



Estadística para Administración y Economía  
Autores: Newbold, P., Carlson, W. & Thorne, B.  
Pearson Prentice Hall, Madrid, 2008, págs. 431-496  
ISBN: 978-84-8322-403-8

Esta obra está protegida por el derecho de autor y su reproducción y comunicación pública, en la modalidad puesta a disposición, se han realizado con autorización de CEDRO. Queda prohibida su posterior reproducción, distribución, transformación y comunicación pública en cualquier medio y de cualquier forma, con excepción de una única reproducción mediante impresora por cada usuario autorizado.

## Regresión simple

### Esquema del capítulo

- 12.1. Análisis de correlación  
Contraste de hipótesis de la correlación
- 12.2. Modelo de regresión lineal
- 12.3. Estimadores de coeficientes por el método de mínimos cuadrados  
Cálculo por computador del coeficiente de regresión
- 12.4. El poder explicativo de una ecuación de regresión lineal  
El coeficiente de determinación  $R^2$
- 12.5. Inferencia estadística: contrastes de hipótesis e intervalos de confianza  
Contraste de hipótesis del coeficiente de la pendiente poblacional utilizando la distribución  $F$
- 12.6. Predicción
- 12.7. Análisis gráfico

### Introducción

Hasta ahora hemos centrado la atención en el análisis y la inferencia relacionados con una única variable. En este capítulo extendemos nuestro análisis a las relaciones entre variables. Comenzamos con una breve introducción al análisis de correlación, seguido de la presentación del análisis de regresión simple. Nuestra presentación es paralela a la del Capítulo 3, en el que hicimos hincapié en las relaciones descriptivas, incluido el uso de diagramas de puntos dispersos, coeficientes de correlación y la regresión lineal como instrumentos para describir las relaciones entre variables. Suponemos que el lector está familiarizado con ese capítulo.

En el análisis de los procesos empresariales y económicos se utilizan a menudo las relaciones entre variables. Estas relaciones se expresan en términos matemáticos de la forma siguiente:

$$Y = f(X)$$

donde la función puede adoptar muchas formas lineales y no lineales. En algunos de esos casos, la forma de la relación no se conoce exactamente. Aquí presentamos análisis que se basan en relaciones lineales. En muchos casos, las relaciones lineales constituyen un buen modelo del proceso. En otros casos, nos interesa una parte limitada de una relación no lineal a la que podemos aproximarnos mediante una relación lineal. En el apartado 13.7 mostramos que algunas relaciones no lineales importantes también pueden analizarse utilizando el análisis de regresión. Por lo tanto, los métodos de correlación y de regresión pueden aplicarse a una amplia variedad de problemas.

Las relaciones lineales son muy útiles para muchas aplicaciones empresariales y económicas, como indican los siguientes ejemplos. El presidente de Materiales de Construcción, S.A., fabricante de placas de yeso, cree que la cantidad anual media de placas de edificación expedidos durante el año anterior. Un vendedor de cereales quiere saber cómo afecta la producción total al precio por tonelada. Está desarrollando un modelo de predicción que utiliza datos históricos. El departamento de marketing necesita saber cómo afecta el precio de la gasolina a sus ventas totales. Utilizando datos semanales sobre los precios y las ventas, planea desarrollar un modelo lineal que muestre cuánto varían las ventas cuando varía el precio.

Con la aparición de muchos y buenos paquetes estadísticos y hojas de cálculo como Excel, hoy es posible para casi todo el mundo calcular estadísticos de correlación y de regresión. Desgraciadamente, también sabemos que no todo el mundo sabe interpretar y utilizar correctamente estos resultados obtenidos por computador. Aquí el lector aprenderá algunas ideas fundamentales que lo ayudarán a utilizar el análisis de regresión. Comenzaremos examinando el análisis de correlación.

## 12.1. Análisis de correlación

---

En este apartado utilizamos los coeficientes de correlación para estudiar las relaciones entre variables. En el Capítulo 3 utilizamos el coeficiente de correlación muestral para describir la relación entre variables indicada en los datos. En el 5 y en el 6 aprendimos lo que era la correlación poblacional. Aquí presentamos métodos inferenciales que utilizan el coeficiente de correlación para estudiar relaciones lineales entre variables.

En principio, dos variables aleatorias pueden estar relacionadas de diversas formas. Es útil postular al comienzo del análisis una forma funcional de su relación. A menudo es razonable suponer, como buena aproximación, que la relación es lineal. Si se examina un par de variables aleatorias,  $X$  e  $Y$ , entre las que existe una relación lineal, en un diagrama de puntos dispersos las observaciones conjuntas sobre este par de variables tenderán a estar concentradas en torno a una línea recta. Y a la inversa, si no existe una relación lineal, no estarán concentradas en torno a una línea recta. No todas las relaciones que estudiaremos estarán muy concentradas en torno a una línea recta. El diagrama de puntos dispersos de muchas relaciones importantes muestra una tendencia hacia una relación lineal, pero con una considerable desviación con respecto a una línea recta. En los diagramas de puntos dispersos del Capítulo 2 vimos algunos ejemplos.

Las correlaciones tienen muchas aplicaciones en el mundo de la empresa y en la economía. En muchos problemas económicos aplicados, afirmamos que hay una variable independiente o exógena  $X$ , cuyos valores son determinados por actividades realizadas fuera del sistema económico examinado y que hay una variable dependiente o endógena  $Y$ , cuyo valor depende del valor de  $X$ . Si preguntamos si las ventas aumentan cuando bajan los precios, estamos analizando una situación en la que un vendedor ajusta de una forma deliberada e independiente los precios en sentido ascendente o descendente y observa cómo varían las ventas. Supongamos ahora que los precios y las cantidades vendidas son el resultado de equilibrios de la oferta y la demanda como propone el modelo económico básico. En ese caso, podríamos analizar los precios y las cantidades como variables aleatorias y preguntarnos si estas dos variables aleatorias están relacionadas entre sí. El coeficiente de correlación puede utilizarse para averiguar si existe una relación entre variables en cualquiera de estas dos situaciones.

Supongamos que tanto  $X$  como  $Y$  son determinados simultáneamente por factores que se encuentran fuera del sistema económico analizado. Por lo tanto, suele ser más realista plantear un modelo en el que tanto  $X$  como  $Y$  sean variables aleatorias. En el Capítulo 5 presentamos el coeficiente de correlación  $\rho_{xy}$  como medida de la relación entre dos variables aleatorias,  $X$  e  $Y$ . En esos casos, utilizamos el coeficiente de correlación poblacional,  $\rho_{xy}$ , para indicar la existencia de una relación lineal sin que ello quisiera decir que una de las variables era independiente y la otra dependiente. En las situaciones en las que una de las variables es dependiente lógicamente de otra, el siguiente paso lógico después del análisis de correlación es la utilización del análisis de regresión para desarrollar el modelo lineal. Éste es el tema del siguiente apartado. Aquí presentamos métodos de inferencia estadística que utilizan correlaciones muestrales para averiguar las características de las correlaciones poblacionales.

## Contraste de hipótesis de la correlación

El coeficiente de correlación muestral

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

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

es una medida descriptiva útil de la fuerza de la relación lineal en una muestra. También podemos utilizar la correlación para contrastar la hipótesis de que no existe una relación lineal en la población entre un par de variables aleatorias; es decir,

$$H_0: \rho = 0$$

Esta hipótesis nula de que no existe una relación lineal entre un par de variables aleatorias es muy interesante en algunas aplicaciones. Cuando calculamos la correlación muestral a partir de datos, es probable que el resultado sea diferente de 0 aunque la correlación poblacional sea 0. Nos gustaría, pues, saber en qué medida debe ser diferente de 0 una correlación muestral para contar con una prueba de que la correlación poblacional no es 0.

Podemos demostrar que cuando la hipótesis nula es verdadera y las variables aleatorias siguen una distribución normal conjunta, la variable aleatoria

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

sigue una distribución  $t$  de Student con  $(n-2)$  grados de libertad. Las ecuaciones 12.1 a 12.3 muestran los contrastes de hipótesis adecuados.

### Contrastes de la correlación poblacional nula

Sea  $r$  el coeficiente de correlación muestral, calculado a partir de una muestra aleatoria de  $n$  pares de observaciones de una distribución normal conjunta. Los siguientes contrastes de la hipótesis nula

$$H_0: \rho = 0$$

tienen un valor de significación  $\alpha$ :

1. Para contrastar  $H_0$  frente a la hipótesis alternativa

$$H_1: \rho > 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2, \alpha} \quad (12.1)$$

2. Para contrastar  $H_0$  frente a la hipótesis alternativa

$$H_1: \rho < 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2, \alpha} \quad (12.2)$$

3. Para contrastar  $H_0$  frente a la hipótesis alternativa bilateral

$$H_1: \rho \neq 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2, \alpha/2} \quad \text{o} \quad \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2, \alpha/2} \quad (12.3)$$

Aquí,  $t_{n-2, \alpha}$  es el número para el que

$$P(t_{n-2} > t_{n-2, \alpha}) = \alpha$$

donde la variable aleatoria  $t_{n-2}$  sigue una distribución  $t$  de Student con  $(n-2)$  grados de libertad.

4. Si introducimos  $t_{n-2, \alpha/2} = 2,0$  en la ecuación 12.3, podemos demostrar que una «regla práctica» aproximada para contrastar la hipótesis anterior de que la correlación poblacional es 0 es

$$|r| > \frac{2}{\sqrt{n}}$$

### EJEMPLO 12.1. Valoración del riesgo político (contraste de hipótesis de la correlación)

Un equipo de investigación estaba intentando averiguar si el riesgo político existente en los países está relacionado con su inflación. En esta investigación, se realizó una encuesta a analistas del riesgo político que permitió elaborar una puntuación media del riesgo político de 49 países (los datos proceden del estudio mencionado en la referencia bibliográfica 2).

#### Solución

Cuanto más alta es la puntuación, mayor es el riesgo político. La correlación muestral entre la puntuación del riesgo político y la inflación de estos países era de 0,43.

Queremos averiguar si la correlación poblacional,  $\rho$ , entre estas medidas es diferente de 0. Concretamente, queremos contrastar

$$H_0: \rho = 0$$

frente a

$$H_1: \rho > 0$$

utilizando la información muestral

$$n = 49 \quad r = 0,43$$

El contraste se basa en el estadístico

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{0,43\sqrt{(49-2)}}{\sqrt{1-(0,43)^2}} = 3,265$$

Dado que hay  $(n-2) = 47$  grados de libertad, vemos en la tabla 8 de la  $t$  de Student del apéndice que

$$t_{47, 0,005} < 2,704$$

Por lo tanto, podemos rechazar la hipótesis nula al nivel de significación del 0,5 por ciento. Tenemos, pues, pruebas contundentes de que existe una relación lineal positiva entre la inflación y la valoración de los expertos del riesgo político de los países. Obsérvese que de este resultado no podemos extraer la conclusión de que una de las variables es la causa de la otra, sólo que están relacionadas.



Antes hemos señalado que la hipótesis nula  $H_0: \rho = 0$  puede rechazarse utilizando la regla práctica aproximada  $|r| > 2/\sqrt{n}$ . Este resultado proporciona un rápido contraste para averiguar si dos variables están relacionadas linealmente cuando se examinan una o más correlaciones muestrales. Así, por ejemplo, en el caso de una muestra de tamaño  $n = 25$ , el valor absoluto de la correlación muestral tendría que ser superior a  $2/\sqrt{25} = 0,40$ . Pero en el caso de una muestra de tamaño  $n = 64$ , el valor absoluto de la correlación muestral tendría que ser superior a  $2/\sqrt{64} = 0,25$  solamente. Se ha observado que este resultado es útil en muchas aplicaciones estadísticas.

## EJERCICIOS

### Ejercicios básicos

**12.1.** Dados los pares siguientes de  $(x, y)$  observaciones, calcule la correlación muestral.

- a) (2, 5), (5, 8), (3, 7), (1, 2), (8, 15).
- b) (7, 5), (10, 8), (8, 7), (6, 2), (13, 15).
- c) (12, 4), (15, 6), (16, 5), (21, 8), (14, 6).
- d) (2, 8), (5, 12), (3, 14), (1, 9), (8, 22).

**12.2.** Contraste la hipótesis nula

$$H_0: \rho = 0 \quad \text{frente a} \quad H_1: \rho \neq 0$$

dada

- a) Una correlación muestral de 0,35 en una muestra aleatoria de tamaño  $n = 40$
- b) Una correlación muestral de 0,50 en una muestra aleatoria de tamaño  $n = 60$

- c) Una correlación muestral de 0,62 en una muestra aleatoria de tamaño  $n = 45$   
 d) Una correlación muestral de 0,60 en una muestra aleatoria de tamaño  $n = 25$

- 12.3. El profesor de un curso de estadística puso un examen final y también pidió a los estudiantes que realizaran un proyecto. La tabla adjunta muestra las calificaciones de una muestra aleatoria de 10 estudiantes. Halle la correlación muestral entre las calificaciones del examen y las del proyecto.

Examen	81	62	74	78	93	69	72	83	90	84
Proyecto	76	71	69	76	87	62	80	75	92	79

### Ejercicios aplicados

- 12.4. En el estudio de 49 países analizado en el ejemplo 12.1, la correlación muestral entre la valoración del riesgo político realizada por los expertos y la tasa de mortalidad infantil de estos países era 0,75. Contraste la hipótesis nula de que no existe ninguna correlación entre estas cantidades frente a la hipótesis alternativa de que existe una correlación positiva.
- 12.5. En una muestra aleatoria de 353 profesores de enseñanza secundaria, se observó que la correlación entre las subidas salariales anuales y las evaluaciones de la docencia era de 0,11. Contraste la hipótesis nula de que estas cantidades no están correlacionadas en la población frente a la hipótesis alternativa de que la correlación poblacional es positiva.
- 12.6. Se observa que la correlación muestral de 68 pares de rendimientos anuales de acciones ordinarias del país A y del país B es de 0,51. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa de que es positiva.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

- 12.7. La tabla adjunta y el fichero de datos **Dow Jones** muestran las variaciones porcentuales ( $x_i$ ) del índice Dow-Jones registradas en los cinco primeros días de sesión de cada uno de los años de un periodo de 13 años y las correspondientes variaciones porcentuales ( $y_i$ ) del índice a lo largo de todo el año.

$x$	$y$	$x$	$y$
1,5	14,9	5,6	2,3
0,2	-9,2	-1,4	11,9
-0,1	19,6	1,4	27,0
2,8	20,3	1,5	-4,3
2,2	-3,7	4,7	20,3
-1,6	27,7	1,1	4,2
-1,3	22,6		

- a) Calcule la correlación muestral.  
 b) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa bilateral.

- 12.8. Una universidad distribuye en todos sus cursos un cuestionario de evaluación para que lo rellenen los estudiantes. La tabla adjunta y el fichero de datos **Student Evaluation** muestran tanto la valoración media del profesor (en una escala de 1 a 5) como la calificación media esperada (en una escala de A = 4 a E = 0) de una muestra aleatoria de 12 cursos.

Valoración del profesor	2,8	3,7	4,4	3,6	4,7	3,5	4,1	3,2	4,9	4,2	3,8	3,3
Calificación esperada	2,6	2,9	3,3	3,2	3,1	2,8	2,7	2,4	3,5	3,0	3,4	2,5

- a) Halle la correlación muestral entre las valoraciones de los profesores y las calificaciones esperadas.  
 b) Contraste al nivel de significación del 10 por ciento la hipótesis de que el coeficiente de correlación poblacional es 0 frente a la hipótesis alternativa de que es positivo.
- 12.9. En un estudio sobre la publicidad, los investigadores querían saber si existía una relación entre el coste per cápita y los ingresos per cápita. Se midieron las siguientes variables en una muestra aleatoria de programas de publicidad:

$x_i$  = coste de la publicidad  $\div$  n.º de preguntas recibidas

$y_i$  = ingresos generados por las preguntas  $\div$  n.º de preguntas recibidas

Los datos muestrales se encuentran en el fichero de datos **Advertising Revenue**. Halle la correlación muestral y contraste la hipótesis nula de que la correlación poblacional es 0 frente a la alternativa bilateral.

## 12.2. Modelo de regresión lineal

Para medir la fuerza de cualquier relación lineal entre un par de variables aleatorias se utilizan coeficientes de correlación. Las variables aleatorias se tratan de una forma totalmente simétrica y da lo mismo que hablemos de «la correlación entre  $X$  e  $Y$ » que de «la correlación entre  $Y$  y  $X$ ». En el resto de este capítulo, continuamos analizando la relación lineal entre un par de variables, pero desde el punto de vista de la dependencia de una de la otra. Ahora dejamos de tratar las variables aleatorias de una forma simétrica. La idea es que, dado que la variable aleatoria  $X$  toma un valor específico, esperamos una respuesta de la variable aleatoria  $Y$ . Es decir, el valor que toma  $X$  influye en el valor de  $Y$ . Podemos pensar que  $Y$  depende de  $X$ . Las variables dependientes o endógenas — $Y$ — tienen valores que dependen de variables independientes o exógenas — $X$ —, cuyos valores son manipulados o influidos, a su vez, por factores externos a un proceso económico específico.



Los modelos lineales no son tan restrictivos como podría parecer para el análisis empresarial y económico aplicado. En primer lugar, los modelos lineales a menudo constituyen una buena aproximación de una relación en el intervalo examinado. En segundo lugar, en los Capítulos 13 y 14 veremos que algunas funciones no lineales pueden convertirse en funciones lineales implícitas para el análisis de regresión.

En este capítulo realizamos un estudio formal del análisis de regresión y de la correspondiente inferencia estadística en el caso de modelos lineales sencillos. En los Capítulos 2 y 3 introdujimos los instrumentos de los diagramas de puntos dispersos, la correlación y la regresión simple para describir datos. En el 13 aplicaremos estas ideas a los modelos de regresión múltiple que tienen más de una variable de predicción y en el 14 presentamos métodos y aplicaciones avanzados que aumentan nuestra capacidad para analizar problemas empresariales y económicos.

Este análisis comienza con un ejemplo que muestra una aplicación representativa del análisis de regresión y el tipo de resultados que pueden obtenerse.

### EJEMPLO 12.2. Predicción sobre las ventas de Northern Household Goods (estimación de un modelo de regresión)

El presidente de Northern Household Goods le ha pedido que desarrolle un modelo que prediga las ventas totales de las nuevas tiendas que se propone abrir. Northern es una cadena de grandes almacenes en rápida expansión y necesita una estrategia racional para averiguar dónde deben abrirse nuevas tiendas. Para realizar este proyecto, necesita estimar una ecuación lineal que prediga las ventas al por menor por hogar en función de la renta disponible del hogar. La empresa ha obtenido datos de una encuesta nacional realizada a los hogares y para desarrollar el modelo se utilizarán las variables de las ventas al por menor ( $Y$ ) y la renta ( $X$ ) por hogar.

#### Solución

La Figura 12.1 es un diagrama de puntos dispersos que muestra la relación entre las ventas al por menor y la renta disponible de las familias. Los datos efectivos se muestran en la Tabla 12.1 y se encuentran en el fichero de datos llamado **Retail Sales**. Según la teoría económica, las ventas deben aumentar cuando aumenta la renta disponible y el diagrama de puntos dispersos apoya en gran medida esa teoría. El análisis de regresión nos proporciona un modelo lineal que puede utilizarse para calcular las ventas al por



**Retail  
Sales**



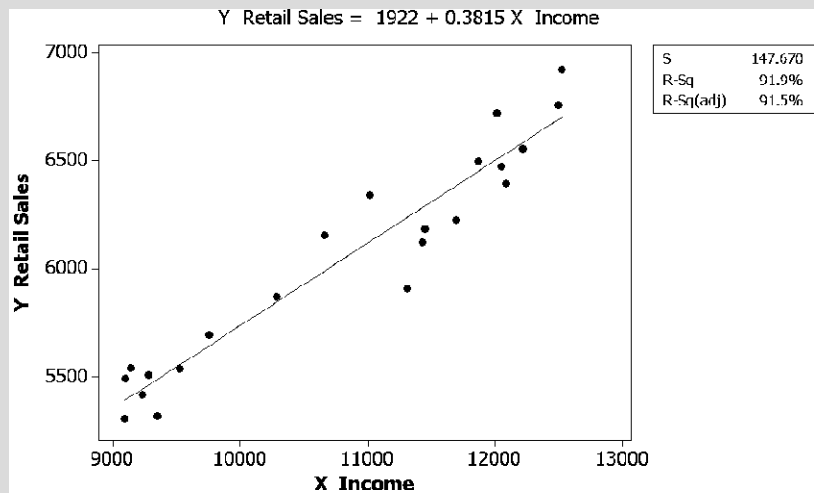


Figura 12.1. Ventas al por menor por hogar en relación con la renta disponible per cápita.

Tabla 12.1. Datos sobre la renta disponible por hogar (X) y ventas al por menor por hogar (Y).

Año	Renta (X)	Ventas al por menor (Y)	Año	Renta (X)	Ventas al por menor (Y)
1	9.098	5.492	12	11.307	5.907
2	9.138	5.540	13	11.432	6.124
3	9.094	5.305	14	11.449	6.186
4	9.282	5.507	15	11.697	6.224
5	9.229	5.418	16	11.871	6.496
6	9.347	5.320	17	12.018	6.718
7	9.525	5.538	18	12.523	6.921
8	9.756	5.692	19	12.053	6.471
9	10.282	5.871	20	12.088	6.394
10	10.662	6.157	21	12.215	6.555
11	11.019	6.342	22	12.494	6.755

menor por hogar correspondientes a varios niveles de renta disponible. La recta del diagrama representa el modelo de regresión simple

$$Y = 1.922,39 + 0,381517X$$

donde Y son las ventas al por menor por hogar y X es la renta disponible por hogar. Por lo tanto, la ecuación de regresión nos proporciona, a partir de los datos, el mejor modelo lineal para predecir las ventas correspondientes a una renta disponible dada. Obsérvese que este modelo nos dice que cada aumento de la renta familiar disponible per cápita de 1 \$, X, va acompañado de un aumento del valor esperado de las ventas al por menor, Y, de 0,38 \$. Es evidente que el resultado es importante para predecir las ventas al por menor. Por ejemplo, observamos que una renta familiar de 50.000 \$ predeciría que las ventas al por menor serán de 20.997 \$ (1.922 + 50.000 × 0,3815).



Llegados a este punto, debemos hacer hincapié en que los resultados de la regresión resumen la información que contienen los datos y no «demuestran» que el aumento de la renta sea la «causa» del aumento de las ventas. La teoría económica sugiere que existe una relación causal y estos resultados apoyan esta teoría. Los diagramas de puntos dispersos, las correlaciones y las ecuaciones de regresión no pueden demostrar la existencia de una relación causal, pero pueden aportar pruebas a su favor. Así pues, para extraer conclusiones, necesitamos conjugar la teoría —la experiencia en la administración de empresas y el análisis económico— con un buen análisis estadístico.

Sabemos por nuestros estudios de la economía que la cantidad comprada de bienes,  $Y$ , en un mercado específico puede representarse por medio de una función lineal de la renta disponible,  $X$ . Si la renta tiene un nivel específico,  $x_i$ , los compradores responden comprando la cantidad  $y_i$ . En el mundo real, sabemos que hay otros factores que influyen en la cantidad efectiva comprada. Son factores identificables como el precio de los bienes en cuestión, la publicidad y los precios de los bienes rivales. También hay otros factores desconocidos que pueden influir en la cantidad efectiva comprada. En una ecuación lineal simple, representamos el efecto de estos factores, salvo la renta, por medio de un término de error llamado  $\varepsilon$ .

La Figura 12.2 muestra un ejemplo de un conjunto de observaciones generadas por un modelo lineal subyacente de un proceso. El nivel medio de  $Y$ , para todo  $X$ , se representa por medio de la ecuación poblacional

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

El modelo de regresión lineal permite hallar el valor esperado de la variable aleatoria  $Y$  cuando  $X$  toma un valor específico. El supuesto de la linealidad implica que esta esperanza puede expresarse de la forma siguiente:

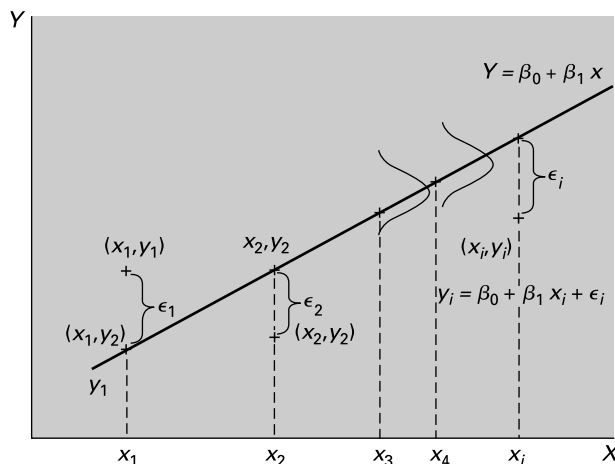
$$E(Y|X = x) = \beta_0 + \beta_1 X$$

donde  $\beta_0$  representa la ordenada en el origen  $Y$  de la ecuación y  $\beta_1$  es la pendiente. El valor observado efectivo de  $Y$  para un valor dado de  $X$  es igual al valor esperado o media poblacional más un error aleatorio,  $\varepsilon$ , que tiene una media 0 y una varianza  $\sigma^2$ :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

El término de error aleatorio  $\varepsilon$  representa la variación de  $Y$  que no es estimada por la relación lineal.

**Figura 12.2.**  
Modelo de  
regresión lineal  
poblacional.



La regresión por mínimos cuadrados nos proporciona un modelo estimado de la relación lineal entre una variable independiente o exógena y una variable dependiente o endógena. Comenzamos el proceso de formulación de la regresión partiendo de un modelo poblacional en el que  $X$  tiene unos valores predeterminados y para todo  $X$  hay un valor medio de  $Y$  más un término de error aleatorio. Utilizamos la ecuación de regresión estimada —mostrada en la Figura 12.1— para estimar el valor medio de  $Y$  para todo valor de  $X$ . Los puntos no están alineados siempre en esta recta debido a que existe un término de error aleatorio que tiene una media 0 y una varianza común para todos los valores de  $X$ . El error aleatorio representa todos los factores que influyen en  $Y$  que no están representados por la relación lineal entre  $Y$  y  $X$ . Los efectos de estos factores, que se supone que son independientes de  $X$ , se comportan como una variable aleatoria cuya media poblacional es 0. Las desviaciones aleatorias  $\varepsilon_i$  en torno al modelo lineal se muestran en la Figura 12.2 y se combinan con la media de  $Y_i$  para todo  $X_i$  para obtener el valor observado  $y_i$ .

### Regresión lineal basada en un modelo poblacional

En la aplicación del análisis de regresión, se representa el proceso estudiado por medio de un modelo poblacional y se calcula un modelo estimado utilizando los datos de que se dispone y realizando una regresión por mínimos cuadrados. El modelo poblacional es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (12.4)$$

donde  $\beta_0$  y  $\beta_1$  son los coeficientes del modelo poblacional y  $\varepsilon$  es un término de error aleatorio. Para todo valor observado,  $x_i$ , el modelo poblacional genera un valor observado,  $y_i$ . Para realizar la inferencia estadística, como veremos en el apartado 12.4, se supone que  $\varepsilon$  sigue una distribución normal de media 0 y varianza  $\sigma^2$ . Más adelante, veremos que puede utilizarse el teorema del límite central para abandonar el supuesto de la distribución normal. El modelo de la relación lineal entre  $Y$  y  $X$  viene definido por los dos coeficientes,  $\beta_0$  y  $\beta_1$ . La Figura 12.2 lo representa esquemáticamente.



En el modelo de regresión por mínimos cuadrados suponemos que se seleccionan valores de la variable independiente,  $x_i$ , y para cada  $x_i$  existe una media poblacional de  $Y$ . Los valores observados de  $y_i$  contienen la media y la desviación aleatoria  $\varepsilon_i$ . Se observa un conjunto de  $n(x_i, y_i)$  puntos y se utiliza para obtener estimaciones de los coeficientes del modelo utilizando el método de mínimos cuadrados. Ampliamos los conceptos de la inferencia clásica presentados en los Capítulos 8 a 11 para hacer inferencias sobre el modelo poblacional subyacente utilizando el modelo de regresión estimado. En el Capítulo 13 veremos cómo pueden considerarse simultáneamente varias variables independientes utilizando la regresión múltiple.

El modelo de regresión estimado y mostrado esquemáticamente en la Figura 12.3 viene dado por la ecuación

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

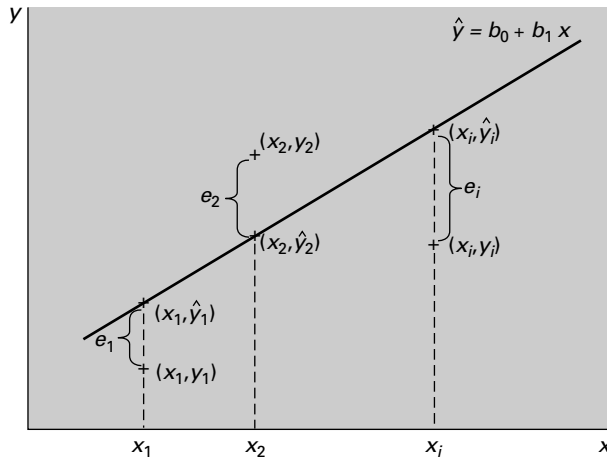
donde  $b_0$  y  $b_1$  son los valores estimados de los coeficientes y  $e$  es la diferencia entre el valor predicho de  $Y$  en la recta de regresión

$$\hat{y}_i = b_0 + b_1 x_i$$

y el valor observado  $y_i$ . La diferencia entre  $y_i$  e  $\hat{y}_i$  para cada valor de  $X$  es el residuo

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

**Figura 12.3.**  
Modelo de  
regresión estimado.



Por lo tanto, para cada valor observado de  $X$  hay un valor predicho de  $Y$  a partir del modelo estimado y un valor observado. La diferencia entre el valor observado de  $Y$  y el predicho es el residuo,  $e_i$ . El residuo,  $e_i$ , no es el error del modelo,  $\varepsilon$ , sino la medida combinada del error del modelo y los errores de la estimación de  $b_0$  y  $b_1$  y, a su vez, los errores de la estimación del valor predicho.

Hallamos el modelo de regresión estimado obteniendo estimaciones,  $b_0$  y  $b_1$ , de los coeficientes poblacionales utilizando el método llamado análisis de mínimos cuadrados, que presentamos en el apartado 12.3. Empleamos, a su vez, estos coeficientes para obtener los valores predichos de  $Y$  para todo valor de  $X$ .

### Resultados de la regresión lineal

La regresión lineal da dos importantes resultados:

1. Los valores predichos de la variable dependiente o endógena en función de la variable independiente o exógena.
2. La variación marginal estimada de la variable endógena provocada por una variación unitaria de la variable independiente o exógena.

## EJERCICIOS

### Ejercicios básicos

**12.10.** Dada la ecuación de regresión

$$Y = 100 + 10X$$

- a) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $+3$ ?
- b) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $-4$ ?
- c) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 12$ ?
- d) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 23$ ?
- e) ¿Demuestra esta ecuación que una variación de  $X$  provoca una variación de  $Y$ ?

**12.11.** Dada la ecuación de regresión

$$Y = -50 + 12X$$

- a) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $+3$ ?
- b) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $-4$ ?
- c) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 12$ ?
- d) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 23$ ?
- e) ¿Demuestra esta ecuación que una variación de  $X$  provoca una variación de  $Y$ ?

**12.12.** Dada la ecuación de regresión

$$Y = 43 + 10X$$

- a) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $+8$ ?
- b) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $-6$ ?
- c) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 11$ ?
- d) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 29$ ?
- e) ¿Demuestra esta ecuación que una variación de  $X$  provoca una variación de  $Y$ ?

12.13. Dada la ecuación de regresión

$$Y = 100 + 21X$$

- a) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $+5$ ?
- b) ¿Cuál es la variación de  $Y$  cuando  $X$  varía en  $-7$ ?
- c) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 14$ ?
- d) ¿Cuál es el valor predicho de  $Y$  cuando  $X = 27$ ?

- e) ¿Demuestra esta ecuación que una variación de  $X$  provoca una variación de  $Y$ ?

### Ejercicios aplicados

- 12.14. ¿Qué diferencia existe entre un modelo lineal poblacional y un modelo de regresión lineal estimado?
- 12.15. Explique la diferencia entre el residuo  $e_i$  y el error del modelo  $\varepsilon_i$ .
- 12.16. Suponga que hemos estimado una ecuación de la regresión de las ventas semanales de «palm pilot» y el precio cobrado durante la semana. Interprete la constante  $b_0$  para el director de la marca.
- 12.17. Se ha estimado un modelo de regresión de las ventas totales de productos alimenticios con respecto a la renta disponible utilizando datos de pequeñas ciudades aisladas del oeste de Estados Unidos. Elabore una lista de los factores que podrían contribuir al término de error aleatorio.

## 12.3. Estimadores de coeficientes por el método de mínimos cuadrados

La recta de regresión poblacional es un útil instrumento teórico, pero para las aplicaciones necesitamos estimar el modelo utilizando los datos de que se disponga. Supongamos que tenemos  $n$  pares de observaciones,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Nos gustaría encontrar la línea recta que mejor se ajusta a estos puntos. Para ello, es necesario encontrar estimadores de los coeficientes desconocidos  $\beta_0$  y  $\beta_1$  de la recta de regresión poblacional.

Hallamos los estimadores de los coeficientes  $b_0$  y  $b_1$  con ecuaciones obtenidas utilizando el método de mínimos cuadrados. Como mostramos en la Figura 12.3, hay una desviación,  $e_i$ , entre el valor observado,  $y_i$ , y el valor predicho,  $\hat{y}_i$ , en la ecuación de regresión estimada para cada valor de  $X$ , donde  $e_i = y_i - \hat{y}_i$ . A continuación, calculamos una función matemática consistente en elevar al cuadrado todos los residuos y sumar las cantidades resultantes. Esta función —cuyo primer miembro se denomina *SCE*— incluye los coeficientes  $b_0$  y  $b_1$ . La cantidad *SCE* se denomina *suma de los cuadrados de los errores*. Los estimadores de los coeficientes  $b_0$  y  $b_1$  son los estimadores que minimizan la suma de los cuadrados de los errores.

### Método de mínimos cuadrados

El método de mínimos cuadrados obtiene estimaciones de los coeficientes de la ecuación lineal  $b_0$  y  $b_1$  en el modelo

$$\hat{y}_i = b_0 + b_1 x_i \quad (12.5)$$

minimizando la suma de los cuadrados de los errores  $e_i$ :

$$SCE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \quad (12.6)$$

Los coeficientes  $b_0$  y  $b_1$  se eligen de tal manera que se minimice la cantidad

$$SCE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 \quad (12.7)$$

Utilizamos el cálculo diferencial para obtener los estimadores de los coeficientes que minimizan la SCE. En el apéndice del capítulo se explica cómo se obtienen los estimadores por medio del cálculo.

El estimador del coeficiente resultante es

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{s_x^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})x_i} y_i \end{aligned}$$

Obsérvese que el numerador del estimador es la covarianza muestral de  $X$  e  $Y$  y el denominador es la varianza muestral de  $X$ . La tercera línea muestra que el coeficiente  $b_1$  es una función lineal de las  $Y$ . Dedicamos mucho tiempo al coeficiente de la pendiente porque este resultado es clave para muchas aplicaciones. El coeficiente de la pendiente  $b_1$  es una estimación de la variación que experimenta  $Y$  cuando  $X$  varía en una unidad. Por ejemplo, si  $Y$  es la producción total y  $X$  es el número de trabajadores, entonces  $b_1$  es una estimación del aumento marginal de la producción por cada nuevo trabajador. Este tipo de resultados explica por qué la regresión se ha convertido en un instrumento analítico tan importante.

Con algunas manipulaciones algebraicas podemos demostrar que el estimador del coeficiente también es igual a

$$b_1 = r \frac{s_y}{s_x}$$

donde  $r_{xy}$  es la correlación muestral y  $s_y$  y  $s_x$  son las desviaciones típicas muestrales de  $X$  e  $Y$ . Este resultado es importante porque indica cómo está relacionada directamente la relación estandarizada entre  $X$  e  $Y$ , la correlación  $r_{xy}$ , con el coeficiente de la pendiente.

En el apéndice del capítulo también mostramos que el estimador de la constante es

$$b_0 = \bar{y} - b_1 \bar{x}$$

Sustituyendo  $b_0$  por este valor en la ecuación lineal, tenemos que

$$\begin{aligned} y &= \bar{y} - b_1 \bar{x} + b_1 x \\ y - \bar{y} &= b_1 (x - \bar{x}) \end{aligned}$$

En esta ecuación vemos que cuando  $x = \bar{x}$ , entonces  $y = \bar{y}$  y que la ecuación de regresión siempre pasa por el punto  $(\bar{x}, \bar{y})$ . El valor estimado de la variable dependiente,  $\hat{y}_i$ , se obtiene utilizando

$$\hat{y}_i = b_0 + b_1 x_i$$

o utilizando

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

Esta última forma pone de relieve que la recta de regresión pasa por las medias de  $X$  e  $Y$ .

### Estimadores de coeficientes por el método de mínimos cuadrados

El estimador del coeficiente de la pendiente es

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

y el estimador de la constante u ordenada en el origen es

$$b_0 = \bar{y} - b_1\bar{x}$$

También señalamos que la recta de regresión siempre pasa por la media  $\bar{x}$ ,  $\bar{y}$ .

El método de mínimos cuadrados podría utilizarse para calcular estimaciones de los coeficientes  $b_0$  y  $b_1$  utilizando cualquier conjunto de datos pareados. Sin embargo, en la mayoría de las aplicaciones queremos hacer inferencias sobre el modelo poblacional subyacente que forma parte de nuestro problema económico o empresarial. Para hacer inferencias, es necesario que estemos de acuerdo en ciertos supuestos. Dados estos supuestos, puede demostrarse que los estimadores de los coeficientes por mínimos cuadrados son insesgados y tienen una varianza mínima.

### Supuestos habituales en los que se basa el modelo de regresión lineal

Para hacer inferencias sobre el modelo lineal poblacional utilizando los coeficientes del modelo estimados se postulan los siguientes supuestos.

1. Las  $Y$  son funciones lineales de  $X$  más un término de error aleatorio

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

2. Las  $x$  son números fijos o son realizaciones de la variable aleatoria  $X$  que son independientes de los términos de error,  $\varepsilon_i$ . En el segundo caso, la inferencia se realiza condicionada a los valores observados de las  $x$ .
3. Los términos de error son variables aleatorias que tienen la media 0 y la misma varianza  $\sigma^2$ . El segundo supuesto se llama homocedasticidad o varianza uniforme.

$$E[\varepsilon_i] = 0 \quad \text{y} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{para } (i = 1, \dots, n)$$

4. Los términos de error aleatorio,  $\varepsilon_i$ , no están correlacionados entre sí, por lo que

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{para todo } i \neq j$$

Generalmente, se considera, con razón, que el segundo de estos supuestos es cierto, aunque en algunos estudios econométricos avanzados es insostenible (el supuesto no se cumple, por ejemplo, cuando no es posible medir  $x_i$  con precisión o cuando la regresión forma parte de un sistema de ecuaciones interdependientes). Sin embargo, aquí consideraremos que se satisface este supuesto.

Los supuestos 3 y 4 se refieren a los términos de error,  $\varepsilon_i$ , de la ecuación de regresión. El término de error esperado es 0 y todos los términos de error tienen la misma varianza. Por lo tanto, no esperamos que las varianzas de los términos de error sean más altas en el caso de algunas observaciones que en el de otras. La Figura 12.2 muestra esta pauta: los errores correspondientes a todos los valores de  $X$  proceden de poblaciones que tienen la misma varianza. Por último, se supone que las discrepancias no están correlacionadas entre sí. Así, por ejemplo, la aparición de una gran discrepancia positiva en un punto de observación no nos ayuda a predecir los valores de ninguno de los demás términos de error. Los supuestos 3 y 4 se satisfacen si los términos de error,  $\varepsilon_i$ , pueden concebirse como una muestra aleatoria procedente de una población que tiene de media 0. En el resto de este capítulo, estos supuestos se cumplen. La posibilidad de abandonar algunos de ellos se examina en el Capítulo 14.

### Cálculo por computador del coeficiente de regresión

La extensa aplicación del análisis de regresión ha sido posible gracias a los paquetes estadísticos y a Excel. Como sospechará el lector, los cálculos para obtener estimaciones de los coeficientes de regresión son tediosos. Las ecuaciones de los estimadores y otros importantes cálculos estadísticos están incluidos en los paquetes informáticos y en Excel y se utilizan para estimar los coeficientes de problemas específicos. El programa Excel puede utilizarse para realizar análisis básicos de regresión sin demasiadas dificultades. Pero si se desea utilizar métodos de análisis de regresión aplicado avanzado o un perspicaz análisis gráfico, debe utilizarse un buen paquete estadístico. Dado que nos interesan principalmente las aplicaciones, nuestra tarea más importante es realizar un análisis adecuado de los cálculos de regresión para estas aplicaciones. Este análisis debe realizarse conociendo las ecuaciones de los estimadores y el análisis relacionado con ellas. Sin embargo, no utilizamos estas ecuaciones para calcular realmente las estimaciones u otros estadísticos de la regresión. *Dejamos los cálculos para los computadores; nuestra tarea es pensar, analizar y hacer recomendaciones.*

La Figura 12.4 muestra una parte de las salidas Minitab y Excel correspondientes al ejemplo de las ventas al por menor. Obsérvese la localización de las estimaciones de la constante,  $b_0$ , y el coeficiente de la pendiente,  $b_1$ , en la salida informática. Los conceptos restantes de cada línea ayudan a interpretar la calidad de las estimaciones y se explican en apartados posteriores.

En esta regresión, la constante estimada,  $b_0$ , es 1.922 y el coeficiente de la pendiente estimado,  $b_1$ , es 0,382. Estos valores se calculan utilizando las ecuaciones de los estimadores de los coeficientes antes presentadas. La ecuación estimada puede expresarse de la forma siguiente:

$$\hat{y} = 1.922 + 0,382x$$

o, utilizando las medias  $\bar{x} = 10.799$  e  $\bar{y} = 6.042$ , de la forma siguiente:

$$\hat{y} = 6.042 + 0,382(x - 10.799)$$



Normalmente, los modelos de regresión sólo deben utilizarse en el rango de los valores observados de  $X$  en el que tenemos información sobre la relación porque la relación puede no ser lineal fuera de este rango. La segunda forma del modelo de regresión está centrada en las medias de los datos con una tasa de variación igual a  $b_1$ . Utilizando esta forma, centramos la atención en la localización media del modelo de regresión y no en la ordenada



**Results for: retail sales.MTW****Regression Analysis: Y Retail Sales versus X Income**

The regression equation is

$$Y \text{ Retail Sales} = 1922 + 0.382 X \text{ Income}$$

Coeficientes  $b_0, b_1$ 

Predictor	Coef	SE Coef	T	P
Constant	1922.4	274.9	6.99	0.000
X Income	0.38152	0.02529	15.08	0.000

 $S = 147.670$        $R\text{-Sq} = 91.9\%$        $R\text{-Sq}(\text{adj}) = 91.5\%$ 

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6697181					
8	Observations	22					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	4961434.406	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		Coeficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08305	2.17E-12	0.320756343	0.4342771
19							

Coeficientes  $b_0, b_1$ 

(b)

**Figura 12.4.** Análisis de regresión de las ventas al por menor (a) por medio de Minitab y (b) por medio de Excel.

en el origen con el eje de las  $Y$ . Los usuarios ingenuos del análisis de regresión a veces intentan hacer interpretaciones de la constante  $b_0$ , extrayendo ciertas conclusiones sobre la variable dependiente cuando la variable independiente tiene un valor de 0. Consideremos la regresión de las ventas al por menor con respecto a la renta disponible del ejemplo. ¿Afirmaríamos realmente que las ventas al por menor son de 1.922 \$ cuando la renta disponible es de 0? En realidad, sencillamente no tenemos datos para afirmar que se vende algo cuando la renta disponible es 0. Éste es otro ejemplo de la importancia de un buen análisis en lugar de interpretaciones tontas. Como analistas profesionales, debemos tener cuidado de no defender resultados que sencillamente no existen.

**EJERCICIOS****Ejercicios básicos**

**12.18.** Calcule los coeficientes de una ecuación de regresión por mínimos cuadrados y formule la ecuación, dados los siguientes estadísticos muestrales:

- a)  $\bar{x} = 50$ ;  $\bar{y} = 100$ ;  $s_x = 25$ ;  $s_y = 75$ ;  $r_{xy} = 0,6$ ;  $n = 60$   
 b)  $\bar{x} = 60$ ;  $\bar{y} = 210$ ;  $s_x = 35$ ;  $s_y = 65$ ;  $r_{xy} = 0,7$ ;  $n = 60$

- c)  $\bar{x} = 20$ ;  $\bar{y} = 100$ ;  $s_x = 60$ ;  $s_y = 78$ ;  $r_{xy} = 0,75$ ;  $n = 60$   
 d)  $\bar{x} = 10$ ;  $\bar{y} = 50$ ;  $s_x = 100$ ;  $s_y = 75$ ;  $r_{xy} = 0,4$ ;  $n = 60$   
 e)  $\bar{x} = 90$ ;  $\bar{y} = 200$ ;  $s_x = 80$ ;  $s_y = 70$ ;  $r_{xy} = 0,6$ ;  $n = 60$

**Ejercicios aplicados**

**12.19.** Una empresa fija un precio distinto para un sistema de DVD en ocho regiones del país. La ta-

bla adjunta muestra los números de unidades vendidas y los precios correspondientes (en cientos de dólares).

<b>Ventas</b>	420	380	350	400	440	380	450	420
<b>Precio</b>	5,5	6,0	6,5	6,0	5,0	6,5	4,5	5,0

- Represente estos datos y estime la regresión lineal de las ventas con respecto al precio.
- ¿Qué efecto sería de esperar que produjera una subida del precio de 100 \$ en las ventas?

**12.20.** Dada una muestra de 20 observaciones mensuales, un analista financiero quiere realizar una regresión de la tasa porcentual de rendimiento ( $Y$ ) de las acciones ordinarias de una empresa con respecto a la tasa porcentual de rendimiento ( $X$ ) del índice Standard and Poor's 500. Dispone de la siguiente información:

$$\sum_{i=1}^{20} y_i = 22,6 \quad \sum_{i=1}^{20} x_i = 25,4$$

$$\sum_{i=1}^{20} x_i^2 = 145,7 \quad \sum_{i=1}^{20} x_i y_i = 150,5$$

- Estime la regresión lineal de  $Y$  con respecto a  $X$ .
- Interprete la pendiente de la recta de regresión muestral.
- Interprete la ordenada en el origen de la recta de regresión muestral.

**12.21.** Una empresa realiza un test de aptitud a todos los nuevos representantes de ventas. La dirección tiene interés en saber en qué medida es capaz este test de predecir su éxito final. La tabla adjunta muestra las ventas semanales medias (en miles de dólares) y las puntuaciones obtenidas en el test de aptitud por una muestra aleatoria de ocho representantes.

<b>Ventas semanales</b>	10	12	28	24	18	16	15	12
<b>Puntuación</b>	55	60	85	75	80	85	65	60

- Estime la regresión lineal de las ventas semanales con respecto a las puntuaciones del test de aptitud.
- Interprete la pendiente estimada de la recta de regresión.

**12.22.** Se ha formulado la hipótesis de que el número de botellas de una cerveza importada que se

vende cada noche en los restaurantes de una ciudad depende linealmente de los costes medios de las cenas en los restaurantes. Se han obtenido los siguientes resultados de una muestra de  $n = 17$  restaurantes que son aproximadamente del mismo tamaño, siendo

$y$  = número de botellas vendidas por noche  
 $x$  = coste medio, en dólares, de una cena

$$\bar{x} = 25,5 \quad \bar{y} = 16,0$$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = 350 \quad \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = 180$$

- Halle la recta de regresión muestral.
- Interprete la pendiente de la recta de regresión muestral.
- ¿Es posible dar una interpretación que tenga sentido de la ordenada en el origen de la recta de regresión muestral? Explique su respuesta.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

**12.23.** Vuelva a los datos del ejercicio 12.7 sobre la variación porcentual ( $X$ ) del índice Dow-Jones en los cinco primeros días de sesión del año y la variación porcentual ( $Y$ ) del índice en el conjunto del año.

- Estime la regresión lineal de  $Y$  con respecto a  $X$ .
- Interprete la ordenada en el origen y la pendiente de la recta de regresión muestral.

**12.24.** El viernes 13 de noviembre de 1989, cayeron vertiginosamente las cotizaciones en la bolsa de Nueva York; el índice Standard and Poor's 500 cayó un 6,1 por ciento ese día. El fichero de datos **New York Stock Exchange Gains and Losses** muestra las *pérdidas* porcentuales ( $y$ ) que experimentaron los 25 mayores fondos de inversión el 13 de noviembre de 1989. También muestra las *ganancias* porcentuales ( $x$ ), suponiendo que los dividendos y las ganancias de capital de estos mismos fondos se reinvertieron en 1989 hasta el 12 de noviembre.

- Estime la regresión lineal de las pérdidas registradas el 13 de noviembre con respecto a las ganancias obtenidas hasta el 12 de noviembre de 1989.
- Interprete la pendiente de la recta de regresión muestral.

**12.25.** Ace Manufacturing está estudiando el absentismo laboral. Los datos del fichero **Employee Absence** se refieren a la variación anual de la tasa total de absentismo y la variación anual de la tasa media de absentismo por enfermedad.

- Estime la regresión lineal de la variación de la tasa media de absentismo por enfermedad con respecto a la variación de la tasa de absentismo.
- Interprete la pendiente estimada de la recta de regresión.

## 12.4. El poder explicativo de una ecuación de regresión lineal

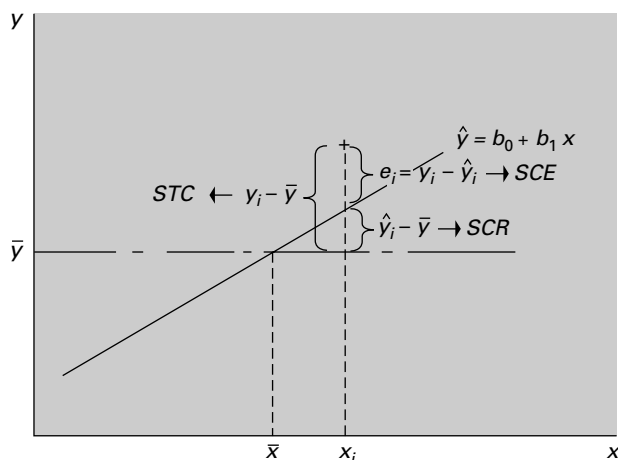
El modelo de regresión estimado que hemos presentado puede concebirse como un intento de explicar los cambios de una variable dependiente  $Y$  provocados por los cambios de una variable independiente  $X$ . Si sólo tuviéramos observaciones de la variable dependiente,  $Y$ , la tendencia central de  $Y$  se representaría por medio de la media  $\bar{y}$  y la variabilidad total en torno a  $Y$  se representaría por medio del numerador del estimador de la varianza muestral,  $\Sigma(y_i - \bar{y})^2$ . Cuando también tenemos medidas de  $X$ , hemos demostrado que la tendencia central de  $Y$  ahora puede expresarse en función de  $X$ . Esperamos que la ecuación lineal esté más cerca de los valores individuales de  $Y$  y que, por lo tanto, la variabilidad en torno a la ecuación lineal sea menor que la variabilidad en torno a la media.

Estamos ya en condiciones de desarrollar medidas que indiquen la eficacia con que la variable  $X$  explica la conducta de  $Y$ . En nuestro ejemplo de las ventas al por menor mostrado en la Figura 12.1, las ventas al por menor,  $Y$ , tienden a aumentar con la renta disponible,  $X$  y, por lo tanto, la renta disponible explica algunas de las diferencias entre las ventas al por menor. Sin embargo, los puntos no están todos en la línea, por lo que la explicación no es perfecta. Aquí desarrollamos medidas basadas en la descomposición de la variabilidad, que miden la capacidad de  $X$  para explicar  $Y$  en una regresión específica.

El análisis de la varianza, ANOVA, para una regresión de mínimos cuadrados se realiza descomponiendo la variabilidad total de  $Y$  en un componente explicado y un componente de error. En la Figura 12.5 mostramos que la desviación de un valor de  $Y$  con respecto a su media puede descomponerse en la desviación del valor predicho con respecto a la media y la desviación del valor observado con respecto al valor predicho

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

**Figura 12.5.**  
Descomposición de la variabilidad.



Elevamos al cuadrado los dos miembros de la ecuación —ya que la suma de las desviaciones en torno a la media es igual a 0— y sumamos el resultado obtenido en los  $n$  puntos

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

Tal vez algunos lectores se hayan dado cuenta de que la elevación al cuadrado del primer miembro debe incluir el producto de los dos términos además de sus cantidades al cuadrado. Puede demostrarse que el término del producto de los dos términos es igual a 0. Esta ecuación puede expresarse de la forma siguiente:

$$STC = SCR + SCE$$

Aquí vemos que la variabilidad total — $STC$ — puede dividirse en un componente — $SCR$ — que representa la variabilidad que es explicada por la pendiente de la ecuación de regresión (la media de  $Y$  es diferente en distintos niveles de  $X$ ). El segundo componente — $SCE$ — se debe a la desviación aleatoria o sin explicar de los puntos con respecto a la recta de regresión. Esta variabilidad es una indicación de la incertidumbre relacionada con el modelo de regresión. El primer miembro es la *suma total de los cuadrados*:

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

La cantidad de variabilidad explicada por la ecuación de regresión es la *suma de los cuadrados de la regresión* y se calcula de la forma siguiente:

$$SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Vemos que la variabilidad explicada por la regresión depende directamente de la magnitud del coeficiente  $b_1$  y de la dispersión de los datos de la variable independiente,  $X$ . Las desviaciones en torno a la recta de regresión,  $e_i$ , que se utilizan para calcular la parte no explicada, o sea, la *suma de los cuadrados de los errores*, pueden definirse utilizando las siguientes formas algebraicas:

$$SCE = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

Dado un conjunto de valores observados de las variables dependientes,  $Y$ , la  $STC$  es fija e igual a la variabilidad total de todas las observaciones con respecto a la media. Vemos que en esta descomposición, cuanto más altos son los valores de  $SCR$  y, por lo tanto, cuanto más bajos son los valores de  $SCE$ , mejor «se ajusta» o se aproxima la ecuación de regresión a los datos observados. Esta descomposición se muestra gráficamente en la Figura 12.5. En la ecuación de  $SCR$  vemos que la variabilidad explicada,  $SCR$ , está relacionada directamente con la dispersión de la variable independiente o  $X$ . Por lo tanto, cuando examinamos aplicaciones del análisis de regresión, sabemos que debemos tratar de obtener datos que tengan un gran rango para la variable independiente de manera que el modelo de regresión resultante tenga una variabilidad sin explicar menor.

### Análisis de la varianza

La variabilidad total en un análisis de regresión,  $STC$ , puede descomponerse en un componente explicado por la regresión,  $SCR$ , y un componente que se debe a un error sin explicar,  $SCE$ :

$$STC = SCR + SCE \quad (12.8)$$

cuyos componentes se definen de la forma siguiente.

Suma total de los cuadrados:

$$STC = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (12.9)$$

Suma de los cuadrados de los errores:

$$SCE = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \quad (12.10)$$

Suma de los cuadrados de la regresión:

$$SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \quad (12.11)$$



#### Retail Sales

Volvamos con esta información a nuestro ejemplo de las ventas al por menor (ejemplo 12.2) con el fichero de datos **Retail Sales** y veamos cómo utilizamos la descomposición de la variabilidad para averiguar en qué medida explica nuestro modelo el proceso estudiado. La Tabla 12.2 muestra los cálculos detallados de los residuos,  $e_i$ ; las desviaciones de  $Y$  con respecto a la media, y las desviaciones de los valores predichos de  $Y$  con respecto a la media. Éstos nos proporcionan los componentes para calcular  $SCE$ ,  $STC$  y  $SCR$ . La suma de los cuadrados de las desviaciones de la columna 5 es  $SCE = 436.127$ . La suma de los cuadrados de las desviaciones de la columna 6 es  $STC = 5.397.561$ . Por último, la suma de los cuadrados de las desviaciones de la columna 7 es  $SCR = 4.961.434$ . La Figura 12.6 presenta las salidas Minitab y Excel del análisis de regresión, incluido el análisis de la varianza.

### El coeficiente de determinación $R^2$

Hemos visto que el ajuste de la ecuación de regresión a los datos mejora cuando aumenta  $SCR$  y disminuye  $SCE$ . El cociente entre la suma de los cuadrados de la regresión,  $SCR$ , y la suma total de los cuadrados,  $STC$ , es una medida descriptiva de la proporción o porcentaje de la variabilidad total que es explicada por el modelo de regresión. Esta medida se llama *coeficiente de determinación* o, en términos más generales,  $R^2$ .

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

A menudo se considera que el coeficiente de determinación es el porcentaje de la variabilidad de  $Y$  que es explicado por la ecuación de regresión. Antes hemos demostrado que  $SCR$  aumenta directamente con la dispersión de la variable independiente  $X$ :

$$SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

**Tabla 12.2.** Valores efectivos y predichos de las ventas al por menor por hogar y residuos calculados a partir de su regresión lineal con respecto a la renta por hogar.

Año	Renta (X)	Ventas al por menor (Y)	Ventas al por menor predichas	Residuo	Desviación observada con respecto a la media	Desviación predicha con respecto a la media
1	9.098	5.492	5.394	98	-550	-649
2	9.138	5.540	5.409	131	-502	-633
3	9.094	5.305	5.392	-87	-737	-650
4	9.282	5.507	5.464	43	-535	-578
5	9.229	5.418	5.444	-26	-624	-599
6	9.347	5.320	5.489	-169	-722	-554
7	9.525	5.538	5.557	-19	-504	-486
8	9.756	5.692	5.645	47	-350	-397
9	10.282	5.871	5.846	25	-171	-197
10	10.662	6.157	5.991	166	115	-52
11	11.019	6.342	6.127	215	300	84
12	11.307	5.907	6.237	-330	-135	194
13	11.432	6.124	6.284	-160	82	242
14	11.449	6.186	6.291	-105	144	248
15	11.697	6.224	6.385	-161	182	343
16	11.871	6.496	6.452	44	454	409
17	12.018	6.718	6.508	210	676	465
18	12.523	6.921	6.701	220	879	658
19	12.053	6.471	6.521	-50	429	479
20	12.088	6.394	6.535	-141	352	492
21	12.215	6.555	6.583	-28	513	541
22	12.494	6.755	6.689	66	713	647
Suma de los cuadrados de los valores				436.127	5.397.561	4.961.434

Vemos, pues, que  $R^2$  también aumenta directamente con la dispersión de la variable independiente. Cuando buscamos datos para estimar un modelo de regresión, es importante elegir las observaciones de la variable independiente que abarquen la mayor dispersión posible de  $X$  con el fin de obtener un modelo de regresión con el mayor  $R^2$ .

### Coefficiente de determinación $R^2$

El coeficiente de determinación de una ecuación de regresión es

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC} \quad (12.12)$$

Esta cantidad varía de 0 a 1 y los valores más altos indican que la regresión es mejor. Las interpretaciones generales de  $R^2$  deben hacerse con cautela, ya que un valor alto puede deberse a que  $SCE$  es bajo o a que  $STC$  es alto o ambas cosas a la vez.

$R^2$  puede variar de 0 a 1, ya que  $STC$  es fijo y  $0 < SCE < STC$ . Cuando  $R^2$  es alto, significa que la regresión es mejor, manteniéndose todo lo demás constante. En la salida del análisis de regresión vemos que el  $R^2$  de la regresión de las ventas al por menor es 0,919, o sea, 91,9 por ciento. Normalmente, se considera que  $R^2$  es la *variabilidad porcentual explicada*.

Results for: retail sales.MTW

Regression Analysis: Y Retail Sales versus X Income

The regression equation is

$$Y \text{ Retail Sales} = 1922 + 0.382 X \text{ Income}$$

Predictor	Coef	SE Coef	T	P
Constant	1922.4	274.9	6.99	0.000
X Income	0.38152	0.02529	15.08	0.000

$s_e$ , Error típico de la estimación

S = 147.670 R-Sq = 91.9% R-Sq(adj) = 91.5%

$R^2$ , Coeficiente de determinación

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4961434	4961434	227.52	0.000
Residual Error	20	436127	21806		
Total	21	5397561			

$s_e^2$ , Varianza del error del modelo

Unusual Observations

Obs	X	Income	Y Retail Sales	Fit	SE Fit	Residual	St Resid
12		11307	5907.0	6236.2	34.0	-329.2	-2.29R

SRC = 4,961,434  
SCE = 436,127  
STC = 5,397,561

R denotes an observation with a large standardized residual.

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6697101					
8	Observations	22					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	4961434.406	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08305	2.17E-12	0.320756343	0.4342771

$s_e$ , Error típico de la estimación

$R^2$ , Coeficiente de determinación

$s_e^2$ , Varianza

(b)

Figura 12.6. Análisis de regresión de las ventas al por menor con respecto a la renta disponible: (a) salida Minitab; (b) salida Excel.

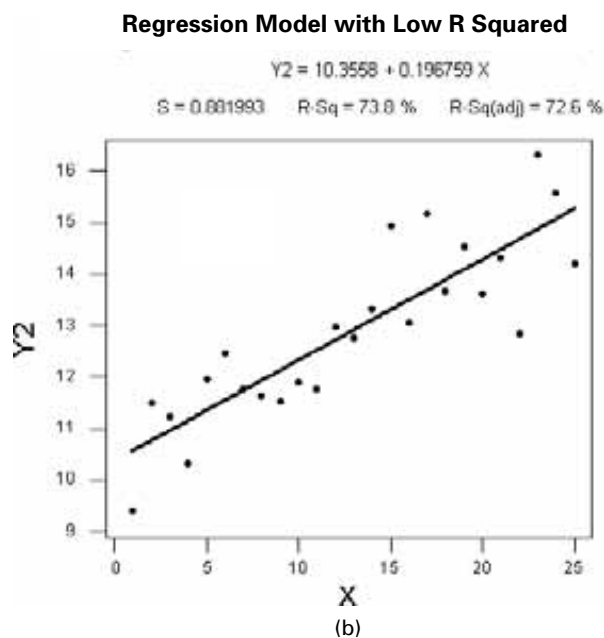
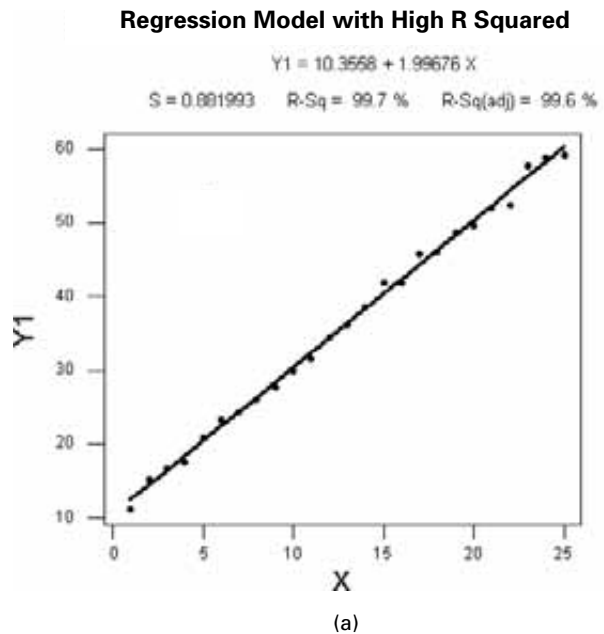


La segunda forma de la ecuación pone de manifiesto que  $R^2$  depende del cociente entre SCE y STC.  $R^2$  puede ser alto porque SCE es bajo —el objetivo deseado— o porque STC es alto o por ambas cosas a la vez. Las interpretaciones generales de  $R^2$  que se aplican a todas las ecuaciones de regresión son peligrosas. Dos modelos de regresión que tengan el mismo conjunto de  $y_i$  observadas siempre pueden compararse utilizando el coeficiente de determinación  $R^2$ , y el modelo cuyo  $R^2$  sea más alto explica mejor la variable Y. Pero las comparaciones generales de  $R^2$  —que afirman que un modelo es bueno porque su  $R^2$  es

superior a un determinado valor— son engañosas. Generalmente, los analistas con experiencia han observado que  $R^2$  es 0,80 o más en los modelos basados en datos de series temporales. En los modelos basados en datos de corte transversal (por ejemplo, ciudades, regiones, empresas), el valor de  $R^2$  oscila entre 0,40 y 0,60 y en los modelos basados en datos de personas individuales a menudo oscila entre 0,10 y 0,20.

Para ilustrar el problema de las interpretaciones generales de  $R^2$ , consideremos dos modelos de regresión —cuyos gráficos se muestran en la Figura 12.7—, cada uno de los cuales se basa en un total de 25 observaciones. En ambos modelos,  $SCE$  es igual a 17,89, por lo

**Figura 12.7.**  
Comparación del  $R^2$   
de dos modelos de  
regresión;  
(a)  $R^2$  alto;  
(b)  $R^2$  bajo.





que el ajuste de la ecuación de regresión a los puntos de datos es el mismo. Pero en el primer modelo, la suma total de los cuadrados es igual a 5.201,05, mientras que en el segundo es igual a 68,22. Los valores de  $R^2$  de los dos modelos son los siguientes.

Modelo 1:

$$R^2 = 1 - \frac{SCE}{STC} = 1 - \frac{17,89}{5.201,05} = 0,997$$

Modelo 2:

$$R^2 = 1 - \frac{SCE}{STC} = 1 - \frac{17,89}{68,22} = 0,738$$

Dado que  $SCE$  es igual en ambos modelos y, por lo tanto, la bondad del ajuste es la misma en los dos, no podemos afirmar que el modelo 1 se ajusta mejor a los datos. Sin embargo, en el modelo 1 el valor de  $R^2$  es mucho más alto que en el modelo 2. Como vemos aquí, la interpretación general de  $R^2$  debe hacerse con mucha cautela. Obsérvese que los dos intervalos diferentes del eje de ordenadas de la Figura 12.7 se deben a valores diferentes de  $STC$ .

También puede establecerse una relación entre el coeficiente de correlación y el  $R^2$ , observando que la correlación al cuadrado es igual al coeficiente de determinación. Otra interpretación de la correlación es que es la raíz cuadrada de la variabilidad porcentual explicada.

### Correlación y $R^2$

El coeficiente de determinación,  $R^2$ , de la regresión simple es igual al cuadrado del coeficiente de correlación simple:

$$R^2 = r^2 \quad (12.13)$$

Este resultado establece una importante conexión entre la correlación y el modelo de regresión.

La suma de los cuadrados de los errores puede utilizarse para obtener una estimación de la varianza del error del modelo  $\varepsilon_i$ . Como veremos, el estimador de la varianza del error del modelo se utiliza para realizar la inferencia estadística en el modelo de regresión. Recuerdese que hemos supuesto que el error poblacional,  $\varepsilon_i$ , es un error aleatorio que tiene una media 0 y una varianza  $\sigma^2$ . El estimador de  $\sigma^2$  se calcula de la forma siguiente:

### Estimación de la varianza del error del modelo

La cantidad  $SCE$  es una medida de la suma total de los cuadrados de las desviaciones en torno a la recta de regresión estimada y  $e_i$  es el residuo. Un estimador de la varianza del error poblacional del modelo es

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SCE}{n-2} \quad (12.14)$$

Se divide por  $n-2$  en lugar de  $n-1$  porque el modelo de regresión simple utiliza dos parámetros estimados,  $b_0$  y  $b_1$ , en lugar de uno. En el siguiente apartado vemos que este estimador de la varianza es la base de la inferencia estadística en el modelo de regresión.

## EJERCICIOS

## Ejercicios básicos

**12.26.** Calcule  $SCR$ ,  $SCE$ ,  $s_e^2$  y el coeficiente de determinación, dados los siguientes estadísticos calculados a partir de una muestra aleatoria de pares de observaciones de  $X$  e  $Y$ :

- a)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 100.000$ ;  $r^2 = 0,50$ ;  $n = 52$
- b)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 90.000$ ;  $r^2 = 0,70$ ;  $n = 52$
- c)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 240$ ;  $r^2 = 0,80$ ;  $n = 52$
- d)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 200.000$ ;  $r^2 = 0,30$ ;  $n = 74$
- e)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 60.000$ ;  $r^2 = 0,90$ ;  $n = 40$

## Ejercicios aplicados

**12.27.** Sea la recta de regresión muestral

$$y_i = b_0 + b_1 x_i + e_i = \hat{y}_i + e_i \quad (i = 1, 2, \dots, n)$$

y sean  $\bar{x}$  y  $\bar{y}$  las medias muestrales de las variables independiente y dependiente, respectivamente.

a) Demuestre que

$$e_i = y_i - \bar{y} - b(x_i - \bar{x})$$

b) Utilizando el resultado del apartado (a), demuestre que

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

c) Utilizando el resultado del apartado (a), demuestre que

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

d) Demuestre que

$$\hat{y}_i - \bar{y} = b_1(x_i - \bar{x})$$

e) Utilizando los resultados de los apartados (c) y (d), demuestre que

$$STC = SCR + SCE$$

f) Utilizando el resultado del apartado (a), demuestre que

$$\sum_{i=1}^n e_i(x_i - \bar{x}) = 0$$

**12.28.** Sea

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

el coeficiente de determinación de la recta de regresión muestral.

a) Utilizando el apartado (d) del ejercicio 12.27, demuestre que

$$R^2 = b_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$


b) Utilizando el resultado del apartado (a), demuestre que el coeficiente de determinación es igual al cuadrado de la correlación muestral entre  $X$  e  $Y$ .


c) Sea  $b_1$  la pendiente de la regresión por mínimos cuadrados de  $Y$  con respecto a  $X$ ,  $b_1^*$  la pendiente de la regresión por mínimos cuadrados de  $X$  con respecto a  $Y$  y  $r$  la correlación muestral entre  $X$  e  $Y$ . Demuestre que


$$b_1 \cdot b_1^* = r^2$$

**12.29.** Halle e interprete el coeficiente de determinación de la regresión de las ventas del sistema de DVD con respecto al precio, utilizando los datos siguientes.

Ventas	420	380	350	400	440	380	450	420
Precio	5,5	6,0	6,5	6,0	5,0	6,5	4,5	5,0

**12.30.**  Halle e interprete el coeficiente de determinación de la regresión de la variación porcentual del índice Dow-Jones en un año con respecto a la variación porcentual del índice en los cinco primeros días de sesión del año, continuando con el análisis del ejercicio 12.7. Compare su respuesta con la correlación muestral obtenida con estos datos en el ejercicio 12.7. Utilice el fichero de datos **Dow Jones**.

**12.31.**  Basándose en los datos del ejercicio 12.24, halle la proporción de la variabilidad muestral de las pérdidas porcentuales experimentadas por los fondos de inversión el 13 de noviembre de 1989 explicada por su dependencia lineal de las ganancias porcentuales obtenidas en 1989 hasta el 12 de noviembre. Utilice el fichero de datos **New York Stock Exchange Gains and Losses**.

**12.32.**  Vuelva a los datos sobre la tasa de absentismo laboral del ejercicio 12.25. Utilice el fichero de datos **Employee Absence**.

- a) Halle los valores predichos,  $\hat{y}_i$ , y los residuos,  $e_i$ , de la regresión por mínimos cuadrados de la variación de la tasa media de absentismo por enfermedad con respecto a la variación de la tasa de desempleo.
- b) Halle las sumas de los cuadrados  $STC$ ,  $SCR$  y  $SCE$  y verifique que

$$STC = SCR + SCE$$

- c) Utilizando los resultados del apartado (a), halle e interprete el coeficiente de determinación.

**12.33.** Vuelva a los datos sobre las ventas semanales y las puntuaciones obtenidas en un test de aptitud por los representantes de ventas del ejercicio 12.21.

- a) Halle los valores predichos,  $\hat{y}_i$ , y los residuos,  $e_i$ , de la regresión por mínimos cua-

drados de las ventas semanales con respecto a las puntuaciones del test de aptitud.

- b) Halle las sumas de los cuadrados  $STC$ ,  $SCR$  y  $SCE$  y verifique que

$$STC = SCR + SCE$$

- c) Utilizando los resultados del apartado (a), halle e interprete el coeficiente de determinación.
- d) Halle directamente el coeficiente de correlación muestral entre las ventas y las puntuaciones del test de aptitud y verifique que su cuadrado es igual al coeficiente de determinación.

**12.34.** En un estudio se demostró que en una muestra de 353 profesores universitarios, la correlación entre las subidas salariales anuales y las evaluaciones de la docencia era de 0,11. ¿Cuál sería el coeficiente de determinación de una regresión de las subidas salariales anuales con respecto a las evaluaciones de la docencia en esta muestra? Interprete su resultado.

## 12.5. Inferencia estadística: contrastes de hipótesis e intervalos de confianza

Una vez desarrollados los estimadores de los coeficientes y un estimador de  $\sigma^2$ , estamos ya en condiciones de hacer inferencias relativas al modelo poblacional. El enfoque básico es paralelo al de los Capítulos 8 a 11. Desarrollamos estimadores de la varianza para los estimadores de los coeficientes,  $b_0$  y  $b_1$ , y utilizamos los parámetros y las varianzas estimados para contrastar hipótesis y para calcular intervalos de confianza utilizando la distribución  $t$  de Student. Las inferencias realizadas a partir del análisis de regresión nos ayudarán a comprender el proceso analizado y a tomar decisiones sobre ese proceso. Suponemos inicialmente que los errores aleatorios del modelo,  $\varepsilon$ , siguen una distribución normal. Más adelante, sustituiremos este supuesto por el del teorema del límite central. Comenzamos desarrollando estimadores de la varianza y formas útiles de contraste. A continuación, los aplicamos utilizando nuestros datos sobre las ventas al por menor.

En el apartado 12.2 definimos la regresión simple correspondiente al modelo poblacional:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

en la que las  $x_i$  tienen valores predeterminados, pero no son variables aleatorias. En los Capítulos 5 y 6 sobre las funciones lineales de variables aleatorias vimos que si  $\varepsilon_i$  es una variable aleatoria que sigue una distribución normal de varianza  $\sigma^2$ , entonces  $y_i$  también sigue una distribución normal que tiene la misma varianza. El segundo miembro es una función lineal de  $X$ , salvo por la variable aleatoria  $\varepsilon_i$ . Si sumamos una función de  $X$  a una

variable aleatoria, no cambiamos la varianza. En el apartado 12.3 observamos que el estimador del coeficiente de la pendiente,  $b_1$ , es

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum \left( \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) y_i \\ &= \sum a_i y_i \end{aligned}$$

donde

$$a_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

En este estimador, vemos que  $b_1$  es una función lineal de la variable aleatoria  $y_i$  cuya varianza es  $\sigma^2$ . Las  $y_i$  son variables aleatorias independientes. Por lo tanto, la varianza de  $b_1$  es una transformación simple de la varianza de  $Y$ . Utilizando los resultados del Capítulo 6, la función lineal puede expresarse de la forma siguiente:

$$\begin{aligned} b_1 &= \sum_{i=1}^n a_i y_i \\ a_i &= \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \sigma_{b_1}^2 &= \sum_{i=1}^n a_i^2 \sigma^2 \\ \sigma_{b_1}^2 &= \sum_{i=1}^n \left( \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} \sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Dado que  $y_i$  sigue una distribución normal y  $b_1$  es una función lineal de variables normales independientes, esta función lineal implica que  $b_1$  también sigue una distribución normal. De este análisis podemos deducir la varianza poblacional y la varianza muestral.

### Distribución en el muestreo del estimador de los coeficientes por mínimos cuadrados

Si se cumplen los supuestos habituales de la estimación por mínimos cuadrados, entonces  $b_1$  es un estimador insesgado de  $\beta_1$  y tiene una varianza poblacional

$$\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2} \quad (12.15)$$

y un estimador insesgado de la varianza muestral

$$s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2} \quad (12.16)$$

El estimador de la constante de la regresión,  $b_0$ , también es una función lineal de la variable aleatoria  $y_i$  y, por lo tanto, puede demostrarse que sigue una distribución normal, y su estimador de la varianza puede obtenerse de la forma siguiente:

$$s_{b_0}^2 = \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) s_e^2$$

Es importante observar que la varianza del coeficiente de la pendiente,  $b_1$ , depende de dos importantes cantidades:

1. La distancia de los puntos con respecto a la recta de regresión medida por  $s_e^2$ . Cuando los valores son más altos, la varianza de  $b_1$  es mayor.
2. La desviación total de los valores de  $X$  con respecto a la media medida por  $(n-1)s_x^2$ . Cuanto mayor es la dispersión de los valores de  $X$ , menor es la varianza del coeficiente de la pendiente.



Estos dos resultados son muy importantes cuando hay que elegir los datos para realizar un modelo de regresión. Antes hemos señalado que cuanto mayor era la dispersión de la variable independiente,  $X$ , mayor era  $R^2$ , lo que indicaba que la relación era más estrecha. Ahora vemos que cuanto mayor es la dispersión de la variable independiente —medida por  $s_x^2$ —, menor es la varianza del coeficiente estimado de la pendiente,  $b_1$ . Por lo tanto, cuanto menores sean los estimadores de la varianza del coeficiente de la pendiente, mejor es el modelo de regresión. También debemos añadir que muchas conclusiones de investigaciones y muchas decisiones de política económica se basan en la variación de  $Y$  que se debe a una variación de  $X$ , estimada por  $b_1$ . Por lo tanto, nos gustaría que la varianza de esta importante variable de decisión,  $b_1$ , fuera lo más pequeña posible.

En el análisis de regresión aplicado, nos gustaría saber primero si existe una relación. En el modelo de regresión, vemos que si  $\beta_1$  es 0, entonces no existe una relación lineal:  $Y$  no aumentaría o disminuiría continuamente cuando aumenta  $X$ . Para averiguar si existe una relación lineal, podemos contrastar la hipótesis

$$H_0: \beta_1 = 0$$

frente a

$$H_1: \beta_1 \neq 0$$

Dado que  $b_1$  sigue una distribución normal, podemos contrastar esta hipótesis utilizando el estadístico  $t$  de Student

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

que se distribuye como una  $t$  de Student con  $n - 2$  grados de libertad. El contraste de hipótesis también puede realizarse con valores de  $\beta_1$  distintos de 0. Una regla práctica es extraer la conclusión de que existe una relación si el valor absoluto del estadístico  $t$  es superior a 2. Este resultado se obtiene exactamente en el caso de un contraste de dos colas con un nivel de significación  $\alpha = 0,05$  y 60 grados de libertad y constituye una buena aproximación cuando  $n > 30$ .

### Base para la inferencia sobre la pendiente de la regresión poblacional

Sea  $\beta_1$  la pendiente de la ecuación poblacional y  $b_1$  su estimación por mínimos cuadrados basada en  $n$  pares de observaciones muestrales. En ese caso, si se cumplen los supuestos habituales del modelo de regresión y puede suponerse también que los errores,  $\varepsilon_p$ , siguen una distribución normal, la variable aleatoria

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad (12.17)$$

se distribuye como una  $t$  de Student con  $(n - 2)$  grados de libertad. Además, el teorema del límite central nos permite concluir que este resultado es aproximadamente válido para una amplia variedad de distribuciones no normales y muestras de un tamaño suficientemente grande,  $n$ .

La mayoría de los programas que se emplean para estimar regresiones calculan normalmente la desviación típica de los coeficientes y el estadístico  $t$  de Student para  $\beta_1 = 0$ . La Figura 12.8 muestra las salidas Minitab y Excel correspondientes al ejemplo de las ventas al por menor.

En el caso del modelo de las ventas al por menor, el coeficiente de la pendiente es  $b_1 = 0,382$  con una desviación típica  $s_{b_1} = 0,02529$ . Para saber si existe relación entre las ventas al por menor,  $Y$ , y la renta disponible,  $X$ , podemos contrastar la hipótesis

$$H_0: \beta_1 = 0$$

frente a

$$H_1: \beta_1 \neq 0$$

En la hipótesis nula, el cociente entre el estimador del coeficiente,  $b_1$ , y su desviación típica sigue una distribución  $t$  de Student. En el ejemplo de las ventas al por menor, observamos que el estadístico  $t$  de Student calculado es

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{0,38152 - 0}{0,02529} = 15,08$$

El estadístico  $t$  de Student resultante,  $t = 15,08$ , mostrado en la salida del análisis de regresión, constituye una prueba contundente para rechazar la hipótesis nula y concluir que existe una estrecha relación entre las ventas al por menor y la renta disponible. También

## Results for: retail sales.MTW

## Regression Analysis: Y Retail Sales versus X Income

The regression equation is

$$Y \text{ Retail Sales} = 1922 + 0.382 X \text{ Income}$$

Predictor	Coef	SE Coef	T	P
Constant	1922.4	274.9	6.99	0.000
X Income	0.38152	0.02529	15.08	0.000

 $t_{b_1}$ , Estadístico  $t$  de Student $s_{b_1}$ , Error típico del coeficiente de la pendiente

S = 147.670 R-Sq = 91.9% R-Sq(adj) = 91.5%

 $s_e$ , Error típico de la estimación

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4961434	4961434	227.52	0.000
Residual Error	20	436127	21806		
Total	21	5397561			

 $s_e^2$ , Varianza del error del modelo

SCR, Suma de los cuadros de la regresión

SCE, Suma de los cuadros de los errores

## Unusual Observations

Obs	X	Income	Y	Retail Sales	Fit	SE Fit	Residual	St Resid
12		11307		5907.0	6236.2	34.0	-329.2	-2.29R

R denotes an observation with a large standardized residual.

 $b_1$ , Coeficiente de la pendiente

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6597181					
8	Observations	22					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	4961434.408	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08305	2.17E-12	0.328756343	0.4342771
19							

$s_e$ , Error típico de la estimación

SCR, Suma de los cuadrados de la regresión

SCE, Suma de los cuadrados de los errores

$s_e$ , Varianza de los errores

 $s_e$ , Error típico de la estimación

SCR, Suma de los cuadros de la regresión

SCE, Suma de los cuadros de los errores

 $s_e$ , Varianza del error del modelo $t_{b_1}$ , Estadístico  $t$  de Student $s_{b_1}$ , Error típico del coeficiente de la pendiente $b_1$ , Coeficiente de la pendiente

(b)

Figura 12.8. Modelos de ventas al por menor: estimadores de las varianzas de los coeficientes:

(a) salida Minitab; (b) salida Excel.

señalamos que el  $p$ -valor de  $b_1$  es 0,000, lo que es una prueba alternativa de que  $\beta_1$  no es igual a 0. Recuérdese que en el Capítulo 10 vimos que el  $p$ -valor es el menor nivel de significación al que puede rechazarse la hipótesis nula.

También podrían realizarse contrastes de hipótesis relativos a la constante de la ecuación,  $b_0$ , utilizando la desviación típica desarrollada antes y mostrada en la salida Minitab. Sin embargo, como normalmente nos interesan las tasas de variación —medidas por  $b_1$ —, los contrastes relativos a la constante generalmente son menos importantes.

Si el tamaño de la muestra es lo suficientemente grande para que se aplique el teorema del límite central, podemos realizar esos contrastes de hipótesis aunque los errores,  $\varepsilon_i$ , no sigan una distribución normal. La cuestión clave es la distribución de  $b_1$ . Si  $b_1$  sigue una distribución normal aproximada, es posible realizar el contraste de hipótesis.

### Contrastes de la pendiente de la regresión poblacional

Si los errores de la regresión,  $\varepsilon_p$ , siguen una distribución normal y se cumplen los supuestos habituales del método de los mínimos cuadrados (o si la distribución de  $b_1$  es aproximadamente normal), los siguientes contrastes tienen un nivel de significación  $\alpha$ .

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \beta_1 = \beta_1^* \quad \text{o} \quad H_0: \beta_1 \leq \beta_1^*$$

frente a la hipótesis alternativa

$$H_1: \beta_1 > \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{b_1 - \beta_1^*}{s_{b_1}} \geq t_{n-2, \alpha} \quad (12.18)$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \beta_1 = \beta_1^* \quad \text{o} \quad H_0: \beta_1 \geq \beta_1^*$$

frente a la hipótesis alternativa

$$H_1: \beta_1 < \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{b_1 - \beta_1^*}{s_b} \leq -t_{n-2, \alpha} \quad (12.19)$$

3. Para contrastar la hipótesis nula

$$H_0: \beta_1 = \beta_1^*$$

frente a la hipótesis alternativa bilateral

$$H_1: \beta_1 \neq \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{b_1 - \beta_1^*}{s_{b_1}} \geq t_{n-2, \alpha/2} \quad \text{o} \quad \frac{b_1 - \beta_1^*}{s_{b_1}} \leq -t_{n-2, \alpha/2} \quad (12.20)$$

Podemos obtener intervalos de confianza para la pendiente  $\beta_1$  de la ecuación poblacional utilizando los estimadores de los coeficientes y de las varianzas que hemos desarrollado y el razonamiento realizado en el Capítulo 8.



### Intervalos de confianza de la pendiente de la regresión poblacional $b_1$

Si los errores de la regresión,  $\varepsilon_i$ , siguen una distribución normal y se cumplen los supuestos habituales del análisis de regresión, se obtiene un intervalo de confianza al  $100(1 - \alpha)\%$  de la pendiente de la recta de regresión poblacional  $\beta_1$  de la forma siguiente:

$$b_1 - t_{n-2, \alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2, \alpha/2} s_{b_1} \quad (12.21)$$

donde  $t_{n-2, \alpha/2}$  es el número para el que

$$P(t_{n-2} > t_{n-2, \alpha/2}) = \alpha/2$$

y la variable aleatoria  $t_{n-2}$  sigue una distribución  $t$  de Student con  $(n - 2)$  grados de libertad.

En la salida del análisis de regresión de las ventas al por menor con respecto a la renta disponible de la Figura 12.8, vemos que

$$n = 22 \quad b_1 = 0,3815 \quad s_b = 0,0253$$

Para obtener el intervalo de confianza al 99 por ciento de  $\beta_1$ , tenemos  $1 - \alpha = 0,99$  y  $n - 2 = 20$  grados de libertad y, por lo tanto, vemos en la tabla 8 del apéndice que

$$t_{n-2, \alpha/2} = t_{20, 0,005} = 2,845$$

Por lo tanto, tenemos el intervalo de confianza al 99 por ciento

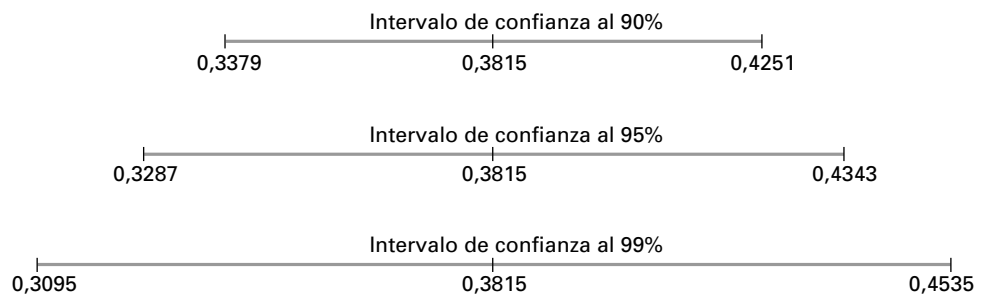
$$0,3815 - (2,845)(0,0253) < \beta_1 < 0,3815 + (2,845)(0,0253)$$

o sea

$$0,3095 < \beta_1 < 0,4535$$

Vemos que el intervalo de confianza al 99 por ciento del aumento esperado de las ventas al por menor por hogar que acompaña a un aumento de la renta disponible por hogar de 1 \$ abarca el intervalo de 0,3095 \$ a 0,4535 \$. La Figura 12.9 muestra los intervalos de confianza al 90, al 95 y al 99 por ciento de la pendiente de la regresión poblacional.

**Figura 12.9.** Intervalos de confianza de la pendiente de la recta de regresión poblacional de las ventas al por menor a los niveles de confianza del 90, el 95 y el 99 por ciento.



## Contraste de hipótesis del coeficiente de la pendiente poblacional utilizando la distribución $F$

Existe otro contraste de la hipótesis de que el coeficiente de la pendiente,  $\beta_1$ , es igual a 0:

$$\begin{aligned}H_0: \beta_1 &= 0 \\H_1: \beta_1 &\neq 0\end{aligned}$$

Este contraste se basa en la descomposición de la variabilidad que hemos presentado en el apartado 12.4. Este contraste parte del supuesto de que, si la hipótesis nula es verdadera, entonces pueden utilizarse tanto  $SCE$  como  $SCR$  para obtener estimadores independientes de la varianza del error del modelo  $\sigma^2$ . Para realizar este contraste, obtenemos dos estimaciones muestrales de la desviación típica poblacional  $\sigma$ , que se denominan términos cuadráticos medios. La suma de los cuadrados de la regresión,  $SCR$ , tiene un grado de libertad, ya que se refiere al coeficiente de la pendiente, y el cuadrado medio de la regresión,  $CMR$ , es

$$CMR = \frac{SCR}{1} = SCR$$

Si la hipótesis nula —ausencia de relación— es verdadera, entonces  $CMR$  es una estimación de la varianza global del modelo,  $\sigma^2$ . También utilizamos la suma de los cuadrados de los errores al igual que antes para hallar el error cuadrático medio,  $ECM$ :

$$ECM = \frac{SCE}{n - 2} = s_e^2$$

En el apartado 11.4 introdujimos la distribución  $F$ , que era el cociente entre estimaciones muestrales independientes de la varianza, dadas varianzas poblacionales iguales. Puede demostrarse que  $CMR$  y  $ECM$  son independientes y que en  $H_0$  ambas son estimaciones de la varianza poblacional,  $\sigma^2$ . Por lo tanto, si  $H_0$  es verdadera, podemos demostrar que el cociente

$$F = \frac{CMR}{ECM} = \frac{SCR}{s_e^2}$$

sigue una distribución  $F$  con 1 grado de libertad en el numerador y  $n - 2$  grados de libertad en el denominador. También debe señalarse que el estadístico  $F$  es igual al cuadrado del estadístico  $t$  del coeficiente de la pendiente. Esta afirmación puede demostrarse algebraicamente. Aplicando la teoría de la distribución, podemos demostrar que una  $t$  de Student al cuadrado con  $n - 2$  grados de libertad y la  $F$  con 1 grado de libertad en el numerador y  $n - 2$  grados de libertad en el denominador son iguales:

$$F_{\alpha, 1, n-2} = t_{\alpha/2, n-2}^2$$

La Figura 12.8(a) muestra el análisis de varianza de la regresión de las ventas al por menor procedente de la salida Minitab. En nuestro ejemplo de las ventas al por menor, la

suma de los cuadrados de los errores se divide por los 20 grados de libertad para calcular el *ECM*:

$$ECM = \frac{436.127}{20} = 21.806$$

A continuación, se calcula el cociente *F*, que es como el cociente entre dos cuadrados medios:

$$F = \frac{CMR}{ECM} = \frac{4.961.434}{21.806} = 227,52$$

Este cociente *F* es considerablemente mayor que el valor crítico de  $\alpha = 0,01$  con 1 grado de libertad en el numerador y 20 grados de libertad en el denominador ( $F_{1,20,0,01} = 8,10$ ) según la Tabla 9 del apéndice. La salida Minitab —Figura 12.8(a)— de la regresión de las ventas al por menor muestra que el *p*-valor de esta *F* calculada es 0,000, lo que constituye una prueba alternativa para rechazar  $H_0$ . Obsérvese también que el estadístico *F* es igual a  $t^2$ , siendo *t* el estadístico del coeficiente de la pendiente,  $b_1$ :

$$F = t^2$$

$$227,52 = 15,08^2$$

### Contraste *F* del coeficiente de regresión simple

Podemos contrastar la hipótesis

$$H_0: \beta_1 = 0$$

frente a la alternativa

$$H_1: \beta_1 \neq 0$$

utilizando el estadístico *F*

$$F = \frac{CMR}{ECM} = \frac{SCR}{s_e^2} \quad (12.22)$$

La regla de decisión es

$$\text{Rechazar } H_0 \text{ si } F \geq F_{1,n-2,\alpha} \quad (12.23)$$

También podemos mostrar que el estadístico *F* es

$$F = t_{b_1}^2 \quad (12.24)$$

en cualquier análisis de regresión simple.

Este resultado muestra que los contrastes de hipótesis relativos al coeficiente de la pendiente poblacional dan exactamente el mismo resultado cuando se utiliza la *t* de Student que cuando se utiliza la distribución *F*. En el Capítulo 13 veremos que la distribución *F* —cuando se utiliza en un análisis de regresión múltiple— también brinda la oportunidad de contrastar la hipótesis de que varios coeficientes poblacionales de la pendiente son simultáneamente iguales a 0.

## EJERCICIOS

### Ejercicios básicos

**12.35.** Dado el modelo de regresión simple

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

y los resultados de la regresión siguientes, contraste la hipótesis nula de que el coeficiente de la pendiente es 0 frente a la hipótesis alternativa de que es mayor que cero utilizando la probabilidad de cometer un error de Tipo I igual a 0,05 y halle los intervalos de confianza bilaterales al 95 y al 99 por ciento.

- a) Una muestra aleatoria de tamaño  $n = 38$  con  $b_1 = 5$  y  $s_{b_1} = 2,1$
- b) Una muestra aleatoria de tamaño  $n = 46$  con  $b_1 = 5,2$  y  $s_{b_1} = 2,1$
- c) Una muestra aleatoria de tamaño  $n = 38$  con  $b_1 = 2,7$  y  $s_{b_1} = 1,87$
- d) Una muestra aleatoria de tamaño  $n = 29$  con  $b_1 = 6,7$  y  $s_{b_1} = 1,8$

**12.36.** Utilice un modelo de regresión simple para contrastar la hipótesis

$$H_0: \beta_1 = 0$$

frente a

$$H_1: \beta_1 \neq 0$$

suponiendo que  $\alpha = 0,05$ , dados los siguientes estadísticos de la regresión:

- a) El tamaño de la muestra es 35,  $STC = 100.000$  y la correlación entre  $X$  e  $Y$  es 0,46.
- b) El tamaño de la muestra es 61,  $STC = 123.000$  y la correlación entre  $X$  e  $Y$  es 0,65.
- c) El tamaño de la muestra es 25,  $STC = 128.000$  y la correlación entre  $X$  e  $Y$  es 0,69.

### Ejercicios aplicados

**12.37.** Considere la regresión lineal de las ventas del sistema DVD con respecto al precio del ejercicio 12.29.

- a) Utilice un método de estimación insesgado para hallar una estimación de la varianza de los términos de error en la regresión poblacional.
- b) Utilice un método de estimación insesgado para hallar una estimación de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
- c) Halle el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.

**12.38.** Una cadena de comida rápida decidió realizar un experimento para averiguar la influencia de los gastos publicitarios en las ventas. Se introdujeron diferentes cambios relativos en los gastos publicitarios en comparación con el año anterior en ocho regiones del país y se observaron los cambios que experimentaron las ventas como consecuencia. La tabla adjunta muestra los resultados.

Aumento de los gastos publicitarios (%)	0	4	14	10	9	8	6	1
Aumento de las ventas (%)	2,4	7,2	10,3	9,1	10,2	4,1	7,6	3,5

- a) Estime por mínimos cuadrados la regresión lineal del aumento de las ventas con respecto al aumento de los gastos publicitarios.
- b) Halle el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.

**12.39.** Un vendedor de bebidas alcohólicas al por mayor tiene interés en averiguar cómo afecta el precio de un whisky escocés a la cantidad vendida. En una muestra aleatoria de datos sobre las ventas de ocho semanas se obtuvieron los resultados de la tabla adjunta sobre el precio, en dólares, y las ventas, en cajas.

Precio	19,2	20,5	19,7	21,3	20,8	19,9	17,8	17,2
Ventas	25,4	14,7	18,6	12,4	11,1	15,7	29,2	35,2

Halle el intervalo de confianza al 95 por ciento de la variación esperada de las ventas provocada por una subida del precio de 1 \$.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

**12.40.** Continúe el análisis del ejercicio 12.30 de la regresión de la variación porcentual del índice Dow-Jones en un año con respecto a la variación porcentual del índice en los cinco primeros días de sesión del año. Utilice el fichero de datos **Dow Jones**.

- a) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de los términos de error de la regresión poblacional.

- b) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
  - c) Halle e interprete el intervalo de confianza al 95 por ciento de la pendiente de la recta de regresión poblacional.
  - d) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.
- 12.41.** Considere el modelo de las pérdidas experimentadas por los fondos de inversión el 13 de noviembre de 1980 del ejercicio 12.24. Utilice el fichero de datos **New York Stock Exchange Gains and Losses**.
- a) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de los términos de error de la regresión poblacional.
  - b) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
  - c) Halle los intervalos de confianza al 90, al 95 y al 99 por ciento de la pendiente de la recta de regresión poblacional.

## 12.6. Predicción

Los modelos de regresión pueden utilizarse para hacer predicciones o previsiones sobre la variable dependiente, partiendo de un valor futuro supuesto de la variable independiente. Supongamos que queremos predecir el valor de la variable dependiente, dado que la variable independiente es igual a un valor específico,  $x_{n+1}$ , y que la relación lineal entre la variable dependiente y la variable independiente continúa manteniéndose. El valor correspondiente de la variable dependiente será, entonces,

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

que, dado  $x_{n+1}$ , tiene la esperanza

$$E[y_{n+1} | x_{n+1}] = \beta_0 + \beta_1 x_{n+1}$$

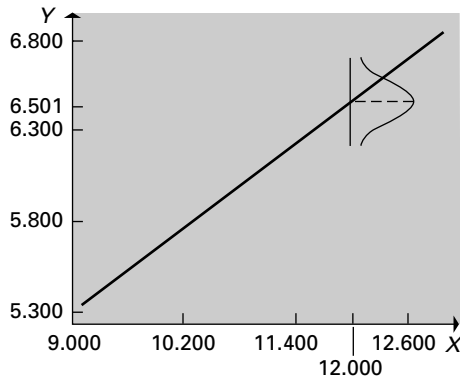
Existen dos opciones interesantes:

1. Podemos querer estimar el valor efectivo que se obtendrá con una única observación,  $y_{n+1}$ . Esta opción se muestra en la Figura 12.10.
2. Podemos querer estimar el valor esperado condicionado,  $E[y_{n+1} | x_{n+1}]$ , es decir, el valor medio de la variable dependiente cuando la variable independiente es fija e igual a  $x_{n+1}$ . Esta opción se muestra en la Figura 12.11.

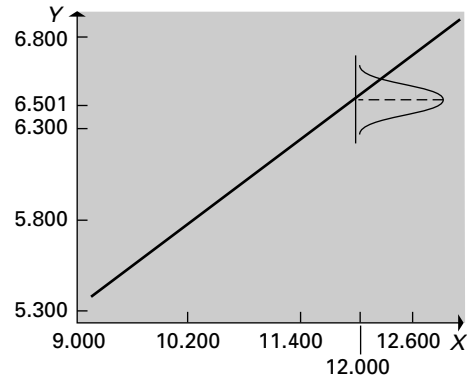
Dado que los supuestos habituales del análisis de regresión continúan cumpliéndose, se obtiene la misma estimación puntual en las dos opciones. Sustituimos simplemente los  $\beta_0$  y  $\beta_1$  desconocidos por sus estimaciones por mínimos cuadrados,  $b_0$  y  $b_1$ . Es decir, estimamos  $(\beta_0 + \beta_1 x_{n+1})$  por medio de  $(b_0 + b_1 x_{n+1})$ . Sabemos que el estimador correspondiente es el mejor estimador insesgado lineal de  $Y$ , dado  $X$ . En la primera opción, nos interesa saber cuál es la mejor predicción de una observación del proceso. Pero en la segunda opción, nos interesa saber cuál es el valor esperado o media a largo plazo del proceso. En ambas opciones, un buen estimador puntual con nuestros supuestos es

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

ya que no sabemos nada útil sobre la variable aleatoria,  $\varepsilon_{n+1}$ , salvo que su media es 0. Por lo tanto, sin otra información utilizaremos 0 como estimación puntual.



**Figura 12.10.** Recta de regresión estimada por mínimos cuadrados de las ventas al por menor con respecto a la renta disponible: aplicación a un único valor observado.



**Figura 12.11.** Recta de regresión estimada por mínimos cuadrados de las ventas al por menor con respecto a la renta disponible: valor esperado.

Sin embargo, normalmente queremos intervalos, además de estimaciones puntuales, y para eso las dos opciones son diferentes, ya que los estimadores de la varianza de dos cantidades diferentes estimadas son diferentes. Los resultados de estos estimadores diferentes de la varianza llevan a los dos intervalos diferentes. En la primera opción, el intervalo generalmente es un intervalo de predicción porque estamos prediciendo el valor de un único punto. El intervalo de la segunda opción es un intervalo de confianza porque es el intervalo del valor esperado.

### Intervalos de confianza de las predicciones e intervalos de predicción

Supongamos que el modelo de regresión poblacional es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \dots, n + 1)$$

que se cumplen los supuestos habituales del análisis de regresión y que los  $\varepsilon_i$  siguen una distribución normal. Sean  $b_0$  y  $b_1$  las estimaciones por mínimos cuadrados de  $\beta_0$  y  $\beta_1$ , basadas en  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . En ese caso, puede demostrarse que los intervalos al  $100(1 - \alpha)\%$  son los siguientes:

1. Para la predicción del valor efectivo resultante de  $Y_{n+1}$ , el intervalo de predicción es

$$\hat{y}_{n+1} \pm t_{n-2, \alpha/2} \sqrt{\left[ 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] s_e} \quad (12.25)$$

2. Para la predicción de la esperanza condicional  $E(Y_{n+1}|x_{n+1})$ , el intervalo de confianza es

$$\hat{y}_{n+1} \pm t_{n-2, \alpha/2} \sqrt{\left[ \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] s_e} \quad (12.26)$$

donde

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{y} \quad \hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$



**Retail  
Sales**

### EJEMPLO 12.3. Predicción de las ventas al por menor (predicción basada en un modelo de regresión)

Mostramos cómo se calculan los intervalos utilizando el ejemplo 12.2 sobre las ventas al por menor y la renta disponible. Le han pedido que haga una predicción de los valores de las ventas al por menor por hogar cuando la renta disponible por hogar es de 12.000 \$: el valor efectivo del año que viene y el valor esperado a largo plazo. También le han pedido que calcule intervalos de predicción e intervalos de confianza para estas predicciones. Utilice el fichero de datos **Retail Sales**.

#### Solución

Los valores predichos para el próximo año y para el largo plazo son

$$\begin{aligned}\hat{y}_{n+1} &= b_0 + b_1 x_{n+1} \\ &= 1.922 + (0,3815)(12.000) = 6.501\end{aligned}$$

Por lo tanto, observamos que las ventas estimadas son de 6.501 \$ cuando la renta disponible es de 12.000 \$. También observamos que

$$n = 22 \quad \bar{x} = 10.799 \quad \sum (x_i - \bar{x})^2 = 34.110.178 \quad s_e^2 = 21.806$$

Por lo tanto, el error típico de una única observación predicha de  $Y$  es

$$\sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]} s_e = \sqrt{\left[1 + \frac{1}{22} + \frac{(12.000 - 10.799)^2}{34.110.178}\right]} \sqrt{21.806} = 154,01$$

Asimismo, observamos que el error típico del valor esperado de  $Y$  es

$$\sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]} s_e = \sqrt{\left[\frac{1}{22} + \frac{(12.000 - 10.799)^2}{34.110.178}\right]} \sqrt{21.806} = 43,76$$

Supongamos que se necesitan intervalos del 95 por ciento para las predicciones suponiendo que  $\alpha = 0,05$  y

$$t_{n-2, \alpha/2} = t_{20, 0,025} = 2,086$$

Utilizando estos resultados, observamos que el intervalo de predicción al 95 por ciento para las ventas al por menor del próximo año cuando la renta disponible es de 12.000 \$ se calcula de la forma siguiente:

$$\begin{aligned}6.501 \pm (2,086)(154,01) \\ 6.501 \pm 321\end{aligned}$$

Por lo tanto, el intervalo de predicción al 95 por ciento para las ventas de un único año en el que la renta es de 12.000 \$ va de 6.180 \$ a 6.822 \$.

En el caso del intervalo de confianza del valor esperado de las ventas al por menor cuando la renta disponible es de 12.000 \$, tenemos que

$$\begin{aligned}6.501 \pm (2,086)(43,76) \\ 6.501 \pm 91\end{aligned}$$

Por lo tanto, el intervalo de confianza al 95 por ciento del valor esperado va de 6.410 \$ a 6.592 \$.



Las Figuras 12.10 y 12.11 muestran la distinción entre estos dos problemas de estimación de intervalos. Vemos en ambas figuras la recta de regresión estimada para nuestros datos sobre las ventas al por menor y la renta disponible. También vemos en la Figura 12.10 una función de densidad que representa nuestra incertidumbre sobre el valor que tomarán las ventas al por menor en cualquier año específico en el que la renta disponible sea de 12.000 \$. La función de densidad de la Figura 12.11 representa nuestra incertidumbre sobre las ventas al por menor esperadas o medias en los años en los que la renta disponible es de 12.000 \$. Naturalmente, tenemos más incertidumbre sobre las ventas de un único año que sobre las ventas medias y eso se refleja en la forma de las dos funciones de densidad. Vemos que ambas están centradas en las ventas al por menor de 6.501 \$, pero que la función de densidad de la Figura 12.10 tiene una dispersión mayor. Como consecuencia, el intervalo de predicción de un valor específico es mayor que el intervalo de confianza de las ventas al por menor esperadas.

Podemos extraer algunas conclusiones más estudiando las formas generales de los intervalos de predicción y de confianza. Como hemos visto, cuanto más amplio es el intervalo, mayor es la incertidumbre sobre la predicción puntual. Basándonos en estas fórmulas, hacemos cuatro observaciones:

1. Manteniéndose todo lo demás constante, cuanto mayor es el tamaño de la muestra  $n$ , más estrecho es el intervalo de confianza. Vemos, pues, que cuanto más información muestral tengamos, más seguros estaremos de nuestra inferencia.
2. Manteniéndose todo lo demás constante, cuanto mayor es  $s_e^2$ , más amplio es el intervalo de confianza. Una vez más, es de esperar, ya que  $s_e^2$  es una estimación de  $\sigma^2$ , la varianza de los errores de la regresión,  $\varepsilon_i$ . Dado que estos errores

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

representan la discrepancia entre los valores observados de las variables dependientes y sus esperanzas, dadas las variables independientes, cuanto mayor es la magnitud de esta discrepancia, más imprecisa será nuestra inferencia.

3. Consideremos ahora la cantidad  $(\sum_{i=1}^n (x_i - \bar{x})^2)$ . Esta cantidad es simplemente un múltiplo de la varianza muestral de las observaciones de la variable independiente. Cuando la varianza es grande, significa que tenemos información sobre un amplio rango de valores de esta variable, lo que nos permite hacer estimaciones más precisas de la recta de regresión poblacional y, por lo tanto, calcular intervalos de confianza más reducidos.
4. También vemos que cuanto mayores son los valores de la cantidad  $(x_{n+1} - \bar{x})^2$ , más amplios son los intervalos de confianza de las predicciones. Por lo tanto, los intervalos de confianza son más amplios a medida que nos alejamos de la media de la variable independiente,  $X$ . Dado que nuestros datos muestrales están centrados en la media  $\bar{x}$ , es de esperar que podamos hacer inferencias más definitivas cuando la variable independiente está relativamente cerca de este valor central que cuando está a alguna distancia de él.



No se recomienda extrapolar la ecuación de regresión fuera del rango de los datos utilizados para realizar la estimación. Supongamos que se nos pide que hagamos una predicción de las ventas al por menor por hogar en un año en el que la renta disponible es de 30.000 \$. Volviendo a los datos de la Tabla 12.1 y a la recta de regresión de la Figura 12.11, vemos que 30.000 \$ se encuentra muy fuera del rango de los datos utilizados para



desarrollar el modelo de regresión. Un analista sin experiencia podría utilizar los métodos antes presentados para hacer una predicción o estimar un intervalo de confianza. En las ecuaciones podemos ver que los intervalos resultantes serían muy amplios y, por lo tanto, la predicción tendría escaso valor. Sin embargo, las predicciones que se realizan fuera del rango de los datos originales plantean un problema más fundamental: no tenemos sencillamente ninguna prueba que indique cómo es la naturaleza de la relación fuera del rango de los datos. No hay ninguna razón en la teoría económica que exija absolutamente que la relación siga siendo lineal con la misma tasa de variación cuando nos salimos del rango de los datos utilizados para estimar los coeficientes del modelo de regresión. Cualquier extrapolación del modelo fuera del rango de los datos para predecir valores debe basarse en otra información o evidencia, además de la que contiene el análisis de regresión basado en los datos de que se dispone. Cuando los analistas intentan hacer este tipo de extrapolación, pueden cometer graves errores.

## EJERCICIOS

### Ejercicios básicos

- 12.45.** Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 12 + 5x_i$$

y

$$s_e = 9,67 \quad \bar{x} = 8 \quad n = 32 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 500$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que  $x = 13$ .

- 12.43.** Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 14 + 7x_i$$

y

$$s_e = 7,45 \quad \bar{x} = 8 \quad n = 25 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 300$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que  $x = 12$ .

- 12.44.** Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 22 + 8x_i$$

y

$$s_e = 3,45 \quad \bar{x} = 11 \quad n = 22 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 400$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que  $x = 17$ .

- 12.45.** Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 8 + 10x_i$$

y

$$s_e = 11,23 \quad \bar{x} = 8 \quad n = 44 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 800$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que  $x = 17$ .

### Ejercicios aplicados

- 12.46.** Se toma una muestra de 25 obreros de una fábrica. Se pide a cada obrero que valore su satisfacción en el trabajo ( $x$ ) en una escala de 1 a 10. Se averigua también el número de días que estos obreros estuvieron ausentes del trabajo ( $y$ ) el año pasado. Se estima la recta de regresión muestral por mínimos cuadrados para estos datos.

$$\hat{y} = 12,6 - 1,2x$$

También se ha observado que

$$\bar{x} = 6,0 \quad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130,0 \quad SCE = 80,6$$

- a) Contraste al nivel de significación del 1 por ciento la hipótesis nula de que la satisfacción en el trabajo no produce un efecto lineal en el absentismo frente a una hipótesis alternativa bilateral adecuada.
- b) Un obrero tiene un nivel de satisfacción en el trabajo de 4. Halle un intervalo al 90 por

ciento del número de días que este obrero estaría ausente del trabajo en un año.

- 12.47.** Los médicos tienen interés en saber qué relación existe entre la dosis de un medicamento y el tiempo que necesita un paciente para recuperarse. La tabla adjunta muestra las dosis (en gramos) y el tiempo de recuperación (en horas) de una muestra de cinco pacientes. Estos pacientes tienen parecidas características, salvo la dosis del medicamento administrada.

Dosis	1,2	1,0	1,5	1,2	1,4
Tiempo de recuperación	25	40	10	27	16

- a) Estime la regresión lineal del tiempo de recuperación con respecto a la dosis.  
 b) Halle e interprete el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.  
 c) ¿Sería útil la regresión muestral obtenida en el apartado (a) para predecir el tiempo de recuperación de un paciente al que se le administran 2,5 gramos de este medicamento? Explique su respuesta.

- 12.48.** En el caso del problema de la tasa de rendimiento de las acciones del ejercicio 12.20, se observó que

$$\sum_{i=1}^{20} y_i^2 = 196,2$$

- a) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que es positiva.  
 b) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 1 frente a la hipótesis alternativa bilateral.  
**12.49.** Utilizando los datos del ejercicio 12.21, contraste la hipótesis nula de que las ventas semanales de los representantes no están relacionadas linealmente con su puntuación en el test de aptitud frente a la hipótesis alternativa de que existe una relación positiva.  
**12.50.** Vuelva a los datos del ejercicio 12.41. Contraste la hipótesis nula de que las pérdidas que experimentaron los fondos de inversión el viernes 13 de noviembre de 1989 no dependían linealmente de las ganancias obtenidas anterior-

mente en 1989 frente a la hipótesis alternativa bilateral.

- 12.51.** Sea  $r$  la correlación muestral entre un par de variables aleatorias.

- a) Demuestre que

$$\frac{1 - r^2}{n - 2} = \frac{s_e^2}{STC}$$

- b) Utilizando el resultado del apartado (a), demuestre que

$$\frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{b}{s_e / \sqrt{\sum (x_i - \bar{x})^2}}$$

- c) Utilizando el resultado del apartado (b), deduzca que el contraste de la hipótesis nula de la correlación poblacional 0, presentado en el apartado 12.1, es igual que el contraste de la pendiente de la recta de regresión poblacional 0, presentado en el apartado 12.5.

- 12.52.** En el problema del ejercicio 12.22 sobre las ventas de cerveza en los restaurantes se observó que

$$\frac{\sum (y_i - \bar{y})^2}{n - 1} = 250$$

Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.

- 12.53.** En una muestra de 74 observaciones mensuales, se estimó la regresión del rendimiento porcentual del oro ( $y$ ) con respecto a la variación porcentual del índice de precios ( $x$ ). La recta de regresión muestral, obtenida por mínimos cuadrados, era


$$y = -0,003 + 1,11x$$

La desviación típica estimada de la pendiente de la recta de regresión poblacional era 2,31. Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que la pendiente es positiva.

- 12.54.** Vuelva a los datos del ejercicio 12.39. Contraste al nivel del 5 por ciento la hipótesis nula de que las ventas no dependen linealmente del precio de este whisky escocés frente a la hipótesis alternativa bilateral apropiada.

- 12.55.** Vuelva a los datos del ejercicio 12.29.

- a) Halle una estimación puntual del volumen de ventas cuando el precio del sistema DVD es de 480 \$ en una región dada.

- b) Si el precio del sistema se fija en 480 \$, halle intervalos de confianza al 95 por ciento del volumen efectivo de ventas en una región y el número esperado de ventas en esa región.
- 12.56.** Continúe con el análisis del ejercicio 12.7. Si el índice Dow-Jones sube un 1,0 por ciento en los cinco primeros días de sesión de un año, halle intervalos de confianza al 90 por ciento de la variación porcentual *efectiva* y la *esperada* del índice en todo el año. Analice la distinción entre estos intervalos.
- 12.57.**  Vuelva a los datos del ejercicio 12.25 (archivo de datos **Employee Absence**). Halle para un año en el que no varía la tasa de desempleo intervalos de confianza al 90 por ciento de la variación *efectiva* de la tasa media de absentismo laboral por enfermedad y de la variación *esperada*.
- 12.58.** Utilice los datos del ejercicio 12.20 para hallar intervalos de confianza al 90 y al 95 por ciento del rendimiento esperado de las acciones de la empresa cuando la tasa de rendimiento del índice Standard and Poor's 500 es del 1 por ciento.
- 12.59.** Un nuevo representante de ventas de la empresa del ejercicio 12.21 obtiene 70 puntos en el test de aptitud. Halle intervalos de confianza al 80 y al 90 por ciento del valor de las ventas semanales que conseguirá.

## 12.7. Análisis gráfico

---

Hemos desarrollado los métodos teóricos y analíticos que permiten realizar análisis de regresión y construir modelos lineales. Utilizando contrastes de hipótesis e intervalos de confianza, podemos averiguar la calidad de nuestro modelo e identificar algunas relaciones importantes. Estos métodos inferenciales suponen inicialmente que los errores del modelo siguen una distribución normal. Pero también sabemos que el teorema del límite central nos ayuda a realizar contrastes de hipótesis y a construir intervalos de confianza mientras las distribuciones muestrales de los estimadores de los coeficientes y los valores predichos sean aproximadamente normales. El modelo de regresión también se basa en un conjunto de supuestos. Sin embargo, las aplicaciones del análisis de regresión pueden ser erróneas por muchas razones, incluidos los supuestos que no se satisfacen si los datos no siguen las pautas supuestas.

El ejemplo de la regresión de las ventas al por menor con respecto a la renta disponible —Figura 12.1— tiene un diagrama de puntos dispersos que sigue la pauta supuesta en el análisis de regresión. Sin embargo, esa pauta no siempre se produce cuando se estudian nuevos datos. Una de las mejores formas de detectar posibles problemas en el análisis de regresión simple es realizar diagramas de puntos dispersos y observar la pauta. Aquí examinamos algunos instrumentos analíticos y ejemplos de análisis de regresión que pueden ayudarnos a preparar mejores aplicaciones del análisis de regresión.

En este apartado utilizamos el análisis gráfico para mostrar cómo afectan al análisis de regresión los puntos que tienen valores extremos de  $X$  y los puntos que tienen valores de  $Y$  que se desvían considerablemente de la ecuación de regresión por mínimos cuadrados. En capítulos posteriores mostramos cómo puede utilizarse el análisis de los residuos para examinar otras desviaciones con respecto a las pautas normales de los datos.

Los puntos extremos son puntos en los que los valores de  $X$  se desvían considerablemente de los valores de  $X$  de los demás puntos. Volvamos a la ecuación 12.26, que presenta el intervalo de confianza del valor esperado de  $Y$  correspondiente a un valor específico

de  $X$ . Para este intervalo de confianza es fundamental un término llamado normalmente valor de influencia (*leverage*),  $h_i$ , de un punto, que se define de la forma siguiente:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Este valor de influencia aumenta la desviación típica del valor esperado cuando los puntos de datos están más lejos de la media de  $X$  y, por lo tanto, llevan a un intervalo de confianza más amplio. Se considera que un punto  $i$  es un punto extremo si el valor de  $h$  de ese punto es muy diferente de los valores de  $h$  de todos los demás puntos de datos. Vemos en el ejemplo siguiente que el programa Minitab identifica los puntos que tienen un elevado valor de influencia con una  $X$  si  $h_i > 3p/n$ , donde  $p$  es el número de predictores, incluida la constante. La mayoría de los paquetes estadísticos buenos permiten identificar estos puntos, pero no así el programa Excel. Utilizando esta opción, es posible identificar los puntos extremos, como muestra el ejemplo 12.4.

Los puntos atípicos son los puntos que se desvían considerablemente en la dirección de  $Y$  con respecto al valor predicho. Normalmente, estos puntos se identifican calculando el residuo normalizado de la forma siguiente:

$$e_{is} = \frac{e_i}{s_e \sqrt{1 - h_i}}$$

Es decir, el residuo normalizado es el residuo dividido por el error típico del residuo. Obsérvese que en la ecuación anterior los puntos que tienen un elevado valor de influencia —un elevado  $h_i$ — tienen un error típico del residuo menor, porque los puntos que tienen un elevado valor de influencia probablemente influyen en la localización de la recta de regresión estimada y, por lo tanto, el valor observado y el esperado de  $Y$  estarán más cerca. Minitab marca las observaciones que tienen un valor absoluto del residuo normalizado superior a 2,0 con una  $R$  para indicar que son casos atípicos. También las marcan la mayoría de los buenos paquetes estadísticos, pero no el Excel. Utilizando esta opción, es posible identificar los puntos atípicos, como muestra el ejemplo 12.5.



En los dos ejemplos siguientes, veremos que los puntos extremos y los casos atípicos tienen una gran influencia en la ecuación de regresión estimada en comparación con otras observaciones. En cualquier análisis aplicado, estos puntos inusuales forman parte de los datos que representan el proceso estudiado o no forman parte de ellos. En el primer caso, deben incluirse en el conjunto de datos y en el segundo caso no. El analista debe decidir. Normalmente, para tomar estas decisiones hay que comprender bien el proceso y hacer una buena valoración. En primer lugar, debe examinarse detenidamente cada punto y comprobarse su fuente. Estos puntos inusuales podrían deberse a errores de medición o de recogida de datos y, por lo tanto, se eliminarían o se corregirían. Una investigación más profunda puede revelar circunstancias excepcionales que no se espera que formen parte del proceso habitual y eso indicaría la exclusión de los puntos de datos. Las decisiones sobre qué es un proceso habitual y otras decisiones afines exigen una valoración y un examen detenidos de otra información sobre el proceso estudiado. Un buen analista utiliza los cálculos estadísticos anteriores para identificar las observaciones que deben examinarse más detenidamente, pero no se basa exclusivamente en estas medidas de identificación de las observaciones inusuales para tomar la decisión final.

### EJEMPLO 12.4. El efecto de los valores extremos de $X$ (análisis mediante un diagrama de puntos dispersos)

Nos interesa saber cómo afectan los valores extremos de  $X$  a la regresión. En este ejemplo, se analiza el efecto de los puntos que tienen valores de  $X$  que son muy diferentes de los otros puntos utilizando dos muestras que sólo se diferencian en dos puntos. Estos ejemplos comparativos, aunque son algo excepcionales, se utilizan para poner énfasis en el efecto que producen los puntos extremos en un análisis de regresión.

#### Solución

La Figura 12.12 es un diagrama de puntos dispersos con una recta de regresión trazada sobre los puntos y la 12.13 es la salida del análisis de regresión calculada con los datos. La pendiente de la recta de regresión es positiva y  $R^2 = 0,632$ . Pero obsérvese que dos puntos extremos parecen determinar la relación de regresión. Examinemos ahora el efecto de un cambio de los dos puntos de datos extremos, mostrado en las Figuras 12.14 y 12.15.

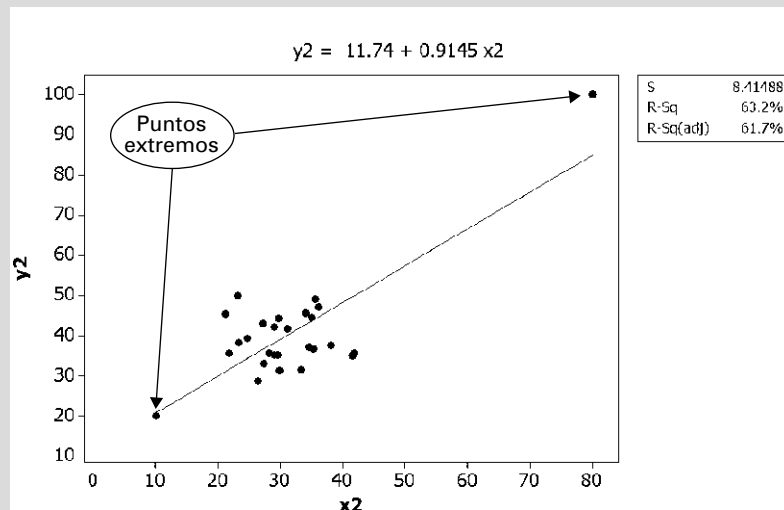


Figura 12.12. Diagrama de puntos dispersos con dos puntos extremos de  $X$ : pendiente positiva.

#### Regression Analysis: Y2 versus x2

The regression equation is  
Y2 = 11.74 + 0.9145 x2

S = 8.41488 R-Sq = 63.2% R-Sq(adj) = 61.7%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3034.80	3034.80	42.86	0.000
Error	25	1770.26	70.81		
Total	26	4805.05			

#### Fitted Line: y2 versus x2

Figura 12.13. Análisis de regresión con dos puntos extremos de  $X$ : pendiente positiva (salida Minitab).

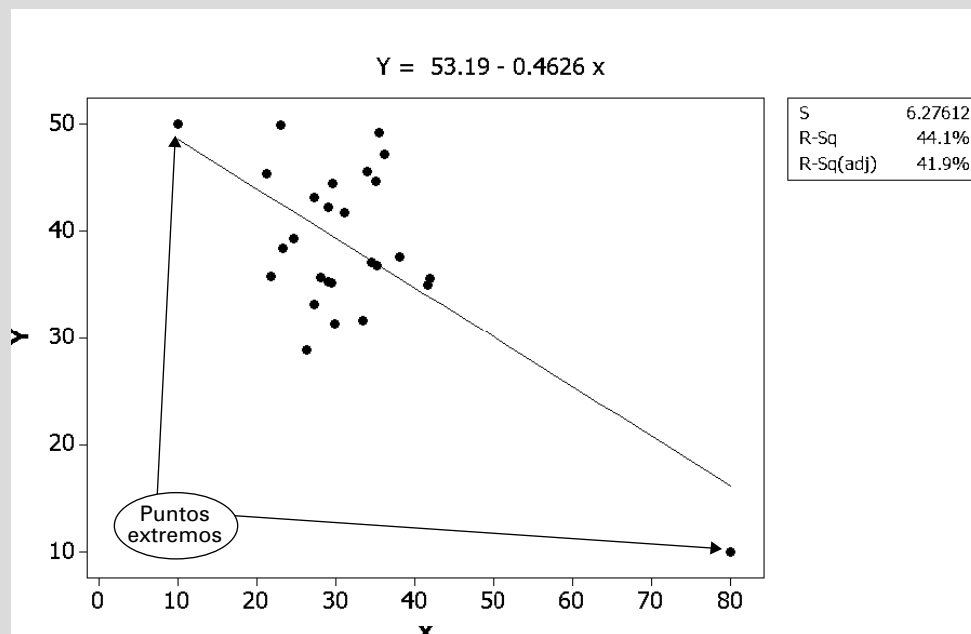


Figura 12.14. Diagrama de puntos dispersos con dos puntos extremos de X: pendiente negativa.

### Regression Analysis: Y versus X

The regression equation is  
Y1 = 53.2 - 0.463 X

Predictor	Coef	SE Coef	T	P
Constant	53.195	3.518	15.12	0.000
X1	-0.4626	0.1042	-4.44	0.000

s = 6.27612 R-Sq = 44.1% R-Sq(adj) = 41.9%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	776.56	776.56	19.71	0.000
Residual Error	25	984.74	39.39		
Total	26	1761.30			

### Unusual Observations

Obs	X	Y	Fit	Se Fit	Residual	St Resid
7	35.5	49.14	36.78	1.27	12.37	2.01R
26	80.0	10.00	16.19	5.17	-6.19	-1.74 X

R denotes an observation with a large standardized residual.  
X denotes an observation whose X value gives it large influence.

La observación 26  
es un punto  
extremo con gran  
influencia

Figura 12.15. Análisis de regresión con dos puntos extremos de X: pendiente negativa (salida Minitab).



Como consecuencia del cambio de dos puntos de datos solamente, ahora la relación tiene una pendiente negativa estadísticamente significativa y las predicciones serían muy diferentes. Sin examinar los diagramas de puntos dispersos, no sabríamos por qué la pendiente que se obtiene es positiva o negativa. Podríamos haber pensado que nuestros resultados representaban una situación de regresión normal como la que hemos visto en el diagrama de puntos dispersos de las ventas al por menor. Obsérvese que en la Figura 12.15 la observación 26 se ha denominado observación extrema mediante el símbolo X.

Este ejemplo muestra un problema que se plantea habitualmente cuando se utilizan datos históricos. Supongamos que  $X$  es el número de trabajadores que trabajan en un turno de producción e  $Y$  es el número de unidades producidas en ese turno. La mayor parte del tiempo la fábrica tiene una plantilla relativamente estable y la producción depende en gran parte de la cantidad de materias primas existentes y de las necesidades de ventas. La producción se ajusta al alza o a la baja en un rango estrecho en respuesta a las demandas y a la plantilla existente,  $X$ . Por lo tanto, vemos que en la mayoría de los casos el diagrama de puntos dispersos cubre un estrecho rango de la variable  $X$ . Pero a veces hay una plantilla muy grande o muy pequeña, o el número de trabajadores se ha registrado incorrectamente. Esos días la producción puede ser excepcionalmente grande o pequeña o puede registrarse incorrectamente. Como consecuencia, tenemos puntos extremos que pueden influir mucho en el modelo de regresión. Estos pocos días determinan los resultados de la regresión. Sin los puntos extremos, la regresión indicaría que la relación es pequeña o nula. Si estos puntos extremos representan extensiones de la relación, el modelo estimado es útil. Pero si estos puntos se deben a condiciones excepcionales o a errores de recogida de datos, el modelo estimado es engañoso.

En una aplicación podemos observar que estos puntos extremos son correctos y deben utilizarse para trazar la recta de regresión. Pero el analista tiene que tomar esa decisión sabiendo que ninguno de los demás puntos de datos apoya la existencia de una relación significativa. De hecho, es necesario realizar un estudio detenido para comprender el sistema y el proceso que generaron los datos y para evaluar los datos de los que se dispone.

### **EJEMPLO 12.5. El efecto de los valores atípicos de la variable $Y$ (análisis mediante un diagrama de puntos dispersos)**

En este ejemplo consideramos el efecto de los valores atípicos en sentido vertical. Recuerdese que el modelo del análisis de regresión supone que toda la variación se produce en el sentido de las  $Y$ . Sabemos, pues, que los valores atípicos en el sentido de las  $Y$  tendrán grandes residuos y éstos residuos darán como resultado una estimación mayor del error del modelo. En este ejemplo, veremos que los efectos pueden ser aún más extremos.

#### **Solución**

Para comenzar, observemos el diagrama de puntos dispersos y el análisis de regresión de las Figuras 12.16 y 12.17. En este ejemplo, tenemos una estrecha relación entre las variables  $X$  e  $Y$ . El diagrama de puntos dispersos apoya claramente la existencia de una relación lineal, estimándose que  $b_1 = 11,88$ . Además, el  $R^2$  del modelo de regresión es cercano a 1 y el estadístico  $t$  de Student es muy alto. Es evidente que tenemos pruebas contundentes para apoyar un modelo lineal.

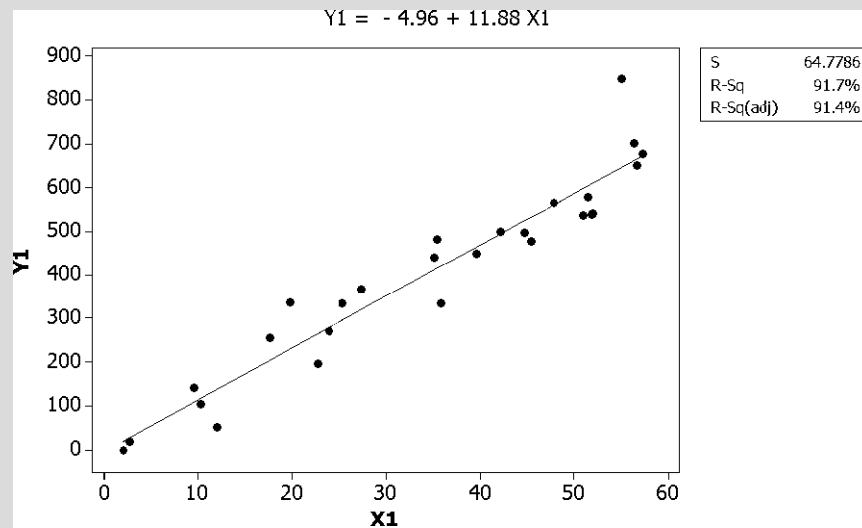


Figura 12.16. Diagrama de puntos dispersos con una pauta prevista.

#### Regression Analysis: Y1 versus X1

The regression equation is  
 $Y1 = -4.96 + 11.88 X1$

s = 64.7786 R-Sq = 91.7% R-Sq(adj) = 91.4%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1160171	1160171	276.48	0.000
Error	25	104907	4196		
Total	26	1265077			

#### Fitted Line: Y1 versus X1

Figura 12.17. Regresión con una pauta prevista (salida Minitab).

Veamos ahora cómo afecta un cambio de dos observaciones a los puntos atípicos, como muestra la Figura 12.18, que podría deberse a un error en la recogida de los datos o a la presencia de unas circunstancias muy poco habituales en el proceso estudiado.

La pendiente de la recta de regresión sigue siendo positiva, pero ahora  $b_1 = 6,40$  y la estimación de la pendiente tiene un error típico mayor, como muestra la Figura 12.19. El intervalo de confianza es mucho más amplio y el valor predicho a partir de la recta de regresión no es tan preciso. Ahora el modelo de regresión correcto no está tan claro. El programa Minitab identifica las observaciones 26 y 27 como observaciones atípicas imprimiendo una R al lado del residuo normalizado. Los residuos normalizados cuyo valor absoluto es superior a 2 se indican en la salida. Si los dos puntos extremos ocurrieron realmente en el funcionamiento normal del proceso, deberíamos incluirlos en



nuestro análisis. Pero el hecho de que se desvíen tanto de la pauta indica que debemos investigar atentamente las situaciones de los datos que generaron esos puntos y estudiar el proceso examinado.

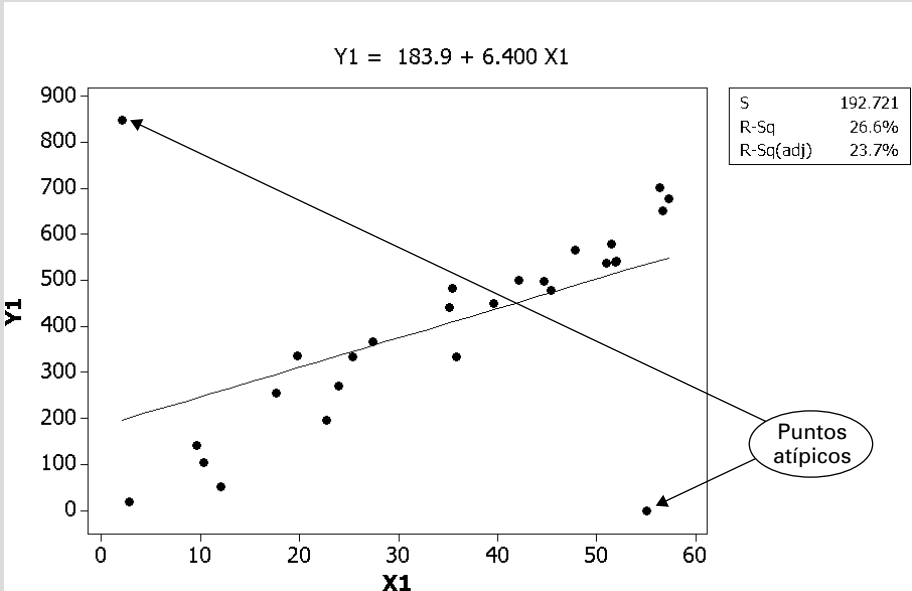


Figura 12.18. Diagrama de puntos dispersos con puntos atípicos de Y.

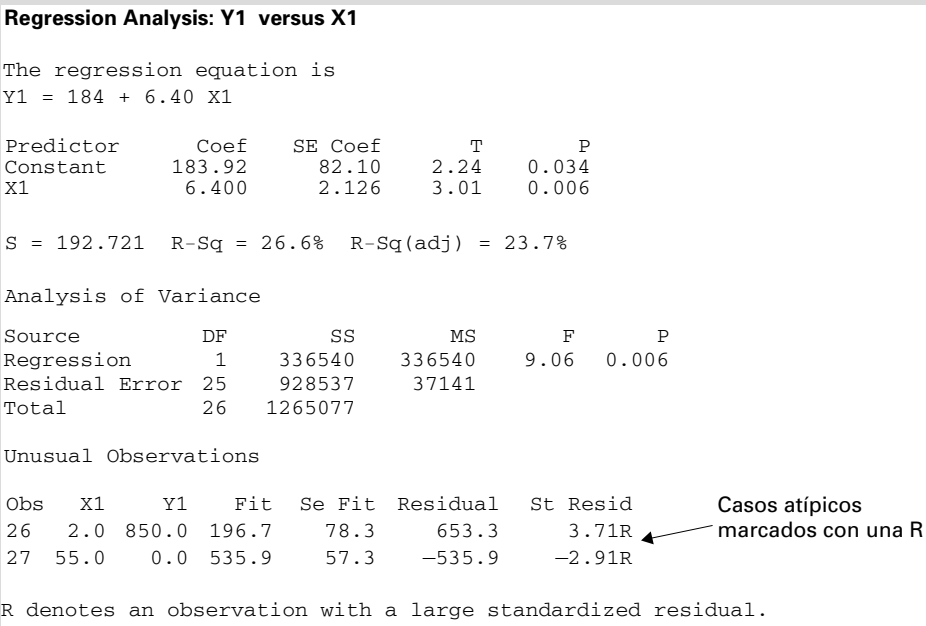


Figura 12.19. Regresión con puntos atípicos de Y (salida Minitab).

Podríamos proponer otros muchos ejemplos. Podríamos observar que el diagrama de puntos dispersos sugiere la existencia de una relación no lineal y, por lo tanto, sería un modelo mejor para un problema específico. En los Capítulos 13 y 14, veremos cómo puede utilizarse la regresión para analizar relaciones no lineales. Observaremos muchas pautas de datos a medida que examinemos distintas aplicaciones del análisis de regresión. Lo importante es que debemos seguir regularmente los métodos del análisis —incluida la realización de diagramas de puntos dispersos— que puedan suministrar la mayor información posible. Como buen analista, debe «¡Conocer sus datos!» En el capítulo siguiente vemos cómo pueden utilizarse también los residuos gráficamente para realizar más contrastes de los modelos de regresión.

## EJERCICIOS

### Ejercicios básicos

- 12.60.** Frank Anscombe, alto ejecutivo encargado de la investigación, le ha pedido que analice los cuatro modelos lineales siguientes utilizando los datos que contiene el fichero de datos **Anscombe**.

$$Y_1 = \beta_0 + \beta_1 X_1$$

$$Y_2 = \beta_0 + \beta_1 X_1$$

$$Y_3 = \beta_0 + \beta_1 X_1$$

$$Y_4 = \beta_0 + \beta_1 X_1$$

Utilice su paquete informático para estimar una regresión lineal para cada modelo. Trace un diagrama de puntos dispersos de los datos utilizados en cada modelo. Escriba un informe, incluyendo los resultados del análisis de regresión y el gráfico, que compare y contraste los cuatro modelos.

### Ejercicio aplicado

- 12.61.** John Foster, presidente de Public Research Inc., le ha pedido ayuda para estudiar el nivel de delincuencia existente en diferentes estados de Estados Unidos antes y después de la realización de elevados gastos federales para reducir la delincuencia. Quiere saber si se puede predecir la tasa de delincuencia en el caso de algunos delitos después de realizados los gastos utilizando la tasa de delincuencia existente antes de realizar los gastos. Le ha pedido que contraste la hipótesis de que la delincuencia existente antes predice la delincuencia posterior en el caso de la tasa total de delincuencia y de las tasas de asesinato, violación y robo. Los datos para su análisis se encuentran en el fichero de datos **Crime Study**. Realice el análisis adecuado y escriba un informe que resuma sus resultados.

## RESUMEN

En este capítulo hemos desarrollado los modelos de dos variables o de mínimos cuadrados simples. Nos hemos basado en algunos de los conceptos descriptivos iniciales presentados en el Capítulo 3. El modelo de regresión simple supone que un conjunto de variables exógenas o independientes tiene una relación lineal con el valor esperado de una variable aleatoria endógena o dependiente. Desarrollando estimaciones de los coeficientes de este modelo, podemos comprender mejor los procesos empresariales y económicos y podemos predecir los valores de la variable endógena en función de la variable exógena. En nuestro estudio, hemos desarrollado estimadores de

los coeficientes y de las variables dependientes. También hemos desarrollado medidas de la bondad del ajuste de la regresión: análisis de la varianza y de  $R^2$ .

Después de ese estudio, hemos presentado métodos de inferencia estadística: contraste de hipótesis e intervalos de confianza de los estimadores de regresión fundamentales. También hemos examinado el análisis de correlación, analizando simplemente la relación entre dos variables. Por último, hemos examinado la importancia de los diagramas de puntos dispersos y el análisis gráfico del desarrollo y el contraste de modelos de regresión.

## TÉRMINOS CLAVE

análisis de la varianza, 450  
base para la inferencia sobre  
la pendiente de la regresión  
poblacional, 459  
coeficiente de determinación,  $R^2$ , 451  
contraste  $F$  para el coeficiente  
de regresión simple, 464  
contrastes de la correlación  
poblacional nula, 433  
contrastes de la pendiente de la  
regresión poblacional, 461

correlación y  $R^2$ , 454  
distribución en el muestreo  
del estimador de los coeficientes  
por mínimos cuadrados, 458  
estimación de la varianza  
del error del modelo, 454  
estimadores de los coeficientes, 442  
intervalos de confianza  
de las predicciones, 467  
intervalos de confianza de la pendiente  
de la regresión poblacional  $b_1$ , 462

método de mínimos  
cuadrados, 442  
regresión lineal basada en un  
modelo poblacional, 440  
resultados de la regresión  
lineal, 441  
supuestos para los estimadores  
de los coeficientes por  
mínimos cuadrados, 442

## EJERCICIOS Y APLICACIONES DEL CAPÍTULO

**12.62.** ¿Qué significa la afirmación de que un par de variables aleatorias están correlacionadas positivamente? Ponga ejemplos de pares de variables aleatorias en los que espera que exista

- a) una correlación positiva
- b) una correlación negativa
- c) una correlación nula

**12.63.** Una muestra aleatoria de cinco conjuntos de observaciones de un par de variables aleatorias dio los resultados de la tabla adjunta.

$X$	4	1	0	1	4
$Y$	-2	-1	0	1	2

- a) Halle el coeficiente de correlación muestral.
- b) Teniendo en cuenta el hecho de que cada valor de  $y_i$  es el cuadrado del valor correspondiente de  $x_i$ , comente su respuesta al apartado (a).

**12.64.** En una muestra aleatoria de 53 tiendas de una cadena de grandes almacenes se observó que la correlación entre las ventas anuales en euros por metro cuadrado de superficie y el alquiler anual en euros por metro cuadrado de superficie era 0,37. Contraste la hipótesis nula de que estas dos cantidades no están correlacionadas en la población frente a la hipótesis alternativa de que la correlación poblacional es positiva.


**12.65.** En una muestra aleatoria de 526 empresas, se observó que la correlación muestral entre la proporción de directivos que son consejeros y una medida del rendimiento de las acciones de la empresa ajustada para tener en cuenta el ries-

go era de 0,1398. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa bilateral.

**12.66.** En una muestra de 66 meses se observó que la correlación entre los rendimientos de los bonos a 10 años de Canadá y de Hong Kong era de 0,293. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa de que es positiva.

**12.67.** En una muestra aleatoria de 192 mujeres trabajadoras, se observó una correlación muestral de -0,18 entre la edad y una medida de la disposición a cambiar de empleo. Basándose únicamente en esta información, extraiga todas las conclusiones que pueda sobre la regresión de la disposición a cambiar de empleo con respecto a la edad.

**12.68.** Basándose en una muestra de  $n$  observaciones,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , se calcula la regresión muestral de  $y$  con respecto a  $x$ . Demuestre que la recta de regresión muestral pasa por el punto  $(x = \bar{x}, y = \bar{y})$ , donde  $\bar{x}$  e  $\bar{y}$  son las medias muestrales.

**12.69.**  Una empresa realiza normalmente un test de aptitud a todo el nuevo personal en formación. Al final del primer año en la empresa, este personal en formación es valorado por sus supervisores inmediatos. En una muestra aleatoria de 12 personas en formación, se obtuvieron los resultados mostrados en el fichero de datos **Employee Test**.

- a) Estime la regresión de la valoración realizada por el supervisor con respecto a la puntuación obtenida en el test de aptitud.

- b) Interprete la pendiente de la recta de regresión muestral.
- c) ¿Es posible dar una interpretación útil a la ordenada en el origen de la recta de regresión muestral?
- d) Halle e interprete el coeficiente de determinación de esta regresión.
- e) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa unilateral obvia.
- f) Halle el intervalo de confianza al 95 por ciento de la valoración que daría el supervisor a una persona en formación que tuviera una puntuación de 70 en el test de aptitud.

**12.70.** Se ha intentado evaluar la tasa de inflación como predictor del tipo al contado en el mercado de letras del Tesoro alemanas. Partiendo de una muestra de 79 observaciones trimestrales, se obtuvo la regresión lineal estimada

$$\hat{y} = 0,0027 + 0,7916x$$

donde

$y$  = variación efectiva del tipo al contado

$x$  = variación del tipo al contado predicha por la tasa de inflación

El coeficiente de determinación era 0,097 y la desviación típica estimada del estimador de la pendiente de la recta de regresión poblacional era 0,2759.

- a) Interprete la pendiente de la recta de regresión estimada.
- b) Interprete el coeficiente de determinación.
- c) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que la verdadera pendiente es positiva e interprete su resultado.
- d) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 1 frente a la hipótesis alternativa bilateral.

**12.71.** La tabla muestra las compras por comprador de ocho cosechas de un vino selecto ( $y$ ) y la valoración del vino realizada por el comprador en un año ( $x$ ).

$x$	3,6	3,3	2,8	2,6	2,7	2,9	2,0	2,6
$y$	24	21	22	22	18	13	9	6

- a) Estime la regresión de las compras por comprador con respecto a la valoración realizada por el comprador.

- b) Interprete la pendiente de la recta de regresión estimada.
- c) Halle e interprete el coeficiente de determinación.
- d) Halle e interprete el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.
- e) Halle el intervalo de confianza al 90 por ciento de las compras esperadas por comprador de una cosecha a la que el comprador da una valoración de 2,0.

**12.72.** En una muestra de 306 estudiantes de un curso básico de estadística, se obtuvo la recta de regresión muestral

$$y = 58,813 + 0,2875x$$

donde

$y$  = calificación final de los estudiantes al terminar el curso

$x$  = calificación en un examen de posición realizado al principio de curso.

El coeficiente de determinación era 0,1158 y la desviación típica estimada del estimador de la pendiente de la recta de regresión poblacional era 0,04566.

- a) Interprete la pendiente de la recta de regresión muestral.
- b) Interprete el coeficiente de determinación.
- c) La información dada permite contrastar la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 de dos formas distintas frente a la hipótesis alternativa de que es positiva. Realice estos contrastes y muestre que llegan a la misma conclusión.

**12.73.** Basándose en una muestra de 30 observaciones, se estimó el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Las estimaciones por mínimos cuadrados obtenidas fueron

$$b_0 = 10,1 \quad y \quad b_1 = 8,4$$

La suma de los cuadrados de la regresión y la suma de los cuadrados de los errores fueron

$$SCR = 128 \quad y \quad SCE = 286$$

- a) Halle e interprete el coeficiente de determinación.
- b) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que  $\beta_1$  es 0 frente a la hipótesis alternativa bilateral.

c) Halle

$$\sum_{i=1}^{30} (x_i - \bar{x})^2$$

- 12.74.** Basándose en una muestra de 25 observaciones, se estimó el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Las estimaciones por mínimos cuadrados obtenidas fueron

$$b_0 = 15,6 \quad y \quad b_1 = 1,3$$

La suma total de los cuadrados y la suma de los cuadrados de los errores fueron

$$STC = 268 \quad y \quad SCE = 204$$

- a) Halle e interprete el coeficiente de determinación.  
 b) Contraste al nivel de significación del 5 por ciento la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.  
 c) Halle el intervalo de confianza al 95 por ciento de  $\beta_1$ .
- 12.75.** Un analista cree que el único determinante importante de los rendimientos de los activos ( $Y$ ) del banco es el cociente entre los préstamos y los depósitos ( $x$ ). En una muestra aleatoria de 20 bancos se obtuvo la recta de regresión muestral

$$Y = 0,97 + 0,47x$$

con el coeficiente de determinación de 0,720.

- a) Halle la correlación muestral entre los rendimientos de los activos y el cociente entre los préstamos y los depósitos.  
 b) Contraste la hipótesis nula de que no existe una relación lineal entre los rendimientos y el cociente frente a una hipótesis alternativa bilateral.  
 c) Halle

$$\frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- 12.76.** Comente la siguiente afirmación:

Si se estima una regresión del rendimiento por acre del maíz con respecto a la cantidad de fertilizante utilizada empleando las cantidades de fertilizante utilizadas normalmente por los agricultores, la pendiente de la recta de regresión estimada será, desde luego, positiva. Sin embargo, es bien sabido que si se utiliza una cantidad muy grande de fertilizante, el rendimiento del maíz es muy bajo. Por lo tanto, las ecuaciones de regresión no son muy útiles para hacer predicciones.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

- 12.77.** El departamento de economía de una universidad está intentando averiguar si los conocimientos verbales o matemáticos son más importantes para predecir el éxito académico en los estudios de economía. El profesorado del departamento ha decidido utilizar como medida del éxito la calificación media (GPA) obtenida por los licenciados en los cursos de economía. Los conocimientos verbales se miden por medio de las calificaciones obtenidas en dos exámenes estandarizados: el SAT verbal y el ACT de inglés. Los conocimientos matemáticos se miden por medio de las calificaciones obtenidas en el SAT de matemáticas y en el ACT de matemáticas. El fichero de datos llamado **Student GPA**, que se encuentra en su disco de datos, contiene los datos de 112 estudiantes. El nombre de las columnas de las variables se indica al comienzo del fichero de datos. Debe utilizar el paquete estadístico que utilice habitualmente para realizar el análisis de este problema.

- a) Represente gráficamente la GPA de economía en relación con cada una de las dos calificaciones de los conocimientos verbales y cada una de las dos calificaciones de los conocimientos matemáticos. ¿Qué variable es el mejor predictor? Observe las pautas poco habituales que haya en los datos.  
 b) Calcule los coeficientes del modelo lineal y los estadísticos del análisis de regresión para los modelos que predicen la GPA de economía en función de cada calificación en conocimientos verbales y cada calificación en conocimientos matemáticos. Utilizando tanto las medidas matemáticas y verbales del SAT como las medidas de matemáticas e inglés del ACT, averigüe si los conocimientos matemáticos o verbales son el mejor predictor de la GPA de economía.  
 c) Compare los estadísticos descriptivos —la media, la desviación típica, el cuartil superior y el inferior, el rango— de las variables consideradas predictoras. Observe las diferencias e indique cómo afectan estas diferencias a la capacidad del modelo lineal para realizar predicciones.

- 12.78.** Los responsables de la National Highway Traffic Safety Administration (NHTSA) de Estados Unidos quieren saber si los diferentes tipos de vehículos de un estado tienen relación con la tasa de mortalidad en carretera del esta-

do. Le han pedido que realice varios análisis de regresión para averiguar si el peso medio de los vehículos, el porcentaje de automóviles importados, el porcentaje de camiones ligeros o la antigüedad media de los automóviles están relacionados con las muertes en accidente ocurridas en automóviles y camionetas. Los datos del análisis se encuentran en el fichero de datos llamado **Crash**, que está en su disco de datos. Las descripciones y las localizaciones de las variables se encuentran en el catálogo del fichero de datos del apéndice.

- a) Represente gráficamente las muertes en accidente en relación con cada una de las variables potenciales de predicción. Observe la relación y cualquier pauta excepcional en los puntos de datos.
- b) Realice un análisis de regresión simple de las muertes totales en accidente con respecto a las variables potenciales de predicción. Indique si alguna de las regresiones muestra una relación significativa y, en caso afirmativo, cuál.
- c) Muestre los resultados de su análisis y ordene las variables de predicción según su relación con las muertes totales en accidente.

**12.79.** El Departamento de Transporte de Estados Unidos desea saber si los estados que tienen un porcentaje mayor de población urbana tienen una tasa más alta de muertes totales en accidente ocurridas en automóviles y camionetas. También quiere saber si existe alguna relación entre la velocidad media a la que se conduce por las carreteras rurales o el porcentaje de carreteras rurales que están asfaltadas y las tasas de muertes en accidente. Los datos de este estudio se encuentran en el fichero de datos **Crash** almacenado en su disco de datos.

- a) Represente gráficamente las muertes en accidente en relación con cada una de las variables potenciales de predicción. Observe la relación y cualquier pauta excepcional en los puntos de datos.
- b) Realice un análisis de regresión simple de las muertes en accidente con respecto a las variables potenciales de predicción.
- c) Muestre los resultados de su análisis y ordene las variables de predicción según su relación con las muertes totales en accidente.

**12.80.** Un economista desea predecir el valor de mercado de las viviendas de pequeñas ciudades del Medio Oeste ocupadas por sus propietarios. Ha reunido un conjunto de datos de 45 peque-

ñas ciudades que se refieren a un periodo de dos años y quiere que los utilice como fuente de datos para el análisis. Los datos se encuentran en el fichero **Citydat**, que están en su disco de datos. Quiere que desarrolle dos ecuaciones de predicción: una que utilice el tamaño de la vivienda como predictor y otra que utilice el tipo impositivo como predictor.

- a) Represente gráficamente el valor de mercado de las viviendas (hseval) en relación con el tamaño de la vivienda (sizense) y en relación con el tipo impositivo (taxrate). Observe cualquier pauta excepcional en los datos.
- b) Realice análisis de regresión para las dos variables de predicción. ¿Qué variable predice mejor el valor de las viviendas?
- c) Un promotor industrial de un estado del Medio Oeste ha afirmado que los tipos del impuesto local sobre bienes inmuebles de las pequeñas ciudades debe bajarse porque, en caso contrario, nadie comprará una vivienda en estas ciudades. Basándose en su análisis de este problema, evalúe la afirmación del promotor.

**12.81.** Stuart Wainwright, vicepresidente de compras para una gran cadena nacional de tiendas de Estados Unidos, le ha pedido que realice un análisis de las ventas al por menor por estados. Quiere saber si el porcentaje de desempleados o la renta personal per cápita están relacionados con las ventas al por menor per cápita. Los datos para realizar este estudio se encuentran en el fichero de datos llamado **Retail**, que está almacenado en su disco de datos y se describe en el catálogo del fichero de datos del apéndice.

- a) Trace gráficos y realice análisis de regresión para averiguar las relaciones entre las ventas al por menor per cápita y el porcentaje de desempleados y la renta personal. Calcule intervalos de confianza al 95 por ciento para los coeficientes de la pendiente de cada ecuación de regresión.
- b) ¿Cómo afecta una disminución de la renta per cápita de 1.000 \$ a las ventas per cápita?
- c) ¿Cuál es el intervalo de confianza al 95 por ciento en la ecuación de la renta per cápita de las ventas al por menor correspondientes a la renta media per cápita y a un nivel que esté 1.000 \$ por encima de la renta media per cápita?

**12.82.** Un importante proveedor nacional de materiales de construcción para la construcción de viviendas está preocupado por las ventas totales

del próximo año. Es bien sabido que las ventas de la empresa están relacionadas directamente con la inversión nacional total en vivienda. Algunos banqueros de Nueva York están prediciendo que los tipos de interés subirán alrededor de 2 puntos porcentuales el próximo año. Le han pedido que realice un análisis de regresión para poder predecir el efecto de las variaciones de los tipos de interés en la inversión en vivienda. Los datos de series temporales para realizar este estudio se encuentran en el fichero de datos llamado **Macro2003**, que está almacenado en su disco de datos y se describe en el apéndice del Capítulo 14.

- a) Desarrolle dos modelos de regresión para predecir la inversión en vivienda utilizando el tipo de interés preferencial para uno y el

tipo de interés de los fondos federales para el otro. Analice los estadísticos de la regresión e indique qué ecuación hace las mejores predicciones.

- b) Halle el intervalo de confianza al 95 por ciento del coeficiente de la pendiente en ambas ecuaciones de regresión.
- c) Basándose en cada modelo, prediga cómo afecta una subida de los tipos de interés de 2 puntos porcentuales a la inversión en vivienda.
- d) Utilizando ambos modelos, calcule intervalos de confianza al 95 por ciento de la variación de la inversión en vivienda provocada por una subida de los tipos de interés de 2 puntos porcentuales.

## Apéndice

En este apéndice mostramos cómo se estiman por mínimos cuadrados los parámetros poblacionales de regresión. Queremos hallar los valores  $b_0$  y  $b_1$  tales que la suma de los cuadrados de las discrepancias

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

sea lo más pequeña posible.

En primer lugar, mantenemos constante  $b_1$  y diferenciamos con respecto a  $b_0$ , lo que nos da

$$\begin{aligned} \frac{\partial SCE}{\partial b_0} &= 2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \\ &= -2 \left( \sum y_i - nb_0 - b_1 \sum x_i \right) \end{aligned}$$

Dado que esta derivada debe ser 0 para obtener un mínimo, tenemos que

$$\sum y_i - nb_0 - b_1 \sum x_i = 0$$

Por lo tanto, dividiendo por  $n$  resulta que

$$b_0 = \bar{y} - b_1 \bar{x}$$

Introduciendo este resultado de  $b_0$  en la expresión anterior, tenemos que

$$SCE = \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2$$

Diferenciando esta expresión con respecto a  $b_1$ , obtenemos

$$\begin{aligned}\frac{\partial SCE}{\partial b_1} &= 2 \sum_{i=1}^n (x_i - \bar{x})[(y_i - \bar{y}) - b_1(x_i - \bar{x})] \\ &= -2 \left( \sum (x_i - \bar{x})(y_i - \bar{y}) - b_1 \sum (x_i - \bar{x})^2 \right)\end{aligned}$$

Esta derivada debe ser 0 para obtener un mínimo, por lo que tenemos que

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = b_1 \sum (x_i - \bar{x})^2$$

Por lo tanto,

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

## Bibliografía

---

1. Dhalla, N. K., «Short-Term Forecasts of Advertising Expenditures», *Journal of Advertising Research*, 19, n.º 1, 1979, págs. 7-14.
2. Mampower, J. L., S. Livingston y T. J. Lee, «Expert Judgments of Political Risk», *Journal of Forecasting*, 6, 1987, págs. 51-65.





## Regresión múltiple

### Esquema del capítulo

- 13.1. El modelo de regresión múltiple
  - Especificación del modelo
  - Desarrollo del modelo
  - Gráficos tridimensionales
- 13.2. Estimación de coeficientes
  - Método de mínimos cuadrados
- 13.3. Poder explicativo de una ecuación de regresión múltiple
- 13.4. Intervalos de confianza y contrastes de hipótesis de coeficientes de regresión
  - individuales
  - Intervalos de confianza
  - Contrastes de hipótesis
- 13.5. Contrastes de los coeficientes de regresión
  - Contrastes de todos los coeficientes
  - Contraste de un conjunto de coeficientes de regresión
  - Comparación de los contrastes  $F$  y  $t$
- 13.6. Predicción
- 13.7. Transformaciones de modelos de regresión no lineales
  - Transformaciones de modelos cuadráticos
  - Transformaciones logarítmicas
- 13.8. Utilización de variables ficticias en modelos de regresión
  - Diferencias entre las pendientes
- 13.9. Método de aplicación del análisis de regresión múltiple
  - Especificación del modelo
  - Regresión múltiple
  - Efecto de la eliminación de una variable estadísticamente significativa
  - Análisis de los residuos

### Introducción

En el Capítulo 12 presentamos el método de regresión simple para obtener una ecuación lineal que predice una variable dependiente o endógena en función de una única variable independiente o exógena; por ejemplo, el número total de artículos vendidos en función del precio. Sin embargo, en muchas situaciones, varias variables independientes influyen conjuntamente en una variable dependiente. La regresión múltiple nos permite averiguar el efecto simultáneo de varias variables independientes en una variable dependiente utilizando el principio de los mínimos cuadrados.

Existen muchas aplicaciones importantes de la regresión múltiple en el mundo de la empresa y en la economía. Entre estas aplicaciones se encuentran las siguientes:

1. La cantidad vendida de bienes es una función del precio, la renta, la publicidad, el precio de los bienes sustitutivos y otras variables.
2. Existe inversión de capital cuando un empresario cree que puede obtener un beneficio. Por lo tanto, la inversión de capital es una función de variables relacionadas con las posibilidades de obtener beneficios, entre las que se encuentran el tipo de interés, el producto interior bruto, las expectativas de los consumidores, la renta disponible y el nivel tecnológico.
3. El salario es una función de la experiencia, la educación, la edad y el puesto de trabajo.
4. Las grandes empresas del comercio al por menor y la hostelería deciden la localización de los nuevos establecimientos basándose en los ingresos previstos por ventas y/o en la rentabilidad. Utilizando datos de localizaciones anteriores que han tenido éxito y que no lo han tenido, los analistas pueden construir modelos que predicen las ventas o los beneficios de una nueva localización posible.

El análisis económico y empresarial tiene algunas características únicas en comparación con el análisis de otras disciplinas. Los científicos naturales trabajan en un laboratorio en el que es posible controlar muchas variables, pero no todas. En cambio, el laboratorio del economista y del directivo es el mundo y las condiciones no pueden controlarse. Por lo tanto, necesitan instrumentos como la regresión múltiple para estimar el efecto simultáneo de varias variables. La regresión múltiple como «instrumento de laboratorio» es muy importante para el trabajo de los directivos y de los economistas. En este capítulo veremos muchas aplicaciones específicas en los ejemplos y los ejercicios.

Los métodos para ajustar modelos de regresión múltiple se basan en el mismo principio de los mínimos cuadrados que aprendimos en el Capítulo 12 y, por lo tanto, las ideas presentadas en ese capítulo se extenderán directamente a la regresión múltiple. Sin embargo, se introducen algunas complejidades debido a las relaciones entre las distintas variables exógenas. Éstas requieren nuevas ideas que se desarrollan en este capítulo.

## 13.1. El modelo de regresión múltiple

---

Nuestro objetivo es aprender a utilizar la regresión múltiple para crear y analizar modelos. Por lo tanto, aprendemos cómo funciona la regresión múltiple y algunas directrices para interpretarla. Comprendiendo perfectamente la regresión múltiple, es posible resolver una amplia variedad de problemas aplicados. Este estudio de los métodos de regresión múltiple es paralelo al de la regresión simple. El primer paso para desarrollar un modelo es la especificación de ese modelo, que consiste en la selección de las variables del modelo y de la forma del modelo. A continuación, se estudia el método de mínimos cuadrados y se analiza la variabilidad para identificar los efectos de cada una de las variables de predicción. Después se estudia la estimación, los intervalos de confianza y el contraste de hipótesis. Se utilizan frecuentemente aplicaciones informáticas para indicar cómo se aplica la teoría a problemas realistas. El estudio de este capítulo será más fácil si se ponen en relación sus ideas con las que presentamos en el Capítulo 12.

### Especificación del modelo

Comenzamos con una aplicación que ilustra la importante tarea de la especificación del modelo de regresión. La especificación del modelo consiste en la selección de las variables exógenas y la forma funcional del modelo.

**EJEMPLO 13.1. Proceso de producción (especificación del modelo de regresión)**

El director de producción de Circuitos Flexibles, S.A., le ha pedido ayuda para estudiar un proceso de producción. Los circuitos flexibles se producen con un rollo continuo de resina flexible que lleva adherida a su superficie una fina película de material conductor hecho de cobre. El cobre se adhiere a la resina pasando la resina por una solución de cobre. El grosor del cobre es fundamental para que los circuitos sean de buena calidad. Depende en parte de la temperatura de la solución de cobre, de la velocidad de la línea de producción, de la densidad de la solución y del grosor de la resina flexible. Para controlar el grosor del cobre adherido a la superficie, el director de producción necesita saber qué efecto produce cada una de estas variables. Le ha pedido ayuda para desarrollar un modelo de regresión múltiple.

**Solución**

La regresión múltiple puede utilizarse para hacer estimaciones del efecto que produce cada variable en combinación con las demás. El desarrollo del modelo comienza con un análisis detenido del contexto del problema. El primer paso en este ejemplo sería una extensa conversación con los ingenieros responsables del diseño del producto y de la producción, con el fin de comprender detalladamente el proceso del que se pretende desarrollar un modelo. En algunos casos, se estudiaría la literatura existente sobre el proceso. Éste debe ser comprendido y aceptado por todos los interesados antes de poder desarrollar un modelo útil utilizando el análisis de regresión múltiple. En este ejemplo, la variable dependiente,  $Y$ , es el grosor del cobre. Las variables independientes son la temperatura de la solución de cobre,  $X_1$ ; la velocidad de la línea de producción,  $X_2$ ; la densidad de la solución,  $X_3$ , y el grosor de la resina flexible,  $X_4$ . Los ingenieros y los científicos que comprendían la tecnología del proceso de recubrimiento identificaron estas variables como posibles predictores del grosor del cobre,  $Y$ . Basándose en el estudio del proceso, la especificación del modelo resultante es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

En el modelo lineal anterior, las  $\beta_j$  son coeficientes lineales constantes de las  $X_j$  que indican el efecto condicionado de cada variable independiente en la determinación de la variable dependiente,  $Y$ , en la población. Por lo tanto, las  $\beta_j$  son parámetros en el modelo de regresión lineal. A continuación, se produciría una serie de lotes para hacer mediciones de distintas combinaciones de las variables independientes y la variable dependiente (véase el análisis del diseño experimental en el apartado 14.2).

**EJEMPLO 13.2. Localización de las tiendas (especificación del modelo)**

El director de planificación de una gran cadena de comercio al por menor estaba insatisfecho con su experiencia en la apertura de nuevas tiendas. En los cuatro últimos años, el 25 por ciento de las nuevas tiendas no había conseguido las ventas previstas en el periodo de prueba de dos años y se había cerrado con cuantiosas pérdidas económicas. El director quería desarrollar mejores criterios para elegir el emplazamiento de las tiendas y llegó a la conclusión de que debía estudiarse la experiencia histórica de las tiendas que habían tenido éxito y las que habían fracasado.

**Solución**

Hablando con un consultor, llegó a la conclusión de que podían utilizarse los datos de las tiendas que habían conseguido las ventas que estaban previstas y los datos de las que no las habían conseguido para desarrollar un modelo de regresión múltiple. El consultor sugirió que debía utilizarse como variable dependiente,  $Y$ , las ventas del segundo año. Se emplearía un modelo de regresión para predecir las ventas del segundo año en función de varias variables independientes que definen la zona que rodea a la tienda. Sólo se abrirían tiendas en los lugares en los que las ventas predichas superaran un nivel mínimo. El modelo también indicaría cómo afectan varias variables independientes a las ventas.

Tras hablar largo y tendido con personas de la empresa, el consultor recomendó las siguientes variables independientes:

1.  $X_1$  = tamaño de la tienda
2.  $X_2$  = volumen de tráfico de la calle en la que se encuentra la tienda
3.  $X_3$  = apertura de la tienda sola o en un centro comercial
4.  $X_4$  = existencia de una tienda rival a menos de 500 metros
5.  $X_5$  = renta per cápita de la población residente a menos de 8 kilómetros
6.  $X_6$  = número total de personas que residen a menos de 8 kilómetros
7.  $X_7$  = renta per cápita de la población que reside a menos de 15 kilómetros
8.  $X_8$  = número total de personas que residen a menos de 15 kilómetros

Se utilizó la regresión múltiple para estimar los coeficientes del modelo de predicción de las ventas a partir de datos recogidos en todas las tiendas abiertas en los ocho últimos años. En el conjunto de datos había tiendas que seguían abiertas y tiendas que se habían cerrado. Se desarrolló un modelo que podía utilizarse para predecir las ventas del segundo año. Este modelo contenía estimadores,  $b'_j$ , de los parámetros del modelo,  $\beta'_j$ . Para aplicar el modelo

$$\hat{y}_i = b_0 + \sum_{j=1}^8 b_j x_{ji}$$

se hicieron mediciones de las variables independientes de cada nueva localización propuesta y se calcularon las ventas predichas de cada localización. Se utilizó el nivel predicho de ventas, junto con el criterio de los analistas de marketing y de un comité de directores de tiendas de éxito, para elegir el lugar en el que se abrirían tiendas.

En la estrategia para especificar un modelo influyen los objetivos del modelo. Uno de los objetivos es la predicción de una variable dependiente o «de resultado». Entre las aplicaciones se encuentran la predicción de las ventas, de la producción, del consumo total, de la inversión total y otros muchos criterios de los resultados empresariales y económicos. El segundo objetivo es estimar el efecto marginal de cada variable independiente. Los economistas y los directivos necesitan saber cómo cambian las medidas de los resultados cuando varían las variables independientes,  $X_j$ , donde  $j = 1, \dots, K$ . Por ejemplo:

1. ¿Cómo varían las ventas como consecuencia de una subida del precio y de los gastos publicitarios?
2. ¿Cómo varía la producción cuando se alteran las cantidades de trabajo y de capital?
3. ¿Disminuye la mortalidad infantil cuando se incrementan los gastos en asistencia sanitaria y en servicios de saneamiento?

## Objetivos de la regresión

La regresión múltiple permite obtener dos importantes resultados:

1. Una ecuación lineal estimada que predice la variable dependiente,  $Y$ , en función de  $K$  variables independientes observadas,  $x_j$ , donde  $j = 1, \dots, K$ .

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_Kx_{Ki}$$

donde  $i = 1, \dots, n$  observaciones.

2. La variación marginal de la variable dependiente,  $Y$ , provocada por las variaciones de las variables independientes, que se estima por medio de los coeficientes,  $b'_j$ . En la regresión múltiple, estos coeficientes dependen de qué otras variables se incluyan en el modelo. El coeficiente  $b'_j$  indica la variación de  $Y$ , dada una variación unitaria de  $x_j$ , descontando al mismo tiempo el efecto simultáneo de las demás variables independientes.

En algunos problemas, ambos resultados son igual de importantes. Sin embargo, normalmente predomina uno de ellos (por ejemplo, la predicción de las ventas de las tiendas,  $Y$ , en el ejemplo de la localización de las tiendas).

La variación marginal es más difícil de estimar porque las variables independientes están relacionadas no sólo con las variables dependientes sino también entre sí. Si dos variables independientes o más varían en una relación lineal directa entre sí, es difícil averiguar el efecto que produce cada variable independiente en la variable dependiente.

Examinaremos detalladamente el modelo del ejemplo 13.2. El coeficiente de  $x_1$  —es decir,  $b_1$ — indica la variación que experimentan las ventas del segundo año por cada variación unitaria del tamaño de la tienda. El coeficiente de  $x_5$  indica la variación que experimentan las ventas por cada variación unitaria de la renta per cápita de la población que reside a menos de 8 kilómetros, mientras que la de  $x_7$  indica la variación de las ventas por cada variación de la renta per cápita de la población que reside a menos de 15 kilómetros. Es probable, por supuesto, que las variables  $x_5$  y  $x_7$  estén correlacionadas. Por lo tanto, en la medida en que estas variables varíen ambas al mismo tiempo, es difícil averiguar la contribución de cada una de ellas a la variación de los ingresos generados por las ventas de las tiendas. Esta correlación entre variables independientes complica el modelo. Es importante comprender que el modelo predice los ingresos generados por las ventas de las tiendas utilizando la combinación de variables que contiene el modelo. El efecto de una variable de predicción es el efecto que produce esa variable cuando se combina con las demás. Por lo tanto, en general, el coeficiente de una variable no indica el efecto que produce esa variable en todas las condiciones. Estas complejidades se analizarán más detenidamente cuando se desarrolle el modelo de regresión múltiple.

## Desarrollo del modelo

Cuando aplicamos la regresión múltiple, construimos un modelo para explicar la variabilidad de la variable dependiente. Para eso queremos incluir las influencias simultáneas e individuales de varias variables independientes. Supongamos, por ejemplo, que queremos desarrollar un modelo que prediga el margen anual de beneficios de las sociedades de ahorro y crédito inmobiliario utilizando los datos recogidos durante un periodo de años. Una especificación inicial del modelo indicaba que el margen anual de beneficios estaba relacionado con los ingresos netos por dólar depositado y el número de oficinas. Se espera que el ingreso neto aumente el margen anual de beneficios y se prevé que el número de oficinas

reducirá el margen anual de beneficios debido al aumento de la competencia. Eso nos llevaría a especificar un modelo de regresión poblacional

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

donde

$Y$  = margen anual de beneficios

$X_1$  = ingresos anuales netos por dólar depositado

$X_2$  = número de oficinas existentes ese año



### Savings and Loan

La Tabla 13.1 y el fichero de datos **Savings and Loan** contienen 25 observaciones por año de estas variables. Utilizaremos estos datos para desarrollar un modelo lineal que prediga el margen anual de beneficios en función de los ingresos por dólar depositado y del número de oficinas (véase la referencia bibliográfica 4).

**Tabla 13.1.** Datos de las asociaciones de ahorro y crédito inmobiliario.

Año	Ingresos por dólar	Número de oficinas	Margen de beneficios	Año	Ingresos por dólar	Número de oficinas	Margen de beneficios
1	3,92	7.298	0,75	14	3,78	6.672	0,84
2	3,61	6.855	0,71	15	3,82	6.890	0,79
3	3,32	6.636	0,66	16	3,97	7.115	0,7
4	3,07	6.506	0,61	17	4,07	7.327	0,68
5	3,06	6.450	0,7	18	4,25	7.546	0,72
6	3,11	6.402	0,72	19	4,41	7.931	0,55
7	3,21	6.368	0,77	20	4,49	8.097	0,63
8	3,26	6.340	0,74	21	4,70	8.468	0,56
9	3,42	6.349	0,9	22	4,58	8.717	0,41
10	3,42	6.352	0,82	23	4,69	8.991	0,51
11	3,45	6.361	0,75	24	4,71	9.179	0,47
12	3,58	6.369	0,77	25	4,78	9.318	0,32
13	3,66	6.546	0,78				

Pero antes de poder estimar el modelo, es necesario desarrollar y comprender el método de regresión múltiple. Para comenzar, examinemos el modelo general de regresión múltiple y observemos sus diferencias con el modelo de regresión simple. El modelo de regresión múltiple es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

donde  $\varepsilon_i$  es el término de error aleatorio que tiene la media 0 y la varianza  $\sigma^2$ , y las  $\beta'_j$  son los coeficientes o efectos marginales de las variables independientes o exógenas,  $x_j$ , donde  $j = 1, \dots, K$ , dados los efectos de las demás variables independientes. Las  $i$  indican las observaciones, siendo  $i = 1, \dots, n$ . Utilizamos las minúsculas  $x_{ji}$  para indicar los valores específicos de la variable  $X_j$  en la observación  $i$ . Suponemos que las  $\varepsilon_i$  son independientes de las  $X_j$  y entre sí para que las estimaciones de los coeficientes y sus varianzas sean correctas. En el Capítulo 14 explicamos qué ocurre cuando se abandonan estos supuestos.

El modelo muestral estimado es

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_Kx_{Ki} + e_i$$

donde  $e_i$  es el residuo o diferencia entre el valor observado de  $Y$  y el valor estimado de  $Y$  obtenido utilizando los coeficientes estimados,  $b_j$ , donde  $j = 1, \dots, K$ . El método de regresión obtiene estimaciones simultáneas,  $b_j$ , de los coeficientes del modelo poblacional,  $\beta_j$ , utilizando el método de mínimos cuadrados.

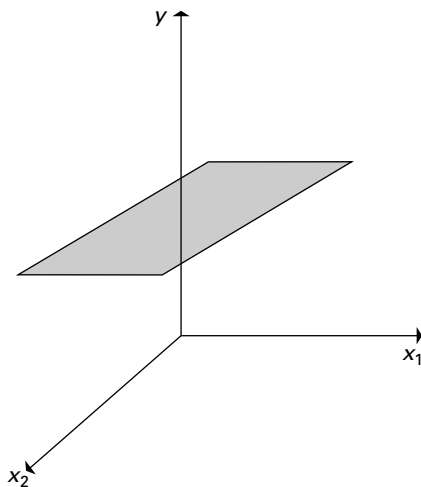
En nuestro ejemplo de las asociaciones de ahorro y crédito inmobiliario, el modelo poblacional para los puntos de datos individuales es

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \varepsilon_i$$

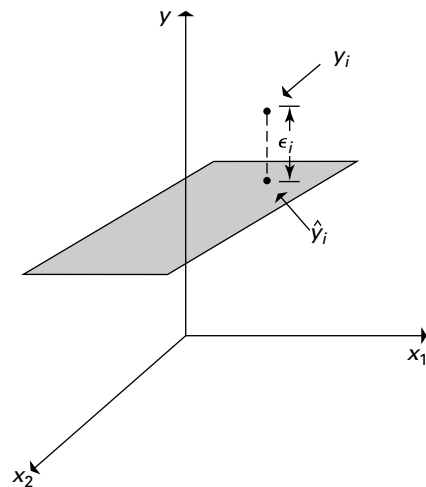
Este modelo reducido con dos variables de predicción solamente brinda la oportunidad de comprender mejor el método de regresión. La función de regresión puede representarse gráficamente en tres dimensiones, como muestra la Figura 13.1. La función de regresión se representa mediante un plano en el que los valores de  $Y$  son una función de los valores de las variables independientes  $X_1$  y  $X_2$ . Para cada par posible,  $x_{1i}$ ,  $x_{2i}$ , el valor esperado de la variable dependiente,  $y_i$ , se encuentra en el plano. La Figura 13.2 ilustra específicamente el ejemplo de las asociaciones de ahorro y crédito inmobiliario. Un aumento de  $X_1$  provoca un aumento del valor esperado de  $Y$ , condicionado al efecto de  $X_2$ . Asimismo, un aumento de  $X_2$  provoca una disminución del valor esperado de  $Y$ , condicionada al efecto de  $X_1$ .

Para completar nuestro modelo, añadimos un término de error  $\varepsilon$ . Este término de error reconoce que no se cumplirá exactamente ninguna relación postulada y que es probable que haya otras variables que también afecten al valor observado de  $Y$ . Por lo tanto, cuando aplicamos el modelo, observamos el valor esperado de la variable dependiente,  $Y$  —representado por el plano en la Figura 13.2—, más un término de error aleatorio,  $\varepsilon$ , que representa la parte de  $Y$  no incluida en el valor esperado. Como consecuencia, el modelo de datos tiene la forma

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \dots + \beta_Kx_{Ki} + \varepsilon_i$$



**Figura 13.1.** El plano es el valor esperado de  $Y$  en función de  $X_1$  y  $X_2$ .



**Figura 13.2.** Comparación del valor observado y el valor esperado de  $Y$  en función de dos variables independientes.



### El modelo de regresión poblacional múltiple

El **modelo de regresión poblacional múltiple** define la relación entre una variable dependiente o endógena,  $Y$ , y un conjunto de variables independientes o exógenas,  $x_j$ , donde  $j = 1, \dots, K$ . Se supone que las  $x_{ji}$  son números fijos;  $Y$  es una variable aleatoria definida para cada observación,  $i$ , donde  $i = 1, \dots, n$ , y  $n$  es el número de observaciones. El modelo se define de la forma siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \varepsilon_i \quad (13.1)$$

donde las  $\beta_j$  son coeficientes constantes y las  $\varepsilon$  son variables aleatorias de 0 y varianza  $\sigma^2$ .

En el ejemplo de las asociaciones de ahorro y crédito inmobiliario, con dos variables independientes, el modelo de regresión poblacional es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Dados valores específicos de los ingresos netos,  $x_{1i}$ , y el número de oficinas,  $x_{2i}$ , el margen de beneficios observado,  $y_i$ , es la suma de dos partes: el valor esperado,  $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ , y el término de error aleatorio,  $\varepsilon_i$ . El término de error aleatorio puede concebirse como la combinación de los efectos de otros muchos factores sin identificar que afectan a los márgenes de beneficios. La Figura 13.2 ilustra el modelo; el plano indica el valor esperado de varias combinaciones de las variables independientes y la  $\varepsilon_i$  es la desviación entre el plano —el valor esperado— y el valor observado de  $Y$  —marcado con un punto grande— de un punto de dato específico. En general, los valores observados de  $Y$  no se encuentran en el plano sino por encima o por debajo de él, debido a los términos de error positivos o negativos,  $\varepsilon_i$ .

La regresión simple, presentada en el capítulo anterior, no es más que un caso especial de la regresión múltiple con una única variable de predicción y, por lo tanto, el plano se reduce a una línea. Así pues, la teoría y el análisis que hemos desarrollado para la regresión simple también se aplican a la regresión múltiple. Sin embargo, existen algunas interpretaciones más que desarrollaremos en nuestro estudio de la regresión múltiple. Una de ellas se ilustra en el siguiente análisis de los gráficos tridimensionales.

### Gráficos tridimensionales

Tal vez sea más fácil comprender el método de regresión múltiple mediante una imagen gráfica simplificada. Observe el rincón de la habitación en la que está sentado. Las líneas formadas por las dos paredes y el suelo representan los ejes de dos variables independientes,  $X_1$  y  $X_2$ . La esquina que forman las dos paredes es el eje de la variable dependiente,  $Y$ . Para estimar una recta de regresión, reunimos conjuntos de puntos ( $x_{1i}$ ,  $x_{2i}$  e  $y_i$ ).

Representemos ahora estos puntos en su habitación utilizando las esquinas de las paredes y el suelo como los tres ejes. Con estos puntos suspendidos en su habitación, buscamos un plano en el espacio que se aproxime a todos ellos. Este plano es la forma geométrica de la ecuación de mínimos cuadrados. Con estos puntos en el espacio, ahora subimos y bajamos un plano y lo hacemos girar en dos direcciones: todos estos movimientos los hacemos simultáneamente hasta que tenemos un plano que está «cerca» de todos los puntos. Recuerde que en el Capítulo 12 hicimos esto con una línea recta en dos dimensiones para obtener una ecuación

$$\hat{y} = b_0 + b_1 x$$

A continuación, extendemos esa idea a tres dimensiones para obtener una ecuación

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

Este proceso es, por supuesto, más complicado que en el caso de la regresión simple. Pero los problemas reales son complicados y la regresión permite analizar mejor la complejidad de estos problemas. Queremos saber cómo varía  $Y$  cuando varía  $X_1$ . Pero sabemos que en estas variaciones influye, a su vez, la forma en que varía  $X_2$ . Y si  $X_1$  y  $X_2$  siempre varían a la vez, no podemos saber cuánto contribuye cada variable a las variaciones de  $Y$ .



Las interpretaciones geométricas de la regresión múltiple son cada vez más complejas a medida que aumenta el número de variables independientes. Sin embargo, la analogía con la regresión simple es extraordinariamente útil. Estimamos los coeficientes minimizando la suma de los cuadrados de las desviaciones de la dimensión  $Y$  en torno a una función lineal de las variables independientes. En la regresión simple, la función es una línea recta en un gráfico bidimensional. Con dos variables independientes, la función es un plano en un espacio tridimensional. Cuando consideramos más de dos variables independientes, tenemos varios hiperplanos complejos que son imposibles de visualizar.

## EJERCICIOS

### Ejercicios básicos

**13.1.** Dado el modelo lineal estimado

$$\hat{y} = 10 + 3x_1 + 2x_2 + 4x_3$$

- a) Calcule  $\hat{y}$  cuando  $x_1 = 20$ ,  $x_2 = 11$  y  $x_3 = 10$ .
- b) Calcule  $\hat{y}$  cuando  $x_1 = 15$ ,  $x_2 = 14$  y  $x_3 = 20$ .
- c) Calcule  $\hat{y}$  cuando  $x_1 = 35$ ,  $x_2 = 19$  y  $x_3 = 25$ .
- d) Calcule  $\hat{y}$  cuando  $x_1 = 10$ ,  $x_2 = 17$  y  $x_3 = 30$ .

**13.2.** Dado el modelo lineal estimado

$$\hat{y} = 10 + 5x_1 + 4x_2 + 2x_3$$

- a) Calcule  $\hat{y}$  cuando  $x_1 = 20$ ,  $x_2 = 11$  y  $x_3 = 10$ .
- b) Calcule  $\hat{y}$  cuando  $x_1 = 15$ ,  $x_2 = 14$  y  $x_3 = 20$ .
- c) Calcule  $\hat{y}$  cuando  $x_1 = 35$ ,  $x_2 = 19$  y  $x_3 = 25$ .
- d) Calcule  $\hat{y}$  cuando  $x_1 = 10$ ,  $x_2 = 17$  y  $x_3 = 30$ .

**13.3.** Dado el modelo lineal estimado

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

- a) Calcule  $\hat{y}$  cuando  $x_1 = 20$ ,  $x_2 = 11$  y  $x_3 = 10$ .
- b) Calcule  $\hat{y}$  cuando  $x_1 = 15$ ,  $x_2 = 24$  y  $x_3 = 20$ .
- c) Calcule  $\hat{y}$  cuando  $x_1 = 20$ ,  $x_2 = 19$  y  $x_3 = 25$ .
- d) Calcule  $\hat{y}$  cuando  $x_1 = 10$ ,  $x_2 = 9$  y  $x_3 = 30$ .

**13.4.** Dado el modelo lineal estimado

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

- a) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_1$  aumenta en 4?
- b) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_3$  aumenta en 1?

- c) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_2$  aumenta en 2?

**13.5.** Dado el modelo lineal estimado

$$\hat{y} = 10 - 2x_1 - 14x_2 + 6x_3$$

- a) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_1$  aumenta en 4?
- b) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_3$  disminuye en 1?
- c) ¿Cuál es la variación de  $\hat{y}$  cuando  $x_2$  disminuye en 2?

### Ejercicios aplicados

**13.6.** Una empresa aeronáutica quería predecir el número de horas de trabajo necesario para acabar el diseño de un nuevo avión. Se pensaba que las variables explicativas relevantes eran la velocidad máxima del avión, su peso y el número de piezas que tenía en común con otros modelos construidos por la empresa. Se tomó una muestra de 27 aviones de la empresa y se estimó el siguiente modelo:

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + \varepsilon_i$$

donde

$y_i$  = esfuerzo de diseño en millones de horas de trabajo

$x_{1i}$  = velocidad máxima del avión, en kilómetros por hora

$x_{2i}$  = peso del avión, en toneladas

$x_{3i}$  = número porcentual de piezas en común con otros modelos

Los coeficientes de regresión estimados eran

$$b_1 = 0,661 \quad b_2 = 0,065 \quad b_3 = -0,018$$

Interprete estas estimaciones.

- 13.7.** En un estudio de la influencia de las instituciones financieras en los tipos de interés de los bonos alemanes, se analizaron datos trimestrales de un periodo de 12 años. El modelo postulado era

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

donde

$y_i$  = variación de los tipos de interés de los bonos en el trimestre

$x_{1i}$  = variación de las compras de bonos por parte de las instituciones financieras en el trimestre

$x_{2i}$  = variación de las ventas de bonos por parte de las instituciones financieras en el trimestre

Los coeficientes de regresión parcial estimados eran

$$b_1 = 0,057 \quad b_2 = -0,065$$

Interprete estas estimaciones.

- 13.8.** Se ajustó el siguiente modelo a una muestra de 30 familias para explicar el consumo de leche por familia:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

donde

$y_i$  = consumo de leche, en litros a la semana

$x_1$  = renta semanal en cientos de dólares

$x_2$  = tamaño de la familia

Las estimaciones de los parámetros de la regresión por mínimos cuadrados eran

$$b_0 = -0,025 \quad b_1 = 0,052 \quad b_2 = 1,14$$

**a)** Interprete las estimaciones  $b_1$  y  $b_2$ .

**b)** ¿Es posible hacer una interpretación de la estimación  $b_0$  que tenga sentido?

- 13.9.** Se ajustó el siguiente modelo a una muestra de 25 estudiantes utilizando datos obtenidos al final de su primer año de universidad. El objetivo era explicar el aumento de peso de los estudiantes.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

donde

$y_i$  = aumento de peso en kilos durante el primer año

$x_{1i}$  = número medio de comidas a la semana

$x_{2i}$  = número medio de horas de ejercicio a la semana

$x_{3i}$  = número medio de cervezas consumidas a la semana

Las estimaciones de los parámetros de la regresión por mínimos cuadrados eran

$$b_0 = 7,35 \quad b_1 = 0,653$$

$$b_2 = -1,345 \quad b_3 = 0,613$$

**a)** Interprete las estimaciones  $b_1$ ,  $b_2$  y  $b_3$ .

**b)** ¿Es posible hacer una interpretación de la estimación  $b_0$  que tenga sentido?

## 13.2. Estimación de coeficientes

Los coeficientes de regresión múltiple se calculan utilizando estimadores obtenidos mediante el método de mínimos cuadrados. Este método de mínimos cuadrados es similar al que presentamos en el Capítulo 12 para la regresión simple. Sin embargo, los estimadores son complicados debido a las relaciones entre las variables independientes  $X_j$  que ocurren simultáneamente con las relaciones entre las variables independientes y la variable dependiente. Por ejemplo, si dos variables independientes aumentan o disminuyen al mismo tiempo —correlación positiva o negativa— mientras que al mismo tiempo la variable dependiente aumenta o disminuye, no podemos saber qué variable independiente está relacionada realmente con la variación de la variable dependiente. Como consecuencia, observamos que los coeficientes de regresión estimados son menos fiables si hay estrechas correlaciones entre dos variables independientes o más. Las estimaciones de los coeficientes y sus varianzas siempre se obtienen por computador. Sin embargo, dedicaremos bastantes esfuerzos a estudiar el álgebra y las formas de calcular la regresión por mínimos cuadrados. Estos esfuerzos permitirán comprender el método y averiguar cómo influyen las diferentes pautas de los datos en los resultados. Comenzamos con los supuestos habituales del modelo de regresión múltiple.