta del modelo

Acabamos de ver si de forma individual las variables exógenas del modelo contribuyen a explicar  $Y$ . Lógicamente, si se detectan variables explicativas, el modelo también lo será. Sin embargo, no tiene que acontecer necesariamente lo contrario: un modelo puede pasar positivamente un contraste de validación y quizá ninguna de las variables sea relevante individualmente; hay muchos factores que pueden explicar esta paradoja, pero ahora no entraremos en ello. 

El hecho de demostrar si un modelo es o no explicativo implica plantear la hipótesis nula y la hipótesis alternativa:

$$H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1: \exists \beta_j \neq 0$$

donde  $H_0$  significa que el modelo, en conjunto, no es explicativo.

El proceso de contrastación se basa en la misma técnica del análisis de la varianza que reasigna la variación total (SCT) a dos componentes: la variación explicada (SCR) y la variación no explicada o residual (SCE); resulta de aquí un estadístico F de Snedecor para el cumplimiento de la prueba.

El siguiente cuadro reúne los cálculos necesarios para la obtención del estadístico de prueba y para la contrastación posterior.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media
$X_2, X_3, \dots, X_k$	SCR	$k-1$	$SCR/(k-1)$
e	SCE	$n-k$	$SCE/(n-k)$
Y	SCT	$n-1$	

¿Hasta qué punto el grupo de variables exógenas es suficientemente importante en la explicación de Y? La respuesta deriva del test siguiente:

$$\frac{\frac{SCR}{K-1}}{\frac{SCE}{n-K}} > F_{\alpha} \rightarrow \text{rechazamos } H_0 \text{ (modelo explicativo)}$$

donde  $F_{\alpha}$  es el punto crítico de la ley F de Snedecor con  $(K-1)(n-K)$  grados de libertad, que deja una cola superior con el área  $\alpha$ :

Para acabar el ejemplo de los ingresos de los ingenieros, donde entre otros cálculos teníamos:

$$n = 5 \quad K = 3 \\ SCT = 14.839,4 \quad SCR = 14.480,1 \quad SCE = 359,3$$

elaboramos el cuadro siguiente de análisis de la varianza a partir del cual podremos obtener el estadístico para la contrastación:

Fuente de variación	Suma de cuadrados	Grados de libertad	Media
$X_2, X_3,$ e	14.480,1 359,3	2 2	7.240,05 179,65
Y	14.839,4	4	

$$\frac{\frac{SCR}{K-1}}{\frac{SCE}{n-K}} = \frac{7.240,05}{179,65} = 40,3$$

Por otro lado, la distribución F con 2,2 grados de libertad proporciona el punto crítico  $F_{\alpha} = 19$  con una significación del 5%, y puesto que  $40,3 > 19$ , tenemos que rechazar la hipótesis nula de no-significación conjunta del modelo, lo cual no quiere decir que cada una de las variables que lo integren sea relevante.

### Actividades

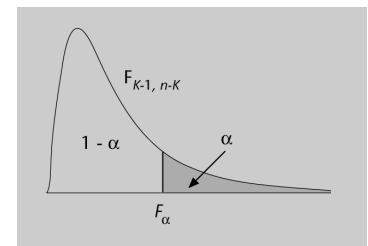
1.7. Justificad de manera analítica la siguiente identidad que tiene que permitir utilizar de manera indistinta una u otra expresión para los contrastes de relevancia conjunta del modelo.

$$\frac{\frac{SCR}{K-1}}{\frac{SCE}{n-K}} = \frac{R^2(n-K)}{(1-R^2)(K-1)}$$

Demostrad, en el ejemplo que acabamos de resolver, cómo ambas fórmulas proporcionan el mismo resultado (40,3).

1.8. Con el objetivo de validar como indicadores del desarrollo económico de una región –aproximado por su renta disponible (YD)– los siguientes: número de entidades de crédito (bancos), inversión industrial anual (inv. ind.) y densidad de población (dens. de pobl.), se ha estimado este modelo de regresión:

$$YD = \beta_1 + \beta_2 \text{ bancos} + \beta_3 \text{ inv. ind.} + \beta_4 \text{ dens. de pobl.} + u$$



con datos referentes a las cuarenta y una comarcas del Principado.

Los resultados han sido:

Regresor	$\hat{\beta}_j$	$S_{\hat{\beta}}$
Constante	- 8,19	3,39
Bancos	0,52	0,04
Inv. ind.	8,08	1,19
Dens. tel.	0,02	0,01

Análisis de la varianza	
Fuente de variación	Sumas
Modelo	3.647.808,0
Residuos	10.380,8
Total	3.658.189,0

Calculad las medidas necesarias para rechazar o no la significación individual de cada uno de los regresores y la bondad conjunta del modelo. Razonad también los signos de las estimaciones de los parámetros.

La solución a la cual llegaréis será:

Regresor	Estadístico t-Student
Constante	- 2,42
Bancos	13,00
Inv. ind.	6,79
Dens. de pobl.	2,00

Análisis de la varianza			
Fuente de variación	Sumas	Grados de libertad	Medias
Modelo	3.647.808,0	3	1.215.936,0
Residuos	10.380,8	37	280,562
Total	3.658.189,0	40	4.333,93

$$R^2 = 0,997$$

Observad que los tres coeficientes son estadísticamente significativos: estadísticamente diferentes de cero. Puesto que  $t_{\hat{\beta}_j} > 2$ , tenemos que rechazar la hipótesis nula  $H_0: \beta_j = 0$ . Para analizar la significación global del modelo, debemos recurrir al estadístico F. Teniendo en cuenta el valor que toma, muy superior al valor crítico que daría una distribución  $F_{3,37} = F_{k-1, n-k}$ ; también tenéis que rechazar la hipótesis no-significativa conjunta. El coeficiente de determinación  $R^2$  muestra, igualmente, un ajuste de los datos al modelo que se considera.

En última instancia, hemos visto que todos los coeficientes estimados presentan signo positivo, lo cual quiere decir que existe una relación positiva entre la variable que hay que explicar, la renta disponible de la comarca y las variables explicativas: entidades bancarias, inversión y densidad de población. Por tanto, un incremento en alguna de estas variables sería un indicador de un incremento del nivel de renta disponible de una región.

1.9. Los datos siguientes corresponden al nivel de producción Q en millones de unidades, al número de trabajadores L, a la cifra de inversiones K en miles de unidades monetarias y a los logaritmos neperianos de estas variables, sobre una muestra de diez empresas de media dimensión de una actividad industrial:

	Q	L	K	lnQ	lnL	lnK
1	9,80	57	107	2,28238	4,04305	4,67283
2	10,55	59	117	2,35613	4,07754	4,76217
3	12,42	67	138	2,51931	4,20469	4,92725
4	18,11	75	211	2,89646	4,31749	5,35186
5	10,78	61	120	2,37769	4,11087	4,78749
6	11,00	59	125	2,39790	4,07754	4,82831
7	17,82	68	203	2,88032	4,21951	5,31321
8	13,43	49	197	2,59749	3,89182	5,28320
9	6,89	38	86	1,93007	3,63759	4,45435
10	9,51	45	95	2,25234	3,80666	4,55388

Nos planteamos la estimación del modelo que permite explicar el nivel de producción a partir de otras dos variables, en primer lugar, especificando una relación lineal del tipo:

$$Q = \beta_1 + \beta_2 L + \beta_3 K + u$$

y, en segundo lugar, en un intento de mejorar la capacidad explicativa del modelo por medio de la relación linealizada

$$\ln Q = \ln A + \gamma \ln L + \lambda \ln K + v$$

Haced las dos estimaciones de forma completa para poder comparar ambos modelos. Os tienen que salir al final los siguientes resultados:

En el caso del primer modelo:

$$\hat{Q} = - 1,97 + 0,104 L + 0,0572 K$$

Predictor	Coef.	Desv. est	Estadístico t
Constante	- 1,972	1,443	- 1,37
L	0,10373	0,03125	3,32
K	0,057237	0,00753	7,59

$$\hat{\sigma} = 0,8253 \quad R\text{-cuadrado} = 95,9\%$$

Fuente	GL	SC	Media	F
Regresión	2	110,392	55,196	81,04
Error	7	4,768	0,681	
Total	9	115,159		

Y para el segundo modelo:

$$\ln \hat{Q} = - 2,66 + 0,449 \ln L + 0,673 \ln K$$

Predictor	Coef.	Desv. est	Estadístico t
Constante	- 2,6598	0,4387	- 6,06
lnL	0,4492	0,1418	3,17
lnK	0,67324	0,09102	7,40

$$\hat{\sigma} = 0,06625 \quad R\text{-cuadrado} = 96,0\%$$

Fuente	GL	SC	Media	F
Regresión	2	0,73458	0,36729	83,70
Error	7	0,03072	0,00439	
Total	9	0,76530		

La función de producción...

... de Cobb-Douglas:  
 $Q = A L^\gamma K^\lambda$   
 se utiliza mucho en economía para vincular el producto final con los factores de producción, gracias a las propiedades operativas que poseen y a la buena adherencia que presenta tanto en datos microeconómicos como en datos macroeconómicos.

La suma de sus coeficientes ( $\gamma + \lambda$ ) determina el grado de homogeneidad de la función y permite distinguir si la empresa o la actividad tiene economías o diseconomías de escala.

Rendimientos crecientes:  
 $\gamma + \lambda > 1$

Rendimientos constantes:  
 $\gamma + \lambda = 1$

Rendimientos decrecientes:  
 $\gamma + \lambda < 1$



## Actividad

1.10. A continuación tenéis datos correspondientes a los diez distritos de Barcelona (Ciutat Vella, Eixample, Sants/Montjuic, Las Corts, Sarriá/S. Gervasio, Gracia, Horta/Guinardó, Nou Barris, San Andrés y San Martín), con las siguientes variables:

Y – Precio medio de venta de viviendas de primera mano (u.m./m<sup>2</sup>).

X<sub>1</sub> – Verde urbano/superficie.

X<sub>2</sub> – Oficinas de La Caixa por cada 1.000 habitantes.

X<sub>3</sub> – Escuelas por cada 1.000 habitantes.

X<sub>4</sub> – Tasa de paro (parados/población activa).

Se trata de analizar si existe una relación de dependencia entre la primera de las variables, el precio de las vivienda y las cuatro restantes, que tienen que ver, respectivamente, con los espacios verdes, con el nivel de servicios y con el estatus socioeconómico de los distritos:

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
210.684	0,13	0,530	0,386	0,162
244.061	0,06	0,489	0,350	0,090
213.413	0,13	0,318	0,412	0,122
266.222	0,09	0,468	0,412	0,077
368.519	0,03	0,513	0,688	0,067
253.706	0,08	0,373	0,443	0,099
214.435	0,06	0,293	0,477	0,112
204.205	0,10	0,281	0,408	0,133
193.401	0,07	0,297	0,400	0,116
199.646	0,13	0,369	0,350	0,129

Los resultados que encontraréis muestran que la primera variable explicativa –los espacios verdes– no es significativa, mientras que las otras sí que lo son. Por otro lado, el modelo es significativo en su conjunto. Hay que destacar el signo negativo del parámetro que acompaña la variable referida a la tasa de paro e indicar que en los distritos con un paro más elevado el nivel socioeconómico es más bajo y los precios de las viviendas son más bajos.