

# CAPÍTULO 14



## Regresión lineal simple

### CONTENIDO

LA ESTADÍSTICA  
EN LA PRÁCTICA:  
ALLIANCE DATA SYSTEMS

#### 14.1 MODELO DE REGRESIÓN LINEAL SIMPLE

Modelo de regresión  
y ecuación de regresión  
Ecuación de regresión estimada

#### 14.2 MÉTODO DE MÍNIMOS CUADRADOS

#### 14.3 COEFICIENTE DE DETERMINACIÓN

Coeficiente de correlación

#### 14.4 SUPOSICIONES DEL MODELO

#### 14.5 PRUEBA DE SIGNIFICANCIA

Estimación de  $\sigma^2$   
Prueba  $t$   
Intervalo de confianza para  $\beta_1$   
Prueba  $F$   
Algunas advertencias acerca de  
la interpretación de las pruebas  
de significancia

#### 14.6 USO DE LA ECUACIÓN DE REGRESIÓN ESTIMADA PARA ESTIMACIONES Y PREDICCIONES

Estimación puntual  
Estimación por intervalo  
Intervalo de confianza  
para el valor medio de  $y$   
Intervalo de predicción para  
un solo valor de  $y$

#### 14.7 SOLUCIÓN POR COMPUTADORAS

#### 14.8 ANÁLISIS RESIDUAL: CONFIRMACIÓN DE LAS SUPOSICIONES DEL MODELO

Gráfica de residuales contra  $x$   
Gráfica de residuales contra  $\hat{y}$   
Residuales estandarizados  
Gráfica de probabilidad normal

#### 14.9 ANÁLISIS DE RESIDUALES: OBSERVACIONES ATÍPICAS Y OBSERVACIONES INFLUYENTES

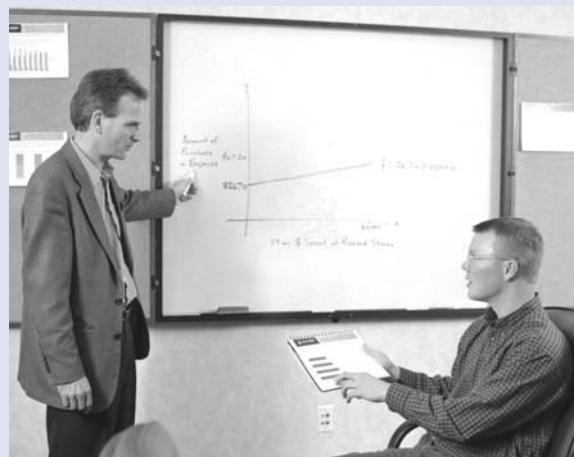
Detección de observaciones  
atípicas  
Detección de observaciones  
influyentes

**LA ESTADÍSTICA *(en)* LA PRÁCTICA**
**ALLIANCE DATA SYSTEMS\***
*DALLAS TEXAS*

Alliance Data Systems (ADS), una empresa de la creciente industria de administración de la relación con el cliente (Customer Relationship Management, CRM) proporciona servicios de transacciones, crédito y mercadotecnia. Los clientes de ADS están concentrados en cuatro industrias: industria minorista, supermercados pequeños, derivados del petróleo/energía eléctrica y transporte. En 1983, Alliance empezó ofreciendo servicios extremo a extremo de tramitación de crédito para la industria minorista, la industria de derivados del petróleo y la industria de restaurantes de categoría media; actualmente ADS emplea a más de 6500 personas que proporcionan servicios a clientes en todo el mundo. ADS, sólo en Estados Unidos, opera más de 140 000 terminales de punto de venta, y procesa más de 2.5 miles de millones de transacciones anuales. En Estados Unidos ADS es la segunda empresa en servicios de crédito de establecimientos locales representando 49 programas de establecimientos locales con casi 72 millones de tarjetahabientes. En 2001, ADS hizo una oferta pública inicial y ahora cotiza en la bolsa de Nueva York.

Uno de los servicios de mercadotecnia ofrecidos por ADS es el de campañas y publicidad directas por correo. La empresa posee una base de datos con información sobre los hábitos de consumo de más de 100 millones de consumidores, lo que le permite dirigir sus acciones a los consumidores que tienen la mayor probabilidad de beneficiarse de la publicidad por correo. El Grupo de desarrollo analítico de ADS emplea el análisis de regresión en la obtención de modelos para medir y predecir la receptividad del consumidor a las campañas de mercadotecnia directa. Algunos modelos de regresión predicen la probabilidad de compra de las personas que reciben la publicidad y otros predicen la cantidad que gastarán cuando realicen una compra.

En una determinada campaña, una cadena de tiendas deseaba atraer a nuevos clientes. Para predecir el efecto de la campaña, los analistas de ADS tomaron de la base de datos una muestra de consumidores, les enviaron material promocional y después recogieron datos sobre la respuesta de los consumidores. Los datos recogidos se referían al monto de la compra realizada por los consumidores que respondieron a la campaña, así como a diversas variables específicas del consumidor, que se consideraron útiles para



Analistas de ADS discuten sobre el uso del modelo de regresión para predecir las ventas en una campaña de comercialización directa. © Cortesía de Alliance Data Systems.

predecir las ventas. La variable del consumidor que más contribuyó a predecir el monto de compra fue la cantidad total de compras a crédito realizadas en tiendas semejantes en los últimos 39 meses. Los analistas de ADS obtuvieron una ecuación de regresión estimada con la que se relacionaba el monto de compra con la cantidad gastada en tiendas semejantes:

$$\hat{y} = 26.7 + 0.00205x$$

donde

$\hat{y}$  = monto de compra

$x$  = monto gastado en tiendas similares

Con esta ecuación, pudieron predecir que una persona que hubiera gastado \$10 000 en tiendas semejantes en los últimos 39 meses, gastaría \$47.20 como respuesta a la publicidad directa por correo. En este capítulo se verá cómo obtener estas ecuaciones de regresión estimada.

En el modelo final que obtuvieron los analistas de ADS también participaban algunas otras variables que incrementaban el poder predictivo de la ecuación de predicción. Entre estas variables se encontraba la existencia o no de una tarjeta de crédito, el ingreso estimado, y la cantidad promedio gastada en cada visita a la tienda seleccionada. En el capítulo siguiente se verá cómo incorporar estas variables adicionales a un modelo de regresión múltiple.

\*Los autores agradecemos a Philip Cleman de Desarrollo analítico de Alliance Data Systems por proporcionarnos este artículo para *La estadística en la práctica*.

En la administración, las decisiones suelen basarse en la relación entre dos o más variables. Por ejemplo, observar la relación entre el gasto en publicidad y las ventas puede permitir a un gerente de mercadotecnia tratar de predecir las ventas correspondientes a un determinado gasto en publicidad. O, una empresa de servicios públicos puede emplear la relación entre la temperatura diaria y la demanda de electricidad para predecir la demanda de electricidad considerando las temperaturas diarias que se esperan el mes siguiente. Algunas veces los directivos se apoyan en la intuición para juzgar la relación entre dos variables. Sin embargo, cuando es posible tener datos, puede emplearse un procedimiento estadístico llamado *análisis de regresión* para obtener una ecuación que indique cuál es la relación entre las variables.

*Sir Francis Galton (1822-1911) fue el primero en emplear los métodos estadísticos para estudiar la relación entre dos variables. Galton estaba interesado en estudiar la relación entre la estatura de padre e hijo. Karl Pearson (1857-1936) analizó esta relación en 1078 pares de padres-hijo.*

En la terminología que se emplea en regresión, a la variable que se va a predecir se le llama **variable dependiente**. A la variable o variables que se usan para predecir el valor de la variable dependiente se les llama **variables independientes**. Por ejemplo, al analizar el efecto de los gastos en publicidad sobre las ventas, como lo que busca el gerente de mercadotecnia es predecir las ventas, esto indica que las ventas serán la variable dependiente.

En este capítulo se estudia el tipo más sencillo de análisis de regresión en el que interviene una variable independiente y una variable dependiente y en el que la relación entre estas variables es aproximada mediante una línea recta. A este tipo de análisis de regresión se le conoce como **regresión lineal simple**. Al análisis de regresión en el que intervienen dos o más variables independientes se le llama análisis de regresión múltiple; el análisis de regresión múltiple y los casos en los que la relación es curvilínea se estudian en los capítulos 15 y 16.

## 14.1

# Modelo de regresión lineal simple

Armand's Pizza Parlors es una cadena de restaurantes de comida italiana. Sus mejores ubicaciones son las que se encuentran cerca de los campus de las universidades. Los gerentes creen que las ventas trimestrales de estos restaurantes (que se denotan por  $y$ ) están directamente relacionadas con el tamaño de la población estudiantil (que se denota  $x$ ); es decir, en los restaurantes que están cerca de campus que tienen una población estudiantil grande se generan más ventas que en los restaurantes situados cerca de campus con una población estudiantil pequeña. Empleando el análisis de regresión, se puede obtener una ecuación que muestre cuál es la relación entre la variable dependiente  $y$  y la variable dependiente  $x$ .

## Modelo de regresión y ecuación de regresión

En el ejemplo de los restaurantes Armand's Pizza Parlors, la población consta de todos los restaurantes Armand. Para cada restaurante de la población, hay un valor  $x$  (población estudiantil) y un correspondiente valor  $y$  (ventas trimestrales). A la ecuación con que se describe cómo se relaciona  $y$  con  $x$  y en la que se da un término para el error, se le llama **modelo de regresión**. El siguiente es el modelo que se emplea en la regresión lineal simple.

### MODELO DE REGRESIÓN LINEAL SIMPLE

$$y = \beta_0 + \beta_1 x + \epsilon$$

(14.1)

$\beta_0$  y  $\beta_1$  se conocen como los parámetros del modelo, y  $\epsilon$  (la letra griega épsilon) es una variable aleatoria que se conoce como término del error. El término del error da cuenta de la variabilidad de  $y$  que no puede ser explicada por la relación lineal entre  $x$  y  $y$ .

La población de los restaurantes Armand's puede verse también como una colección de subpoblaciones, una para cada uno de los valores de  $x$ . Por ejemplo, una subpoblación está formada por todos los campus universitarios de 8000 estudiantes; otra subpoblación consta de todos los restaurantes Armand's localizados cerca de los campus universitarios de 9000 estudiantes; etc. Para cada subpoblación hay una distribución de valores  $y$ . Así, hay una distribución de valores  $y$  que corresponde a los restaurantes localizados cerca de los campus de 8000 estudiantes; hay otra distribución de valores  $y$  que corresponde a los restaurantes ubicados cerca de los campus de 9000 estudiantes, y así sucesivamente. Cada una de estas distribuciones de valores  $y$  tiene su propia media o valor esperado. A la ecuación que describe la relación entre el valor esperado de  $y$ , que se denota  $E(x)$ , y  $x$  se le llama **ecuación de regresión**. La siguiente es la ecuación de regresión para la regresión lineal simple.

### ECUACIÓN DE REGRESIÓN LINEAL SIMPLE

$$E(y) = \beta_0 + \beta_1 x \quad (14.2)$$

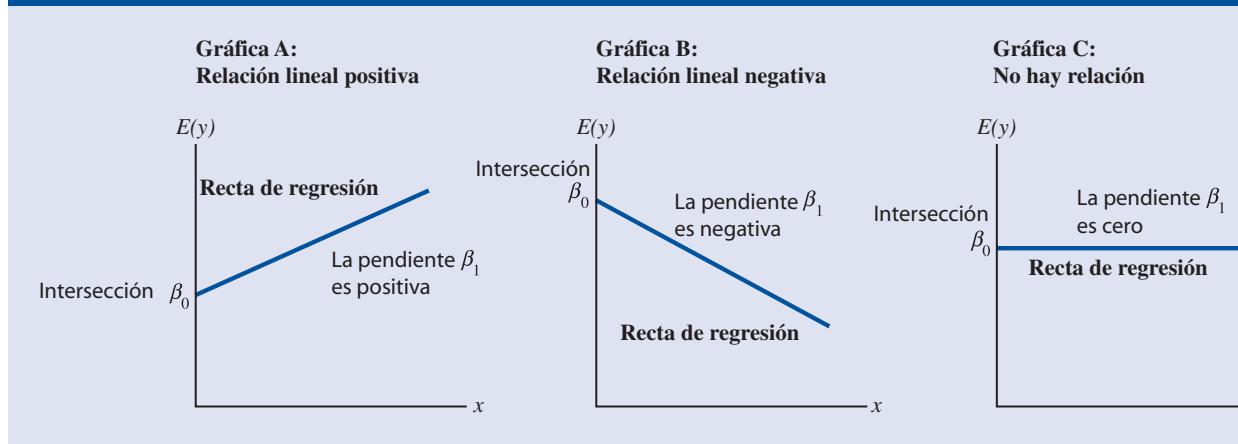
La gráfica de la ecuación de regresión lineal simple es una línea recta;  $\beta_0$  es la intersección de la recta de regresión con el eje  $y$ ,  $\beta_1$  es la pendiente y  $E(y)$  es la media o valor esperado de  $y$  para un valor dado de  $x$ .

En la figura 14.1 se presentan ejemplos de posibles rectas de regresión. La recta de regresión de la gráfica A indica que el valor medio de  $y$  está relacionado positivamente con  $x$ . La recta de regresión de la gráfica B indica que el valor medio de  $y$  está relacionado negativamente con  $x$ , valores menores de  $E(y)$  corresponden a valores mayores de  $x$ . La recta de regresión de la gráfica C muestra el caso en el que el valor medio de  $y$  no está relacionado con  $x$ ; es decir, el valor medio de  $y$  es el mismo para todos los valores de  $x$ .

### Ecuación de regresión estimada

Si se conocieran los valores de los parámetros poblacionales  $\beta_0$  y  $\beta_1$ , se podría emplear la ecuación (14.2) para calcular el valor medio de  $y$  para un valor dado de  $x$ . Sin embargo, en la práctica no se conocen los valores de estos parámetros y es necesario estimarlos usando datos muestrales. Se calculan estadísticos muestrales (que se denotan  $b_0$  y  $b_1$ ) como estimaciones de los parámetros poblacionales  $\beta_0$  y  $\beta_1$ . Sustituyendo en la ecuación de regresión  $b_0$  y  $b_1$  por los

**FIGURA 14.1** EJEMPLOS DE LÍNEAS DE REGRESIÓN EN LA REGRESIÓN LINEAL SIMPLE



valores de los estadísticos muestrales  $\beta_0$  y  $\beta_1$ , se obtiene la **ecuación de regresión estimada**. La ecuación de regresión estimada de la regresión lineal simple se da a continuación.

### ECUACIÓN DE REGRESIÓN LINEAL SIMPLE ESTIMADA

$$\hat{y} = b_0 + b_1 x \quad (14.3)$$

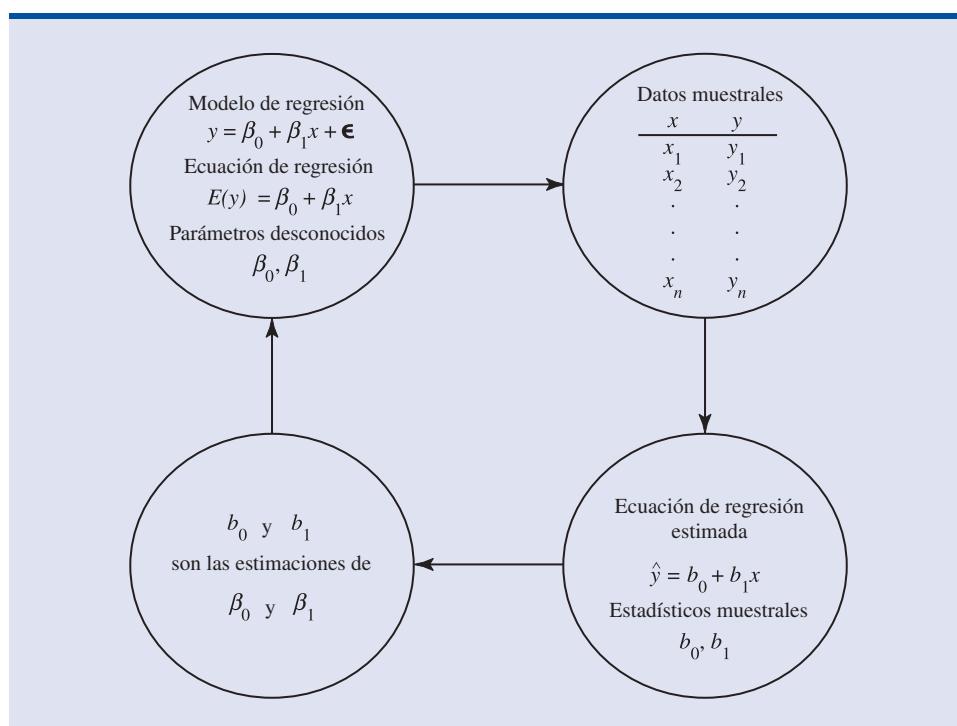
A la gráfica de la ecuación de regresión simple estimada se le llama *recta de regresión estimada*;  $b_0$  es la intersección con el eje y y  $b_1$  es la pendiente. En la sección siguiente se muestra el uso del método de mínimos cuadrados para calcular los valores de  $b_0$  y  $b_1$  para la ecuación de regresión estimada.

En general,  $\hat{y}$  es el estimador puntual de  $E(y)$ , el valor medio de las  $y$  para un valor dado de  $x$ . Por lo tanto, para estimar la media o el valor esperado de las ventas trimestrales de todos los restaurantes situados cerca de los campus de 10 000 estudiantes, Armand's tendrá que sustituir en la ecuación (14.3)  $x$  por 10 000, pero, en algunos casos, a Armand's lo que le interesará será predecir las ventas de un determinado restaurante. Por ejemplo, supóngase que Armand's deseé predecir las ventas trimestrales del restaurante que se encuentra cerca de Talbot Collage, una escuela de 10 000 estudiantes. Resulta que la mejor estimación de la  $y$  que corresponde a un determinado valor de  $x$  es también la proporcionada por  $\hat{y}$ . Por lo tanto, para predecir las ventas trimestrales del restaurante ubicado cerca de Talbot Collage, Armand's también sustituirá la  $x$  de la ecuación (14.3) por 10 000.

Como el valor de  $\hat{y}$  proporciona tanto una estimación puntual de  $E(x)$  para un valor dado de  $x$  como una estimación puntual de un solo valor de  $y$  para un valor dado de  $x$ , a  $\hat{y}$  se le llamará simplemente *valor estimado de  $y$* . En la figura 14.2 se presenta en forma resumida el proceso de estimación en la regresión lineal simple.

**FIGURA 14.2** PROCESO DE ESTIMACIÓN EN LA REGRESIÓN LINEAL SIMPLE

*La estimación de  $\beta_0$  y  $\beta_1$  es un proceso estadístico muy parecido a la estimación de  $\mu$  que se vio en el capítulo 7.  $\beta_0$  y  $\beta_1$  son los parámetros de interés que son desconocidos, y  $b_0$  y  $b_1$  son los estadísticos muestrales que se usan para estimar los parámetros.*



## NOTAS Y COMENTARIOS

1. El análisis de regresión no puede entenderse como un procedimiento para establecer una relación de causa y efecto entre las variables. Este procedimiento sólo indica cómo o en qué medida las variables están relacionadas una con otra. Conclusiones acerca de una relación causa y efecto deben basarse en los conocimientos de los especialistas en la aplicación de que se trate.
2. La ecuación de regresión en la regresión lineal simple es  $E(y) = \beta_0 + \beta_1 x$ . En libros más avanzados sobre análisis de regresión se suele escribir la ecuación de regresión como  $E(y|x) = \beta_0 + \beta_1 x$  enfatizando así que lo que proporciona esta ecuación es el valor medio de las y para un valor dado de x.

### 14.2

## Método de mínimos cuadrados

*En la regresión lineal simple, cada observación consta de dos valores: uno de la variable independiente y otro de la variable dependiente.*

El **método de mínimos cuadrados** es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada. Para ilustrar el método de mínimos cuadrados, supóngase que se recolectan datos de una muestra de 10 restaurantes Armand's Pizza Parlors ubicados todos cerca de campus universitarios. Para la observación  $i$  o el restaurante  $i$  de la muestra,  $x_i$  es el tamaño de la población de estudiantes (en miles) en el campus y  $y_i$  son las ventas trimestrales (en miles de dólares). En la tabla 14.1 se presentan los valores de  $x_i$  y  $y_i$  en esta muestra de 10 restaurantes. Como se ve, el restaurante 1, para el que  $x_1 = 2$  y  $y_1 = 58$ , está cerca de un campus de 2000 estudiantes y sus ventas trimestrales son de \$58 000. El restaurante 2, para el que  $x_2 = 6$  y  $y_2 = 105$ , está cerca de un campus de 6000 estudiantes y sus ventas trimestrales son de \$105 000. El valor mayor es el que corresponde a ventas del restaurante 10, el cual está cerca de un campus de 26 000 estudiantes y sus ventas trimestrales son de \$202 000.

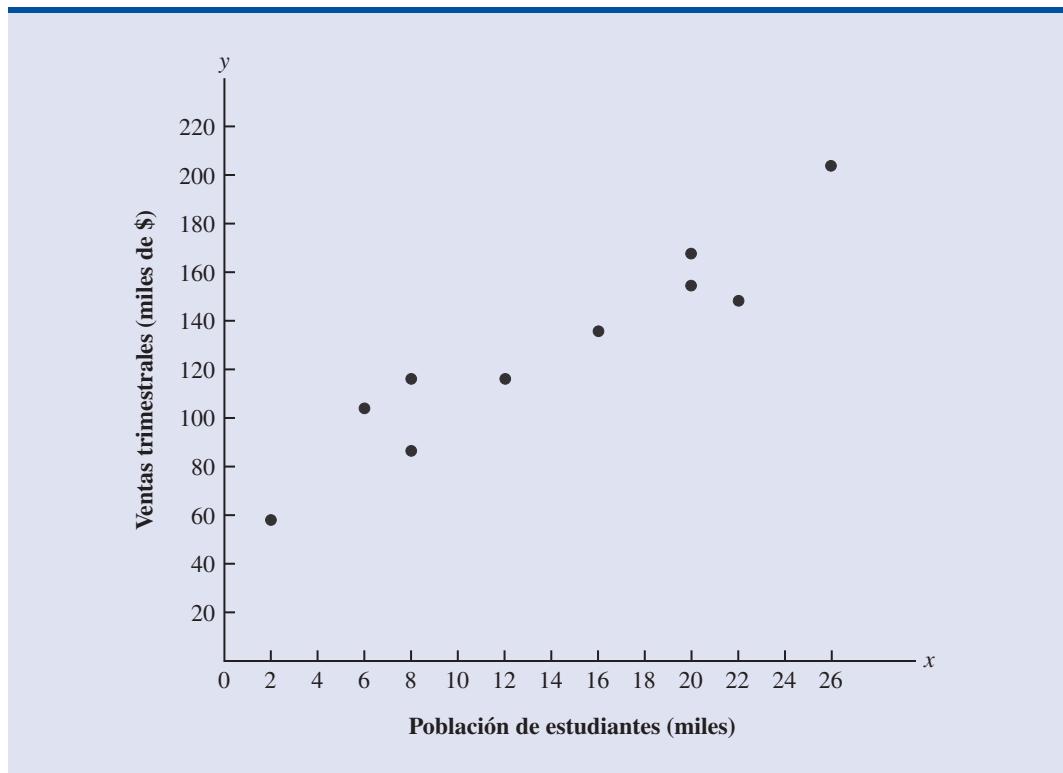
La figura 14.3 es el diagrama de dispersión de los datos de la tabla 14.1. La población de estudiantes se indica en el eje horizontal y las ventas trimestrales en el eje vertical. Los **diagramas de dispersión** para el análisis de regresión se trazan colocando la variable independiente  $x$  en el eje horizontal y la variable dependiente  $y$  en el eje vertical. El diagrama de dispersión permite observar gráficamente los datos y obtener conclusiones acerca de la relación entre las variables.

¿Qué conclusión preliminar se puede obtener de la figura 14.3? Las ventas trimestrales parecen ser mayores cerca de campus en los que la población de estudiantes es mayor. Además, en estos datos se observa que la relación entre el tamaño de la población de estudiantes y las ventas trimestrales parece poder aproximarse mediante una línea recta; en efecto, se observa que hay

**TABLA 14.1 POBLACIÓN DE ESTUDIANTES Y VENTAS TRIMESTRALES EN 10 RESTAURANTES ARMAND'S PIZZA PARLORS**

Restaurante $i$	Población de estudiantes (miles) $x_i$	Ventas trimestrales (miles de \$) $y_i$
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

**FIGURA 14.3** DIAGRAMA DE DISPERSIÓN EN EL QUE SE MUESTRA LA POBLACIÓN DE ESTUDIANTES Y LAS VENTAS TRIMESTRALES DE ARMAND'S PIZZA PARLORS



una relación lineal positiva entre  $x$  y  $y$ . Por tanto, para representar la relación entre ventas trimestrales y la población de estudiantes, se elige el modelo de regresión lineal simple. Decidido esto, la tarea siguiente es usar los datos muestrales de la tabla 14.1 para determinar los valores de  $b_0$  y  $b_1$  en la ecuación de regresión lineal simple. Para el restaurante  $i$ , la ecuación de regresión simple estimada es

$$\hat{y}_i = b_0 + b_1 x_i \quad (14.4)$$

donde

- $\hat{y}_i$  = valor estimado de las ventas trimestrales (en miles de dólares) del restaurante  $i$
- $b_0$  = intersección de la recta de regresión con el eje  $y$
- $b_1$  = pendiente de la recta de regresión
- $x_i$  = tamaño de la población de estudiantes (en miles) del restaurante  $i$

Como para el restaurante  $i$ ,  $y_i$  denota ventas observadas (reales) y  $\hat{y}_i$  denota ventas estimadas mediante la ecuación (14.4), para cada uno de los restaurantes de la muestra habrá un valor de ventas observadas  $y_i$  y un valor de ventas estimadas  $\hat{y}_i$ . Para que la recta de regresión estimada proporcione un buen ajuste a los datos, las diferencias entre los valores observados y los valores estimados deben ser pequeñas.

En el método de mínimos cuadrados se usan los datos muestrales para obtener los valores de  $b_0$  y  $b_1$  que minimicen la *suma de los cuadrados de las desviaciones (diferencias)* entre los valores observados de la variable dependiente  $y_i$  y los valores estimados de la variable dependiente. El criterio que se emplea en el método de mínimos cuadrados es el de la expresión (14.5).

*Carl Friedrich Gauss (1777- 1855) fue quien propuso el método de mínimos cuadrados.*

### CRITERIO DE MÍNIMOS CUADRADOS

$$\min \sum (y_i - \hat{y}_i)^2 \quad (14.5)$$

donde

$y_i$  = valor observado de la variable dependiente en la observación  $i$

$\hat{y}_i$  = valor estimado de la variable independiente en la observación  $i$

Se puede usar cálculos diferenciales para demostrar (véase apéndice 14.1) que los valores de  $b_0$  y  $b_1$  que minimiza la expresión (14.5) se pueden encontrar usando las ecuaciones (14.6) y (14.7).

### PENDIENTE E INTERSECCIÓN CON EL EJE Y DE LA ECUACIÓN DE REGRESIÓN ESTIMADA\*

*Al calcular  $b_1$  con una calculadora, en los cálculos intermedios deben llevarse tantas cifras significativas como sea posible. Se recomienda llevar por lo menos cuatro cifras significativas*

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.7)$$

donde

$x_i$  = valor de la variable independiente en la observación  $i$

$y_i$  = valor de la variable dependiente en la observación  $i$

$\bar{x}$  = media de la variable independiente

$\bar{y}$  = media de la variable dependiente

$n$  = número total de observaciones

En la tabla 14.2 se presentan los cálculos necesarios para obtener la ecuación de regresión estimada en el ejemplo de Armand's Pizza Parlors. Como la muestra es de 10 restaurantes, tenemos 10 observaciones. Dado que en las ecuaciones (14.6) y (14.7) se necesitan  $\bar{x}$  y  $\bar{y}$ , se empieza por calcular  $\bar{x}$  y  $\bar{y}$ .

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Usando las ecuaciones (14.6) y (14.7) y la información de la tabla 14.2, se calcula la pendiente y la intersección con el eje y de la ecuación de regresión de Armand's Pizza Parlors. La pendiente ( $b_1$ ) se calcula como sigue.

\*Otra fórmula de calcular  $b_1$ , es

$$b_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

Esta forma de la ecuación (14.6) suele recomendarse cuando se emplea una calculadora para calcular  $b_1$ .

**TABLA 14.2** ECUACIÓN DE REGRESIÓN ESTIMADA PARA ARMAND'S PIZZA PARLORS OBTENIDA POR EL MÉTODO DE MÍNIMOS CUADRADOS

Restaurante $i$	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totales	140	1300			2840	568
	$\Sigma x_i$	$\Sigma y_i$			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$\begin{aligned} b_1 &= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \\ &= \frac{2840}{568} \\ &= 5 \end{aligned}$$

La intersección con el eje  $y$  ( $b_0$ ) se calcula como sigue.

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 130 - 5(14) \\ &= 60 \end{aligned}$$

Por lo tanto, la ecuación de regresión estimada es

$$\hat{y} = 60 + 5x$$

En la figura 14.4 se muestra esta ecuación graficada sobre el diagrama de dispersión.

La pendiente de la ecuación de regresión estimada ( $b_1 = 5$ ) es positiva, lo que implica que a medida que aumenta el tamaño de la población de estudiantes, aumentan las ventas. Se concluye (basándose en las ventas dadas en miles de \$ y en el tamaño de la población de estudiantes en miles) que un aumento de 1000 en el tamaño de la población de estudiantes corresponde a un aumento esperado de \$5000 en las ventas; es decir, se espera que las ventas trimestrales aumenten \$5 por cada aumento de un estudiante.

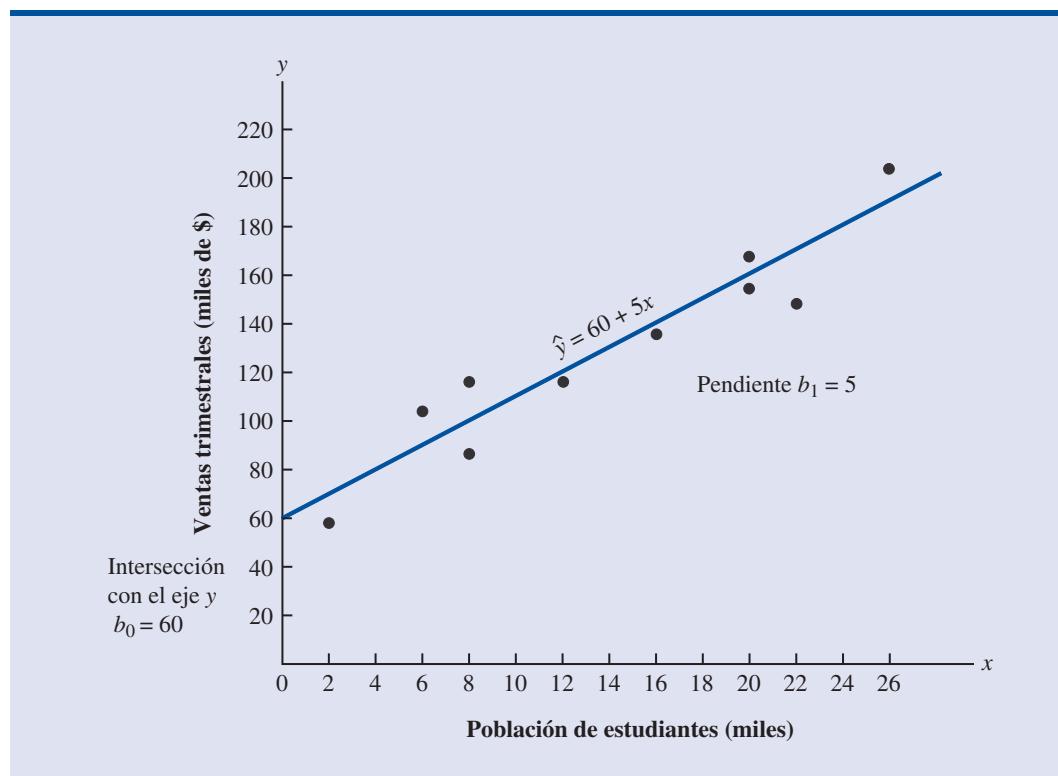
Si se considera que la ecuación de regresión estimada obtenida por el método de mínimos cuadrados describe adecuadamente la relación entre  $x$  y  $y$ , parecerá razonable usar esta ecuación de regresión estimada para estimar el valor de  $y$  para un valor dado de  $x$ . Por ejemplo, si se quisieran predecir las ventas trimestrales de un restaurante ubicado cerca de un campus de 16 000 estudiantes, se calcularía

$$\hat{y} = 60 + 5(16) = 140$$

De manera que las ventas trimestrales pronosticadas para este restaurante serían de \$140 000. En la sección siguiente se verán los métodos para evaluar el uso correcto de la ecuación de regresión para hacer estimaciones y predicciones.

*Debe tenerse mucho cuidado al usar la ecuación de regresión estimada para hacer predicciones fuera del rango de valores de la variable independiente, ya que fuera de ese rango no puede asegurarse que esta relación sea válida.*

**FIGURA 14.4** GRÁFICA DE LA ECUACIÓN DE REGRESIÓN ESTIMADA DE ARMAND'S PIZZA PARLORS:  $\hat{y} = 60 + 5x$



### NOTAS Y COMENTARIOS

El método de mínimos cuadrados proporciona una ecuación de regresión estimada que minimiza la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente  $y_i$  y los valores estimados de la variable dependiente  $\hat{y}_i$ . El criterio de mínimos cuadrados permite obtener la

ecuación de mejor ajuste. Si se empleara otro criterio, como minimizar la suma de las desviaciones absolutas entre  $y_i$  y  $\hat{y}_i$ , se obtendría una ecuación diferente. En la práctica el método de mínimos cuadrados es el método más usado.

### Ejercicios

#### Método

- Dadas las siguientes cinco observaciones de las variables  $x$  y  $y$ .

**Autoexamen**

$x_i$	1	2	3	4	5
$y_i$	3	7	5	11	14

- Trace el diagrama de dispersión correspondiente a estos datos.
- ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?

- c. Trate de aproximar la relación entre  $x$  y  $y$  trazando una línea recta que pase a través de los puntos de los datos.
- d. Con las ecuaciones (14.6) y (14.7) calcule  $b_0$  y  $b_1$  para obtener la ecuación de regresión estimada.
- e. Use la ecuación de regresión estimada para predecir el valor de  $y$  cuando  $x = 4$ .
2. Dadas las siguientes cinco observaciones de las variables  $x$  y  $y$ .

$x_i$	3	12	6	20	14
$y_i$	55	40	55	10	15

- a. Trace, con estos datos, el diagrama de dispersión.
- b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?
- c. Trate de aproximar la relación entre  $x$  y  $y$  trazando una línea recta a través de los puntos de los datos.
- d. Con las ecuaciones (14.6) y (14.7) calcule  $b_0$  y  $b_1$ , para obtener la ecuación de regresión estimada.
- e. Use la ecuación de regresión estimada para predecir el valor de  $y$  cuando  $x = 4$ .
3. Dadas las observaciones siguientes sobre estas dos variables obtenidas en un estudio de regresión.

$x_i$	2	6	9	13	20
$y_i$	7	18	9	26	23

- a. Con estos datos trace el diagrama de dispersión.
- b. Obtenga la ecuación de regresión estimada correspondiente a estos datos.
- c. Use la ecuación de regresión estimada para predecir el valor de  $y$  cuando  $x = 4$ .

## Aplicaciones

**Autoexamen**

4. Los datos siguientes son estaturas y pesos de nadadoras.

Estatura	68	64	62	65	66
Peso	132	108	102	115	128

- a. Trace el diagrama de dispersión de estos datos usando la estatura como variable independiente.
- b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?
- c. Trate de aproximar la relación entre estatura y peso trazando una línea recta a través de los puntos de los datos.
- d. Obtenga la ecuación de regresión estimada calculando  $b_0$  y  $b_1$ .
- e. Si la estatura de una nadadora es 63 pulgadas, ¿cuál será su peso estimado?
5. Los adelantos tecnológicos han hecho posible fabricar botes inflables. Estos botes de goma inflables, que pueden enrollarse formando un paquete no mayor que una bolsa de golf, tienen tamaño suficiente para dos pasajeros con su equipo de excursionismo. La revista *Canoe & Kayak* probó los botes de nueve fabricantes para ver su funcionamiento en un recorrido de tres días. Uno de los criterios de evaluación fue su capacidad para equipaje que se evaluó utilizando una escala de 4 puntos, siendo 1 la puntuación más baja y 4 la puntuación más alta. Los datos siguientes muestran la evaluación que obtuvieron respecto a capacidad para equipaje y los precios de los botes (*Canoe Kayak*, marzo 2003).

archivo  
en CD  
Boats

Bote	Capacidad para equipaje	Precio (\$)
S14	4	1595
Orinoco	4	1399
Outside Pro	4	1890
Explorer 380X	3	795
River XK2	2.5	600
Sea Tiger	4	1995
Maverik II	3	1205
Starlite 100	2	583
Fat Pack Cat	3	1048

- Trace el diagrama de dispersión de estos datos empleando la capacidad para equipaje como variable independiente.
  - ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre capacidad para equipaje y precio?
  - A través de los puntos de los datos trace una línea recta para aproximar la relación lineal entre capacidad para equipaje y precio.
  - Utilice el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
  - Dé una interpretación de la pendiente de la ecuación de regresión estimada.
  - Diga cuál será el precio de un bote que tenga 3 en la evaluación de su capacidad para equipaje.
6. Wageweb realiza estudios sobre datos salariales y presenta resúmenes de éstos en su sitio de la Red. Basándose en datos salariales desde el 1 de octubre de 2002 Wageweb publicó que el salario anual promedio de los vicepresidentes de ventas era \$142 111 con una gratificación anual promedio de \$15 432 (Wageweb.com, 13 de mayo de 2003). Suponga que los datos siguientes sean una muestra de salarios y bonos anuales de 10 vicepresidentes de ventas. Los datos se dan en miles de dólares.

archivo  
en CD  
VPSalary

Vicepresidente	Salario	Gratificación
1	135	12
2	115	14
3	146	16
4	167	19
5	165	22
6	176	24
7	98	7
8	136	17
9	163	18
10	119	11

- Trace un diagrama de dispersión con estos datos tomando como variable independiente los salarios.
  - ¿Qué indica el diagrama de dispersión del inciso a) acerca de la relación entre salario y gratificación?
  - Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
  - Dé una interpretación de la ecuación de regresión estimada.
  - ¿Cuál será la gratificación de un vicepresidente que tenga un salario anual de \$120 000?
7. ¿Esperaría que los automóviles más confiables fueran los más caros? *Consumer Reports* evaluó 15 de los mejores automóviles sedán. La confiabilidad se evaluó con una escala de 5 puntos: mala (1), regular (2), buena (3), muy buena (4) y excelente (5). Los precios y la evaluación sobre la confiabilidad de estos 15 automóviles se presenta en la tabla siguiente (*Consumer Reports*, febrero de 2004).

archivo  
en  
**CD**  
Cars

Marca y modelo	Confiabilidad	Precio (\$)
Acura TL	4	33 150
BMW 330i	3	40 570
Lexus IS300	5	35 105
Lexus ES330	5	35 174
Mercedes-Benz C320	1	42 230
Lincoln LS Premium (V6)	3	38 225
Audi A4 3.0 Quattro	2	37 605
Cadillac CTS	1	37 695
Nissan Maxima 3.5 SE	4	34 390
Infiniti I35	5	33 845
Saab 9-3 Aero	3	36 910
Infiniti G35	4	34 695
Jaguar X-Type 3.0	1	37 995
Saab 9-5 Arc	3	36 955
Volvo S60 2.5T	3	33 890

- Trace un diagrama de dispersión con estos datos tomando como variable independiente las evaluaciones de confiabilidad.
  - Dé la ecuación de regresión obtenida por el método de mínimos cuadrados.
  - De acuerdo con este análisis, ¿cree usted que los automóviles más confiables sean más caros?
  - Estime el precio de un automóvil sedán cuya evaluación de confiabilidad sea 4.
8. Las bicicletas de montaña que actualmente cuestan menos de \$1000 tienen muchos de los componentes de alta calidad que hasta hace poco sólo tenían los modelos de alta calidad. Hoy, incluso modelos de menos de \$1000 suelen ofrecer suspensión flexible, pedales clipless y cuadro muy bien diseñado. Una cuestión interesante es si precios más altos corresponden a mayor facilidad de manejo, medida a través del agarre lateral de la bicicleta. Para medir el agarre lateral, *Outside Magazine* empleó una escala de evaluación del 1 al 5, en la que el 1 correspondía a mala y 5 a promedio. A continuación se presenta el agarre lateral y los precios de 10 bicicletas de montaña probadas por *Outside Magazine* (*Outside Magazine Buyer's Guide*, 2001)

archivo  
en  
**CD**  
MtnBikes

Fabricante y modelo	Agarre lateral	Precio (\$)
Raleigh M80	1	600
Marin Bear Valley Feminina	1	649
GT Avalanche 2.0	2	799
Kona Jake the Snake	1	899
Schwinn Moab 2	3	950
Giant XTC NRS 3	4	1100
Fisher Paragon Genesisters	4	1149
Jamis Dakota XC	3	1300
Trek Fuel 90	5	1550
Specialized Stumpjumper M4	4	1625

- Trace un diagrama de dispersión con estos datos tomando como variable independiente el agarre lateral.
- ¿Parecen indicar estos datos que los modelos más caros sean de más fácil manejo? Explique.
- Dé la ecuación de regresión estimada obtenida por el método de mínimos cuadrados.
- ¿Cuál es el precio estimado de una bicicleta de montaña cuyo agarre lateral tenga una evaluación de 4?

9. Un gerente de ventas recolectó los datos siguientes sobre ventas anuales y años de experiencia.

archivo en CD  
Sales

Vendedor	Años de experiencia	Ventas anuales (miles de \$)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

- a. Elabore un diagrama de dispersión con estos datos, en el que la variable independiente sean los años de experiencia.
  - b. Dé la ecuación de regresión estimada que puede emplearse para predecir las ventas anuales cuando se conocen los años de experiencia.
  - c. Use la ecuación de regresión estimada para pronosticar las ventas anuales de un vendedor de 9 años de experiencia.
10. Bergans of Norway ha estado fabricando equipo para excursionismo desde 1908. En los datos que se presentan en la tabla siguiente se da la temperatura ( $^{\circ}\text{F}$ ) y el precio (\$) de 11 modelos de sacos de dormir fabricados por Bergans (*Backpacker 2006 Gear Guide*)

archivo en CD  
SleepingBags

Modelo	Temperatura	Precio
Ranger 3-Seasons	12	319
Ranger Spring	24	289
Ranger Winter	3	389
Rondane 3-Seasons	13	239
Rondane Summer	38	149
Rondane Winter	4	289
Senja Ice	5	359
Senja Snow	15	259
Senja Zero	25	229
Super Light	45	129
Tight & Light	25	199

- a. Trace un diagrama de dispersión con estos datos, en el que la variable independiente sea la temperatura ( $^{\circ}\text{F}$ ).
  - b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre temperatura y precio?
  - c. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
  - d. Prediga cuál será el precio de un saco de dormir si el índice de temperatura ( $^{\circ}\text{F}$ ) es 20.
11. Aunque actualmente en los aeropuertos grandes los retrasos son menos frecuentes, es útil saber en qué aeropuertos es más probable que le echen a perder a uno sus planes. Además, si su vuelo llega con retraso a un determinado aeropuerto en el que tiene que hacer un trasbordo, ¿cuál es la probabilidad de que se retrase la salida y que pueda hacer así el trasbordo? En la tabla siguiente se muestra el porcentaje de llegadas y salidas retrasadas durante el mes de agosto en 13 aeropuertos (*Business 2.0*, febrero 2002).

Aeropuerto	Llegadas retrasadas (%)	Salidas retrasadas (%)
Atlanta	24	22
Charlotte	20	20
Chicago	30	29
Cincinnati	20	19
Dallas	20	22
Denver	23	23
Detroit	18	19
Houston	20	16
Minneapolis	18	18
Phoenix	21	22
Pittsburgh	25	22
Salt Lake City	18	17
St. Louis	16	16

- Trace un diagrama de dispersión con estos datos, en el que la variable independiente sean las llegadas retrasadas.
  - ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre llegadas retrasadas y salidas retrasadas?
  - Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
  - ¿Cómo se debe interpretar la pendiente de la ecuación de regresión estimada?
  - Suponga que en el aeropuerto de Filadelfia hubo 22% de llegadas retrasadas. ¿Cuál es el porcentaje estimado de salidas retrasadas?
12. Una moto acuática personal (personal watercraft, PWC) es una embarcación a motor dentro de borda diseñada para ser conducida por una persona sentada, de pie o arrodillada. Al principio de los años 80, Kawasaki Motors Corp. USA introdujo la moto acuática JET SKI®, la primera moto acuática comercial. Hoy *jet ski* se usa como término genérico para motos acuáticas personales. En la tabla siguiente se dan pesos (redondeados a la decena de libra más cercana) y precios (redondeados a los 50 dólares más cercanos) de 10 motos acuáticas personales de tres plazas ([www.jetskinews.com](http://www.jetskinews.com), 2006).

Fabricante y modelo	Peso (lb)	Precio (\$)
Honda AquaTrax F-12	750	9 500
Honda AquaTrax F-12X	790	10 500
Honda AquaTrax F-12X GPScape	800	11 200
Kawasaki STX-12F JetSKI	740	8 500
Yamaha FX Cruiser Waverunner	830	10 000
Yamaha FX High Output Waverunner	770	10 000
Yamaha FX Waverunner	830	9 300
Yamaha VX110 Deluxe Waverunner	720	7 700
Yamaha VX110 Sport Waverunner	720	7 000
Yamaha XLT1200 Waverunner	780	8 500

- Trace el diagrama de dispersión correspondiente a estos datos, empleando el peso como variable independiente.
- ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre peso y precio?
- Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
- Indique cuál será el precio de una moto acuática de tres plazas cuyo peso sea 750 libras.
- La Honda Aqua Trax F-12 pesa 750 libras y su precio es \$9500. ¿No debería ser el precio pronosticado en el inciso d) también de \$950?

- f. La JetSKI Kawasaki SX-R 800 tiene asiento para una persona y pesa 350 libras. ¿Cree usted que la ecuación de regresión estimada obtenida en el inciso c) deba emplearse para predecir su precio?
13. Para la Dirección general de impuestos internos de Estados Unidos el que las deducciones parezcan razonables depende del ingreso bruto ajustado del contribuyente. Deducciones grandes que comprenden deducciones por donaciones de caridad o por atención médica son más probables en contribuyentes que tengan un ingreso bruto ajustado grande. Si las deducciones de un contribuyente son mayores que las correspondientes a un determinado nivel de ingresos, aumentan las posibilidades de que se le realice una auditoría.

Ingreso bruto ajustado (miles de \$)	Monto razonable de las deducciones (miles de \$)
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

- a. Trace un diagrama de dispersión con estos datos empleando como variable independiente el ingreso bruto ajustado.
- b. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada
- c. Si el ingreso bruto ajustado de un contribuyente es \$52 500, estime el monto razonable de deducciones. Si el contribuyente tiene deducciones por \$20 400, ¿estarán justificadas una auditoría? Explique.
14. Los salarios iniciales de contadores y auditores en Rochester, Nueva York, corresponden a los de muchos ciudadanos de Estados Unidos. En la tabla siguiente se presentan salarios iniciales (en miles de dólares) y el índice del costo de vida en Rochester y en otras nueve zonas metropolitanas (*Democrat and Chronicle*, 1 de septiembre de 2002).

Área metropolitana	Índice	Salario (miles de \$)
Oklahoma City	82.44	23.9
Tampa/St. Petersburg/Clearwater	79.89	24.5
Indianapolis	55.53	27.4
Buffalo/Niagara Falls	41.36	27.7
Atlanta	39.38	27.1
Rochester	28.05	25.6
Sacramento	25.50	28.7
Raleigh/Durham/Chapel Hill	13.32	26.7
San Diego	3.12	27.8
Honolulu	0.57	28.3

- a. Elabore un diagrama de dispersión con estos datos empleando como variable independiente el índice del costo de vida.
- b. Obtenga la ecuación de regresión para relacionar el índice del costo de vida con el salario inicial.
- c. Estime el salario inicial en una zona metropolitana en la que el índice del costo de vida es 50.

14.3

## Coeficiente de determinación

En el ejemplo de Armand Pizza Parlors para aproximar la relación lineal entre el tamaño de la población de estudiantes  $x$  y las ventas trimestrales  $y$  se obtuvo la ecuación de regresión estimada  $\hat{y} = 60 + 5x$ . Ahora la pregunta es: ¿qué tan bien se ajusta a los datos la ecuación de regresión estimada? En esta sección se muestra que una medida de la bondad de ajuste de la ecuación de regresión estimada (lo bien que se ajusta la ecuación a los datos) es el **coeficiente de determinación**.

A la diferencia que existe, en la observación  $i$ , entre el valor observado de la variable dependiente  $y_i$ , y el valor estimado de la variable dependiente  $\hat{y}_i$ , se le llama **residual  $i$** . El residual  $i$  representa el error que existe al usar  $\hat{y}_i$  para estimar  $y_i$ . Por lo tanto, para la observación  $i$ , el residual es  $y_i - \hat{y}_i$ . La suma de los cuadrados de estos residuales o errores es la cantidad que se minimiza empleando el método de los mínimos cuadrados. Esta cantidad, también conocida como *suma de cuadrados debida al error*, se denota por SCE.

### SUMA DE CUADRADOS DEBIDA AL ERROR

$$\text{SCE} = \sum (y_i - \hat{y}_i)^2 \quad (14.8)$$

El valor de SCE es una medida del error al utilizar la ecuación de regresión estimada para estimar los valores de la variable dependiente en los elementos de la muestra.

En la tabla 14.3 se muestran los cálculos que se requieren para calcular la suma de cuadrados debida al error en el ejemplo de Armand's Pizza Parlors. Por ejemplo, los valores de las variables independiente y dependiente para/del restaurante 1 son  $x_1 = 2$  y  $y_1 = 58$ . El valor estimado para las ventas trimestrales del restaurante 1 obtenido con la ecuación de regresión estimada es  $\hat{y}_1 = 60 + 5(2) = 70$ . Por lo tanto, para el restaurante 1, el error al usar  $\hat{y}_1$  para estimar  $y_1$  es  $y_1 - \hat{y}_1 = 58 - 70 = -12$ . El error elevado al cuadrado,  $(-12)^2 = 144$ , aparece en la última columna de la tabla 14.3. Despues de calcular y elevar al cuadrado los residuales de cada uno de los restaurantes de la muestra, se suman y se obtiene que SCE = 1530. Por lo tanto, SCE = 1530 mide el error que existe al utilizar la ecuación de regresión estimada  $\hat{y} = 60 + 5x$  para predecir las ventas.

Ahora supóngase que se pide una estimación de las ventas trimestrales sin saber cuál es el tamaño de la población de estudiantes. Sin tener conocimiento de ninguna otra variable relacionada con las ventas trimestrales, se emplearía la media muestral como una estimación de las ven-

**TABLA 14.3** CÁLCULO DE SCE EN EL EJEMPLO ARMAND'S PIZZA PARLORS

Restaurante $i$	$x_i$ = población de estudiantes (miles)	$y_i$ = ventas trimestrales (miles de \$)	Ventas pronosticadas $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Error al cuadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	-12	144
2	6	105	90	15	225
3	8	88	100	-12	144
4	8	118	100	18	324
5	12	117	120	-3	9
6	16	137	140	-3	9
7	20	157	160	-3	9
8	20	169	160	9	81
9	22	149	170	-21	441
10	26	202	190	12	144
$\text{SCE} = 1530$					

**TABLA 14.4** CÁLCULO DE LA SUMA TOTAL DE CUADRADOS EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Restaurante <i>i</i>	$x_i =$ población de estudiantes (miles)	$y_i =$ ventas trimestrales (miles de \$)	Desviación $y_i - \bar{y}$	Desviación al cuadrado $(y_i - \bar{y})^2$
1	2	58	-72	5 184
2	6	105	-25	625
3	8	88	-42	1 764
4	8	118	-12	144
5	12	117	-13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				STC = 15 730

tas trimestrales de/en cualquiera de los restaurantes. En la tabla 14.2 se mostró que de acuerdo con los datos de las ventas,  $\sum y_i = 1300$ . Por lo tanto, la media de las ventas trimestrales en la muestra de los 10 restaurantes Armand's es  $\bar{y} = \sum y_i/n = 1300/10 = 130$ . En la tabla 14.4 se presenta la suma de las desviaciones al cuadrado que se obtiene cuando se usa la media muestral  $\bar{y} = 130$  para estimar el valor de las ventas trimestrales de cada uno de los restaurantes de la muestra. Para el *i*-ésimo restaurante de la muestra, la diferencia  $y_i - \bar{y}$  proporciona una medida del error que hay al usar  $\bar{y}$  para estimar las ventas. La correspondiente suma de cuadrados, llamada *suma total de cuadrados*, se denota STC.

#### SUMA TOTAL DE CUADRADOS

$$\text{STC} = \sum (y_i - \bar{y})^2 \quad (14.9)$$

La suma debajo de la última columna de la tabla 14.4 es la suma total de cuadrados en el ejemplo de Armand's Pizza Parlors; esta suma es  $\text{STC} = 15 730$ .

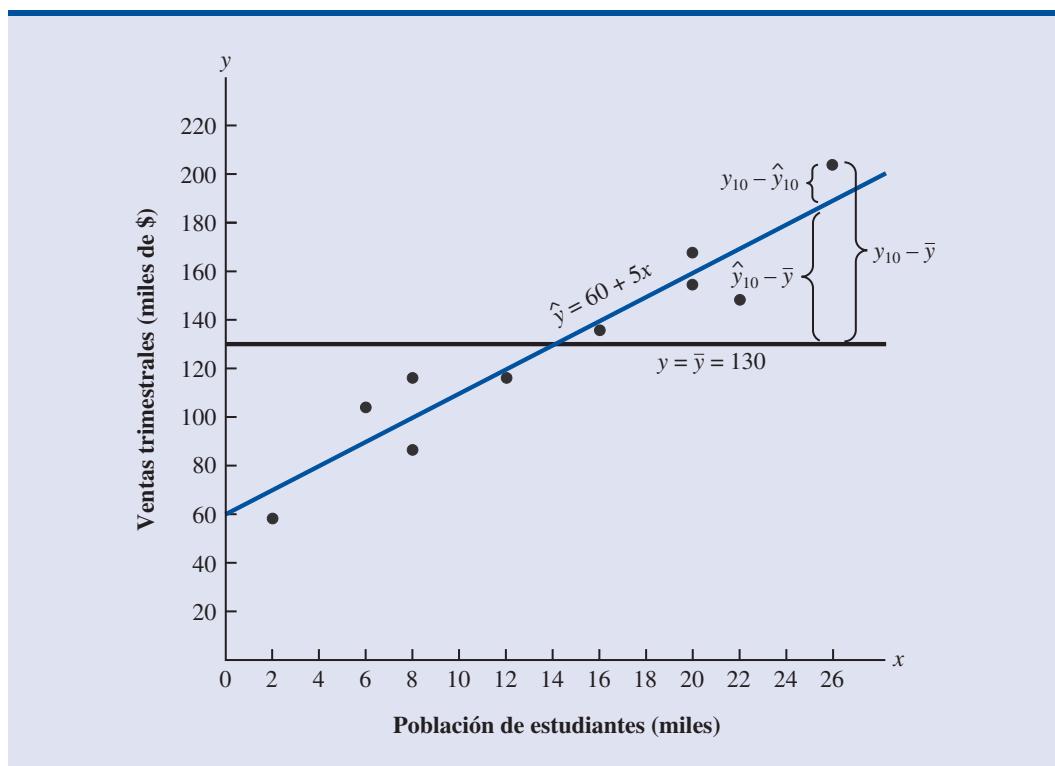
En la figura 14.5 se muestra la línea de regresión estimada  $\hat{y} = 60 + 5x$  y la línea correspondientes a  $\bar{y} = 130$ . Obsérvese que los puntos se encuentran más agrupados en torno a la recta de regresión estimada que en torno a la recta  $\bar{y} = 130$ . Por ejemplo, se ve que para el 10o. restaurante de la muestra, el error es mucho más grande cuando se usa  $\bar{y} = 130$  para estimar  $y_{10}$  que cuando se usa  $\hat{y}_{10} = 60 + 5(26) = 190$ . Se puede entender STC como una medida de qué tanto se agrupan las observaciones en torno a la recta  $\bar{y}$  y SCE como una medida de qué tanto se agrupan las observaciones en torno de la recta  $\hat{y}$ .

Para medir qué tanto se desvían de  $\bar{y}$  los valores  $\hat{y}_i$  de la recta de regresión, se calcula otra suma de cuadrados. A esta suma se le llama *suma de cuadrados debida a la regresión* y se denota SCR.

#### SUMA DE CUADRADOS DEBIDA A LA REGRESIÓN

$$\text{SCR} = \sum (\hat{y}_i - \bar{y})^2 \quad (14.10)$$

**FIGURA 14.5** DESVIACIONES RESPECTO A LA LÍNEA DE REGRESIÓN ESTIMADA Y A LA LÍNEA  $y = \bar{y}$  EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS



Por lo antes dicho, se esperaría que hubiera alguna relación entre STC, SCR y SCE. En efecto, y la relación entre estas tres sumas de cuadrados constituye uno de los resultados más importantes de la estadística.

#### RELACIÓN ENTRE STC, SCR Y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \quad (14.11)$$

donde

STC = suma total de cuadrados

SCR = suma de cuadrados debida a la regresión

SCE = suma de cuadrados debida al error

*La SCR puede entenderse como la parte explicada de la STC, y la SCE puede entenderse como la parte no explicada de la STC.*

La ecuación (14.11) muestra que la suma total de cuadrados puede ser dividida en dos componentes, la suma de los cuadrados debida a la regresión y la suma de cuadrados debida al error. Por lo tanto, si se conocen dos cualesquiera de estas sumas de cuadrados, es fácil calcular la tercera suma de cuadrados. Por ejemplo, en el ejemplo de Armand's Pizza Parlors, se conocen SCE = 1530 y STC 15 730; por lo tanto, despejando de la ecuación (14.11) SCR, se encuentra que la suma de los cuadrados debidos a la regresión es

$$\text{SCR} = \text{STC} + \text{SCE} = 15\,730 - 1530 = 14\,200$$

Ahora se verá cómo se usan estas tres sumas de cuadrados, STC, SCR y SCE, para obtener una medida de la bondad de ajuste de la ecuación de regresión estimada. La ecuación de regresión estimada se ajustaría perfectamente a los datos si cada uno de los valores de la variable independiente  $y_i$  se encontraran sobre la recta de regresión. En este caso para todas las observaciones se tendría que  $y_i - \hat{y}_i$  sería igual a cero, con lo que  $SCE = 0$ . Como  $STC = SCR + SCE$  se ve que para que haya un ajuste perfecto  $SCR$  debe ser igual a  $STC$ , y el cociente ( $SCR/STC$ ) debe ser igual a uno. Cuando los ajustes son malos, se tendrán valores altos para  $SCE$ . Si en la ecuación (14.11) se despeja  $SCE$ , se tiene que  $SCE = STC - SCR$ . Por lo tanto, los valores más grandes de  $SCE$  (y por lo tanto un peor ajuste) se presentan cuando  $SCR = 0$  y  $SCE = STC$ .

El cociente  $SCR/STC$ , que toma valores entre cero y uno, se usa para evaluar la bondad de ajuste de la ecuación de regresión estimada. A este cociente se le llama *coeficiente de determinación* y se denota  $r^2$ .

#### COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{SCR}{STC} \quad (14.12)$$

En el ejemplo de Armand's Pizza Parlors, el valor del coeficiente de determinación es

$$r^2 = \frac{SCR}{STC} = \frac{14\,200}{15\,730} = 0.9027$$

Si se expresa el coeficiente de determinación en forma de porcentaje,  $r^2$  se puede interpretar como el porcentaje de la suma total de cuadrados que se explica mediante el uso de la ecuación de regresión estimada. En el ejemplo de Armand's Pizza Parlors, se concluye que 90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre el tamaño de la población de estudiantes y las ventas. Sería bueno que la ecuación de regresión tuviera un ajuste tan bueno.

#### Coeficiente de correlación

En el capítulo 3 se presentó el **coeficiente de correlación** como una medida descriptiva de la intensidad de la relación lineal entre dos variables  $x$  y  $y$ . Los valores del coeficiente de correlación son valores que van desde  $-1$  hasta  $+1$ . El valor  $+1$  indica que las dos variables  $x$  y  $y$  están perfectamente relacionadas en una relación lineal positiva. Es decir, los puntos de todos los datos se encuentran en una línea recta que tiene pendiente positiva. El valor  $-1$  indica que  $x$  y  $y$  están perfectamente relacionadas, en una relación lineal negativa, todos los datos se encuentran en una línea recta que tiene pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que  $x$  y  $y$  no están relacionadas linealmente.

En la sección 3.5 se presentó la ecuación para calcular el coeficiente de correlación muestral. Cuando se ha realizado un análisis de regresión y se ha calculado el coeficiente de determinación  $r^2$ , el coeficiente de correlación muestral se puede calcular como se indica a continuación.

#### COEFICIENTE DE CORRELACIÓN MUESTRAL

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1) \sqrt{\text{Coeficiente de determinación}} \\ &= (\text{signo de } b_1) \sqrt{r^2} \end{aligned} \quad (14.13)$$

donde

$$b_1 = \text{pendiente de la ecuación de regresión estimada } \hat{y} = b_0 + b_1x$$

El signo del coeficiente de regresión muestral es positivo si la ecuación de regresión tiene pendiente positiva ( $b_1 > 0$ ) y es negativo si la ecuación de regresión estimada tiene pendiente negativa ( $b_1 < 0$ ).

En el ejemplo de Armand's Pizza Parlor, el valor del coeficiente de determinación correspondiente a la ecuación de regresión estimada  $\hat{y} = 60 + 5x$  es 0.9027. Como la pendiente de la ecuación de regresión estimada es positiva, la ecuación (14.13) indica que el coeficiente de correlación muestral es  $+\sqrt{0.9027} = +0.9501$ . Con este coeficiente de correlación muestral,  $r_{xy} = +0.9501$ , se concluye que existe una relación lineal fuerte entre  $x$  y  $y$ .

En el caso de una relación lineal entre dos variables, tanto el coeficiente de determinación como el coeficiente de correlación muestral proporcionan medidas de la intensidad de la relación. El coeficiente de determinación proporciona una medida cuyo valor va desde cero hasta uno, mientras que el coeficiente de correlación muestral proporciona una medida cuyo valor va desde  $-1$  hasta  $+1$ . El coeficiente de correlación lineal está restringido a la relación lineal entre dos variables, pero el coeficiente de determinación puede emplearse para relaciones no lineales y para relaciones en las que hay dos o más variables independientes. Por tanto, el coeficiente de determinación tiene un rango más amplio de aplicaciones.

## NOTAS Y COMENTARIOS

- Al obtener la ecuación de regresión estimada mediante el método de mínimos cuadrados y calcular el coeficiente de determinación, no se hizo ninguna suposición probabilística acerca del término del error  $\epsilon$  ni tampoco una prueba de significancia para la relación entre  $x$  y  $y$ . Los valores grandes de  $r^2$  implican que la recta de mínimos cuadrados se ajusta mejor a los datos; es decir, las observaciones se encuentran más cerca de la recta de mínimos cuadrados. Sin embargo, usando únicamente  $r^2$  no se pueden sacar conclusiones acerca de si la relación entre  $x$  y  $y$  es estadísticamente significativa. Tal conclusión debe basarse en consideraciones que

implican el tamaño de la muestra y las propiedades de la distribución muestral adecuada de los estimadores de mínimos cuadrados.

- Para fines prácticos, cuando se trata de datos que se encuentran en las ciencias sociales, valores de  $r^2$  tan pequeños como 0.25 suelen considerarse útiles. En datos de la física o de las ciencias de la vida, suelen encontrarse valores de  $r^2$  de 0.60 o mayores; en algunos casos pueden encontrarse valores mayores de 0.90. En las aplicaciones a los negocios, los valores de  $r^2$  varían enormemente dependiendo de las características particulares de cada aplicación.

## Ejercicios

### Método

- Los datos a continuación son los datos del ejercicio 1.

$x_i$	1	2	3	4	5
$y_i$	3	7	5	11	14

La ecuación de regresión estimada para estos datos es  $\hat{y} = 0.20 + 2.60x$ .

- Empleando las ecuaciones (14.8), (14.9) y (14.10) calcule SCE, STC y SCR.
- Calcule el coeficiente de determinación  $r^2$ . Haga un comentario sobre la bondad del ajuste.
- Calcule el coeficiente de correlación muestral.

16. Los datos a continuación son los datos del ejercicio 2.

$x_i$	3	12	6	20	14
$y_i$	55	40	55	10	15

La ecuación de regresión estimada para estos datos es  $\hat{y} = 68 - 3x$ .

- a. Calcule SCE, STC y SCR.
  - b. Calcule el coeficiente de determinación  $r^2$ . Haga un comentario sobre la bondad del ajuste.
  - c. Calcule el coeficiente de correlación muestral.
17. Los datos a continuación son los datos del ejercicio 3.

$x_i$	2	6	9	13	20
$y_i$	7	18	9	26	23

La ecuación de regresión estimada para estos datos es  $\hat{y} = 7.6 + 0.9x$ . ¿Qué porcentaje de la suma total de cuadrados puede explicarse mediante la ecuación de regresión estimada? ¿Cuál es el valor del coeficiente de correlación muestral?

## Aplicaciones



18. En los datos siguientes,  $y$  corresponde a los salarios mensuales y  $x$  es el promedio obtenido por los estudiantes que terminaron la licenciatura de administración con especialidad en sistemas de información. La ecuación de regresión obtenida con estos datos es  $\hat{y} = 1790.5 + 581.1x$ .

Promedio	Salario mensual (\$)
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

- a. Calcule SCE, STC y SCR.
  - b. Calcule el coeficiente de determinación  $r^2$ . Haga un comentario sobre la bondad del ajuste.
  - c. Calcule el coeficiente de correlación muestral.
19. Los datos a continuación son los datos del ejercicio 7.



Cars

Fabricante y modelo	$x = \text{confiabilidad}$	$y = \text{precio} (\$)$
Acura TL	4	33 150
BMW 330i	3	40 570
Lexus IS300	5	35 105
Lexus ES330	5	35 174
Mercedes-Benz C320	1	42 230
Lincoln LS Premium (V6)	3	38 225
Audi A4 3.0 Quattro	2	37 605
Cadillac CTS	1	37 695
Nissan Maxima 3.5 SE	4	34 390
Infiniti I35	5	33 845
Saab 9-3 Aero	3	36 910
Infiniti G35	4	34 695
Jaguar X-Type 3.0	1	37 995
Saab 9-5 Arc	3	36 955
Volvo S60 2.5T	3	33 890

- La ecuación de regresión estimada para estos datos es  $\hat{y} = 40\,639 - 1301.2x$ . ¿Qué porcentaje de la suma total de cuadrados puede explicarse mediante la ecuación de regresión estimada? Haga un comentario sobre la bondad del ajuste. ¿Cuál es el valor del coeficiente de correlación muestral?
20. *Consumer Reports* publica pruebas y evaluaciones sobre televisores de alta definición. Para cada modelo se elaboró una evaluación general basada principalmente en la calidad de la imagen. Una evaluación más alta indica un mejor funcionamiento. En los datos siguientes se dan evaluación general y precio de televisores de plasma de 45 pulgadas (*Consumer Reports*, marzo 2006).



Marca	Precio	Puntuación en la valuación
Dell	2800	62
Hisense	2800	53
Hitachi	2700	44
JVC	3500	50
LG	3300	54
Maxent	2000	39
Panasonic	4000	66
Phillips	3000	55
Proview	2500	34
Samsung	3000	39

- a. Use estos datos para obtener una ecuación de regresión estimada que pueda emplearse para estimar la puntuación en la evaluación general de una televisión de 42 pulgadas dado el precio.
- b. Calcule  $r^2$ . ¿Proporcionó un buen ajuste la ecuación de regresión estimada?
- c. Estime la puntuación en la evaluación general de un televisor cuyo precio es \$3200.
21. Una aplicación importante del análisis de regresión a la contaduría es la estimación de costos. Con datos sobre volumen de producción y costos y empleando el método de mínimos cuadrados para obtener la ecuación de regresión estimada que relacione volumen de producción y costos, los contadores pueden estimar los costos correspondientes a un determinado volumen de producción. Considere la siguiente muestra de datos sobre volumen de producción y costos totales de una operación de fabricación.

Volumen de producción (unidades)	Costos totales (\$)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

- a. Con estos datos obtenga la ecuación de regresión estimada para pronosticar los costos totales dado un volumen de producción.
- b. ¿Cuál es el costo por unidad producida?
- c. Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación en los costos totales puede ser explicada por el volumen de producción?
- d. De acuerdo con el programa de producción de la empresa, el mes próximo se deberán producir 500 unidades. ¿Cuál es el costo total estimado de esta operación?
22. *PC World* publicó evaluaciones de las cinco mejores impresoras láser de oficina y de las cinco mejores impresoras láser corporativas (*PC World*, febrero 2003). La impresora de oficina mejor evaluada fue la Minolta-QMS PagePro 1250W, que en la evaluación general obtuvo una puntuación de 91 puntos. La impresora láser corporativa mejor evaluada fue la Xerox Phaser 4400/N, que



en la evaluación general obtuvo una puntuación de 83 puntos. En la tabla siguiente se da rapidez, en páginas por minuto (ppm), en la impresión de texto y precio de cada impresora.

Nombre	Tipo	Velocidad (ppm)	Precio (\$)
Minolta-QMS PagePro 1250W	Oficina	12	199
Brother HL-1850	Oficina	10	499
Lexmark E320	Oficina	12.2	299
Minolta-QMS PagePro 1250E	Oficina	10.3	299
HP Laserjet 1200	Oficina	11.7	399
Xerox Phaser 4400/N	Corporativa	17.8	1850
Brother HL-2460N	Corporativa	16.1	1000
IBM Infoprint 1120n	Corporativa	11.8	1387
Lexmark W812	Corporativa	19.8	2089
Oki Data B8300n	Corporativa	28.2	2200

- Dé la ecuación de regresión estimada empleando velocidad como variable independiente.
- Calcule  $r^2$ . ¿Qué porcentaje de la variación del precio puede ser explicado por la velocidad de la impresora?
- ¿Cuál es el coeficiente de correlación muestral entre velocidad y precio? ¿Refleja este coeficiente una relación fuerte o débil entre la velocidad de la impresora y el costo?

## 14.4

## Suposiciones del modelo

En un análisis de regresión se empieza por hacer una suposición acerca del modelo apropiado para la relación entre las variables dependientes e independientes. En el caso de la regresión lineal simple, se supone que el modelo de regresión es

$$y = \beta_0 + \beta_1 x + \epsilon$$

Después empleando el método de mínimos cuadrados se obtienen los valores de  $b_0$  y  $b_1$ , que son las estimaciones de los parámetros  $\beta_0$  y  $\beta_1$ , respectivamente, del modelo. Así se llega la ecuación de regresión estimada

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

Como se vio, el valor del coeficiente de determinación ( $r^2$ ) es una medida de la bondad de ajuste de la ecuación de regresión estimada. Sin embargo, aun cuando se obtenga un valor grande para  $r^2$ , la ecuación de regresión estimada no debe ser usada hasta que se realice un análisis para determinar si el modelo empleado es adecuado. Un paso importante para ver si el modelo empleado es adecuado es probar la significancia de la relación. Las pruebas de significancia en el análisis de regresión están basadas en las suposiciones siguientes acerca del término del error  $\epsilon$ .

### SUPOSICIONES ACERCA DEL TÉRMINO DEL ERROR EN EL ANÁLISIS DE REGRESIÓN

$$y = \beta_0 + \beta_1 x + \epsilon$$

- El término del error  $\epsilon$  es una variable aleatoria cuya media, o valor esperado, es cero; es decir,  $E(\epsilon) = 0$ .

*Implicación:*  $\beta_0$  y  $\beta_1$  son constantes, por lo tanto  $E(\beta_0) = \beta_0$  y  $E(\beta_1) = \beta_1$ ; así, para un valor dado de  $x$ , el valor esperado de  $y$  es

$$E(y) = \beta_0 + \beta_1 x \quad (14.14)$$

(continúa)

Como ya se indicó, a la ecuación (14.14) se le conoce como ecuación de regresión.

2. La varianza de  $\epsilon$ , que se denota  $\sigma^2$ , es la misma para todos los valores de  $x$ .

*Implicación:* La varianza de  $y$  respecto a la recta de regresión es igual a  $\sigma^2$  y es la misma para todos los valores de  $x$ .

3. Los valores de  $\epsilon$  son independientes.

*Implicación:* El valor de  $\epsilon$  correspondiente a un determinado valor de  $x$  no está relacionado con el valor de  $\epsilon$  correspondiente a ningún otro valor de  $x$ ; por lo tanto, el valor de  $y$  correspondiente a un determinado valor de  $x$  no está relacionado con el valor de  $y$  de ningún otro valor de  $x$ .

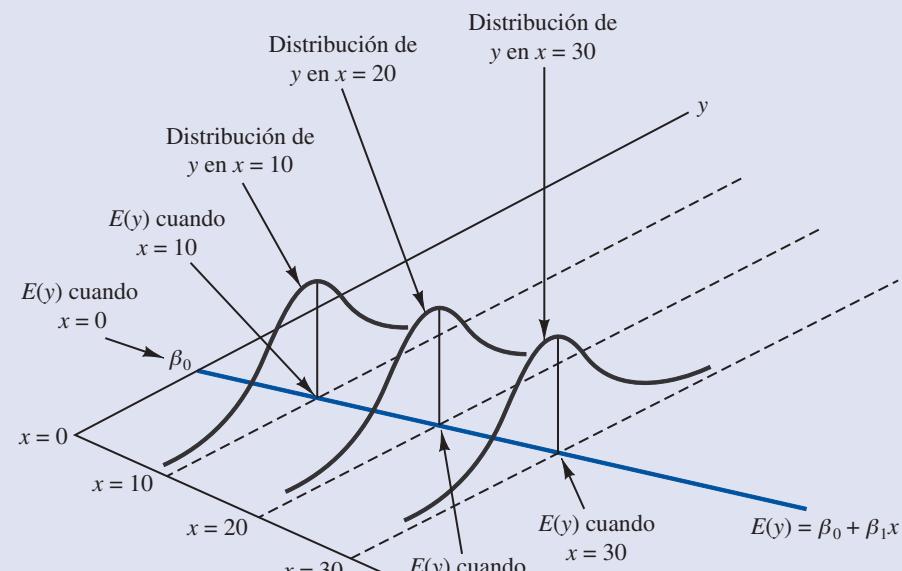
4. El término del error  $\epsilon$  es una variable aleatoria distribuida normalmente.

*Implicación:* como  $y$  es función lineal de  $\epsilon$ , también  $y$  es una variable aleatoria distribuida normalmente.

En la figura 14.6 se muestran las suposiciones del modelo y sus implicaciones; obsérvese que en esta interpretación gráfica, el valor de  $E(y)$  cambia de acuerdo con el valor de  $x$  que se considere. Sin embargo, sea cual sea el valor de  $x$ , la distribución de probabilidad de  $\epsilon$ , y por tanto la distribución de probabilidad de  $y$ , son distribuciones normales, que tienen, todas, la misma varianza. El valor específico del error  $\epsilon$  en cualquier punto depende de si el valor real de  $y$  es mayor o menor que  $E(y)$ .

En este punto, hay que tener presente que también se hace una suposición o se tienen una hipótesis acerca de la forma de la relación entre  $x$  y  $y$ . Es decir, se supone que la base de la relación entre las variables es una recta representada por  $\beta_0 + \beta_1x$ . No se debe perder de vista el

**FIGURA 14.6 SUPOSICIONES DEL MODELO DE REGRESIÓN**



*Nota:* Para cada uno de los valores de  $x$  las distribuciones  $y$  tienen la misma forma.

hecho de que puede haber algún otro modelo, por ejemplo  $y = \beta_0 + \beta_1x^2 + \epsilon$ , que resulte ser un mejor modelo para la relación en estudio.

14.5

## Prueba de significancia

En una ecuación de regresión lineal simple, la media o valor esperado de  $y$  es una función lineal de  $x$ :  $E(y) = \beta_0 + \beta_1x$ . Pero si el valor de  $\beta_1$  es cero,  $E(y) = \beta_0 + (0)x = \beta_0$ . En este caso, el valor medio de  $y$  no depende del valor de  $x$  y por lo tanto se puede concluir que  $x$  y  $y$  no están relacionadas linealmente. Pero si el valor de  $\beta_1$  es distinto de cero, se concluirá que las dos variables están relacionadas. Por lo tanto, para probar si existe una relación de regresión significante, se debe realizar una prueba de hipótesis para determinar si el valor de  $\beta_1$  es distinto de cero. Hay dos pruebas que son las más usadas. En ambas, se requiere una estimación de  $\sigma^2$ , la varianza de  $\epsilon$  en el modelo de regresión.

### Estimación de $\sigma^2$

De acuerdo con el modelo de regresión y con sus suposiciones, se puede concluir que  $\sigma^2$ , la varianza de  $\epsilon$ , representa también la varianza de los valores de  $y$  respecto a la recta de regresión. Recuérdese que a las desviaciones de los valores de  $y$  de la recta de regresión estimada se les conoce como residuales. Por lo tanto, SCE, la suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada. El **error cuadrado medio** (ECM) proporciona una estimación de  $\sigma^2$ ; esta estimación es SCE dividida entre sus grados de libertad.

Como  $\hat{y}_i = b_0 + b_1x_i$ , SCE se puede expresar como

$$\text{SCE} = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 - b_1x_i)^2$$

A cada suma de cuadrados le corresponde un número llamado sus grados de libertad. Se ha demostrado que SCE tiene  $n - 2$  grados de libertad porque para calcular SCE es necesario estimar dos parámetros ( $\beta_0$  y  $\beta_1$ ). Por lo tanto, el cuadrado medio se calcula dividiendo SCE entre  $n - 2$ . ECM proporciona un estimador insesgado de  $\sigma^2$ . Como el valor del ECM proporciona un estimado de  $\sigma^2$ , se emplea también la notación  $s^2$ .

#### ERROR CUADRADO MEDIO (ESTIMACIÓN DE $\sigma^2$ )

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2} \quad (14.15)$$

En la sección 14.3 se encontró que en el ejemplo de Armand's Pizza Parlors, SCE = 1530; por lo tanto,

$$s^2 = \text{ECM} = \frac{1530}{8} = 191.25$$

es un estimador insesgado de  $\sigma^2$ .

Para estimar  $\sigma$  se saca la raíz cuadrada de  $s^2$ . Al valor que se obtiene,  $s$ , se le conoce como el **error estándar de estimación**.

#### ERROR ESTÁNDAR DE ESTIMACIÓN

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n - 2}} \quad (14.16)$$

En el ejemplo de Armand's Pizza Parlors,  $s = \sqrt{\text{ECM}} = \sqrt{191.25} = 13.829$ . El error estándar de estimación se emplea en la discusión siguiente acerca de las pruebas de significancia de la relación entre  $x$  y  $y$ .

## Prueba $t$

El modelo de regresión lineal simple es  $y = \beta_0 + \beta_1 x + \epsilon$ . Si  $x$  y  $y$  están relacionadas linealmente, entonces  $\beta_1 \neq 0$ . El objetivo de la prueba  $t$  es determinar si se puede concluir que  $\beta_1 \neq 0$ . Para probar la hipótesis siguiente acerca del parámetro  $\beta_1$  se emplearán los datos muestrales.

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

Si se rechaza  $H_0$ , se concluirá que  $\beta_1 \neq 0$  y que entre las dos variables existe una relación estadísticamente significante. La base para esta prueba de hipótesis la proporcionan las propiedades de la distribución muestral de  $b_1$ , el estimador de  $\beta_1$ , obtenido mediante el método de mínimos cuadrados.

Primero, considérese que es lo que ocurriría si para el mismo estudio de regresión se usara otra muestra aleatoria simple. Supóngase, por ejemplo, que Armand's Pizza Parlors usa una muestra de las ventas de otros 10 restaurantes. El análisis de regresión de esta otra muestra dará como resultado una ecuación de regresión parecida a la ecuación de regresión anterior  $\hat{y} = 60 + 5x$ . Sin embargo, no puede esperarse que se obtenga exactamente la misma ecuación (una ecuación en la que la intersección con el eje  $y$  sea exactamente 60 y la pendiente sea exactamente 5). Los estimadores  $b_0$  y  $b_1$ , obtenidos por el método de mínimos cuadrados, son estadísticos muestrales que tienen su propia distribución muestral. A continuación se presentan las propiedades de la distribución muestral de  $b_1$ .

### DISTRIBUCIÓN MUESTRAL DE $b_1$

*Valor esperado*

$$E(b_1) = \beta_1$$

*Desviación estándar*

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} \quad (14.17)$$

*Forma de la distribución*

Normal

Obsérvese que el valor esperado de  $b_1$  es  $\beta_1$ , por lo que  $b_1$  es un estimador insesgado de  $\beta_1$ .

Como no se conoce el valor de  $\sigma$ , se obtiene una estimación de  $\sigma_{b_1}$ , que se denota  $s_{b_1}$ , estimando  $\sigma$  mediante  $s$  en la ecuación (14.17). De esta manera se obtiene el estimador siguiente de  $\sigma_{b_1}$ .

A la desviación estándar de  $b_1$ , se le conoce también como error estándar de  $b_1$ . Por lo tanto,  $s_{b_1}$  proporciona una estimación del error estándar de  $b_1$ .

### DESVIACIÓN ESTÁNDAR ESTIMADA DE $b_1$

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \quad (14.18)$$

En el ejemplo de Armand's Pizza Parlors,  $s = 13.829$ . Por lo tanto, dado que  $\sum(x_i - \bar{x})^2 = 568$  como se muestra en la tabla 14.2, se tiene que

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.5803$$

es la desviación estándar estimada de  $b_1$ .

La prueba  $t$  para determinar si la relación es significativa se basa en el hecho de que el estadístico de prueba

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

sigue una distribución  $t$  con  $n - 2$  grados de libertad. Si la hipótesis nula es verdadera, entonces  $\beta_1 = 0$  y  $t = b_1/s_{b_1}$ .

Ahora se realizará esta prueba de significancia con los datos de Armand's Pizza Parlors, empleando como nivel de significancia  $\alpha = 0.01$ . El estadístico de prueba es

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

*En los apéndices 14.3 y 14.4 se muestra el uso de Minitab y de Excel para calcular el valor- $p$*

En las tablas de la distribución  $t$  se encuentra que para  $n - 2 = 10 - 2 = 8$  grados de libertad,  $t = 3.355$  da un área de 0.005 en la cola superior. Por lo tanto, el área en la cola superior de la distribución  $t$  correspondiente al valor del estadístico de prueba  $t = 8.62$  debe ser menor a 0.005. Como esta prueba es una prueba de dos colas, este valor se duplica y se concluye que el valor- $p$  para  $t = 8.62$  debe ser menor a  $2(0.005) = 0.01$ . Empleando Excel o Minitab se encuentra valor- $p = 0.000$ . Dado que el valor- $p$  es menor a  $\alpha = 0.01$  se rechaza  $H_0$  y se concluye que  $\beta_1$  no es igual a cero. Esto es suficiente evidencia para concluir que existe una relación significativa entre la población de estudiantes y las ventas trimestrales. A continuación se presenta un resumen de la prueba  $t$  de significancia para la regresión lineal simple.

#### PRUEBA $t$ DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL SIMPLE

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

#### ESTADÍSTICO DE PRUEBA

$$t = \frac{b_1}{s_{b_1}} \tag{14.19}$$

#### REGLA DE RECHAZO

- Método del valor- $p$ : Rechazar  $H_0$  si valor- $p \leq \alpha$   
 Método del valor crítico: Rechazar  $H_0$  si  $t \leq -t_{\alpha/2}$  o si  $t \geq t_{\alpha/2}$

donde  $t_{\alpha/2}$  se toma de la distribución  $t$  con  $n - 2$  grados de libertad.

### Intervalo de confianza para $\beta_1$

La fórmula para un intervalo de confianza para  $\beta_1$  es la siguiente:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

El estimador puntual es  $b_1$  y el margen de error es  $t_{\alpha/2}s_{b_1}$ . El coeficiente de confianza para este intervalo es  $1 - \alpha$  y  $t_{\alpha/2}$  es el valor  $t$  que proporciona un área  $\alpha/2$  en la cola superior de la distribución  $t$  con  $n - 2$  grados de libertad. Supóngase, por ejemplo, que en el caso de Armand's Pizza Parlors se desea obtener una estimación de  $\beta_1$  mediante un intervalo de 99% de confianza. En la tabla 2 del apéndice B se encuentra que el valor  $t$  correspondiente a  $\alpha = 0.01$  y  $n - 2 = 10 - 2 = 8$  grados de libertad es  $t_{0.005} = 3.355$ . Por lo tanto, la estimación mediante un intervalo de 99% de confianza es

$$b_1 \pm t_{\alpha/2}s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

o el intervalo que va de 3.05 a 6.95.

Al emplear la prueba  $t$  de significancia la hipótesis probada fue

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

Empleando  $\alpha = 0.01$  como nivel de significancia, se puede usar el intervalo de 99% de confianza como alternativa para llegar a la conclusión de la prueba de hipótesis que se obtiene con los datos de Armand's. Como 0, que es el valor hipotético de  $\beta_1$ , no está comprendido en el intervalo de confianza (3.05 a 6.95), se rechaza  $H_0$  y se concluye que entre el tamaño de la población de estudiantes y las ventas trimestrales sí existe una relación estadísticamente significativa. En general, se puede usar un intervalo de confianza para probar cualquier hipótesis de dos colas acerca de  $\beta_1$ . Si el valor hipotético de  $\beta_1$  está contenido en el intervalo de confianza, no se rechaza  $H_0$ . De lo contrario, se rechaza  $H_0$ .

## Prueba $F$

Una prueba  $F$ , basada en la distribución de probabilidad  $F$  puede emplearse también para probar la significancia en la regresión. Cuando sólo se tiene una variable independiente, la prueba  $F$  lleva a la misma conclusión que la prueba  $t$ ; es decir, si la prueba  $t$  indica que  $\beta_1 \neq 0$  y por lo tanto que existe una relación significante, la prueba  $F$  también indicará que existe una relación significante. Pero cuando hay más de una variable independiente, sólo la prueba  $F$  puede usarse para probar que existe una relación significante general.

La lógica detrás del uso de la prueba  $F$  para determinar si la relación de regresión es estadísticamente significativa se basa en la obtención de dos estimaciones independiente de  $\sigma^2$ . Se explicó cómo ECM proporciona una estimación de  $\sigma^2$ . Si la hipótesis nula  $H_0: \beta_1 = 0$  es verdadera, la suma de cuadrados debida a la regresión, SCR, dividida entre sus grados de libertad proporciona otra estimación independiente de  $\sigma^2$ . A esta estimación se le llama el *cuadrado medio debido a la regresión* o simplemente el *cuadrado medio de la regresión*. Y se denota CMR. En general,

$$\text{CMR} = \frac{\text{SCR}}{\text{Grados de libertad de la regresión}}$$

En los modelos que se consideran en este texto, el número de grados de libertad de la regresión es siempre igual al número de variables independientes en el modelo:

$$\text{CMR} = \frac{\text{SCR}}{\text{Número de variables independientes}} \tag{14.20}$$

Como en este capítulo sólo se consideran modelos de regresión con una sola variable independiente, se tiene  $\text{CMR} = \text{SCR}/1 = \text{SCR}$ . Por lo tanto, en el ejemplo de Armand's Pizza Parlors,  $\text{CMR} = \text{SCR} = 14.200$ .

Si la hipótesis nula es verdadera ( $H_0: \beta_1 = 0$ ), CMR y ECM son dos estimaciones independientes de  $\sigma^2$  y la distribución muestral de CMR/ECM sigue una distribución  $F$  en la que el nú-

mero de grados de libertad en el numerador es igual a uno y el número de grados de libertad en el denominador es igual a  $n - 2$ . Por lo tanto, si  $\beta_1 = 0$  el valor de CMR/ECM deberá ser un valor cercano a uno. Pero, si la hipótesis nula es falsa, ( $\beta_1 \neq 0$ ), CMR sobreestimará  $\sigma^2$  y el valor de CMR/ECM se inflará; de esta manera valores grandes de CMR/ECM conducirán al rechazo de  $H_0$  y a la conclusión de que la relación entre  $x$  y  $y$  es estadísticamente significante.

A continuación se realizará la prueba  $F$  en el ejemplo de Armand's Pizza Parlors. El estadístico de prueba es

$$F = \frac{\text{CMR}}{\text{ECM}} = \frac{14\,200}{191.25} = 74.25$$

*En la regresión lineal simple, la prueba  $F$  y la prueba  $t$  proporcionan resultados idénticos.*

En la tabla de la distribución  $F$  (tabla 4 del apéndice B) se observa que con un grado de libertad en el numerador y  $n - 2 = 10 - 2 = 8$  grados de libertad en el denominador,  $F = 11.26$  proporciona un área de 0.01 en la cola superior. Por lo tanto, el área en la cola superior de la distribución  $F$  que corresponde al estadístico de prueba  $F = 74.25$  debe de ser menor a 0.01. Por lo tanto, se concluye que el valor- $p$  debe de ser menor a  $\alpha = 0.01$ . Empleando Excel o Minitab se encuentra que valor- $p = 0.000$ . Como el valor- $p$  es menor a  $\alpha = 0.01$ , se rechaza  $H_0$  y se concluye que entre el tamaño de la población de estudiantes y las ventas trimestrales, existe una relación significante. A continuación se presenta un resumen de la prueba  $F$  de significancia para la regresión lineal simple.

#### PRUEBA $F$ DE SIGNIFICANCIA EN EL CASO DE LA REGRESIÓN LINEAL SIMPLE

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

#### ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{ECM}} \quad (14.21)$$

#### REGLA DE RECHAZO

Método del valor- $p$ : Rechaza  $H_0$  si valor- $p \leq \alpha$

Método del valor crítico: Rechaza  $H_0$  si  $F \geq F_\alpha$

donde  $F_\alpha$  es un valor de la distribución  $F$  con 1 grado de libertad en el numerador y  $n - 2$  grados de libertad en el denominador.

En el capítulo 13 se vio el análisis de varianza (ANOVA) y el uso de la **tabla de ANOVA** para proporcionar una visión resumida de los cálculos que se emplean en el análisis de varianza. Para resumir los cálculos de la prueba  $F$  de significancia para la regresión se emplea una tabla ANOVA similar. En la tabla 14.5 se presenta la forma general de una tabla ANOVA para la regresión lineal simple. En la tabla 14.6 se presenta la tabla ANOVA con los cálculos para la prueba  $F$  del ejemplo de Armand's Pizza Parlors. Regresión, error y total son los rótulos de las tres fuentes de variación, y SCR, SCE y STC las sumas de cuadrados correspondientes que aparecen en la columna 2. En la columna 3 aparecen los grados de libertad 1 para SCR,  $n - 2$  para SCE y  $n - 1$  para STC. Los valores de CMR y ECM aparecen en la columna 4. En la columna 5 aparece el valor de  $F = \text{CMR}/\text{ECM}$ , y en la columna 6 aparece el valor- $p$  que corresponde al valor de  $F$  de la columna 5. Casi todos los resultados proporcionados por computadoras para el análisis de regresión presentan una tabla ANOVA de la prueba  $F$  de significancia.

**TABLA 14.5** FORMA GENERAL DE LA TABLA ANOVA PARA LA REGRESIÓN LINEAL SIMPLE

*En toda tabla para el análisis de varianza, la suma total de cuadrados es la suma de la suma de cuadrados de la regresión más la suma de cuadrados del error; además, el total de los grados de libertad es la suma de los grados de libertad de la regresión más los grados de libertad del error.*

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	SCR	1	$CMR = \frac{SCR}{1}$	$F = \frac{CMR}{CME}$	
Error	SCE	$n - 2$	$CME = \frac{SCE}{n - 2}$		
Total	STC	$n - 1$			

### Algunas advertencias acerca de la interpretación de las pruebas de significancia

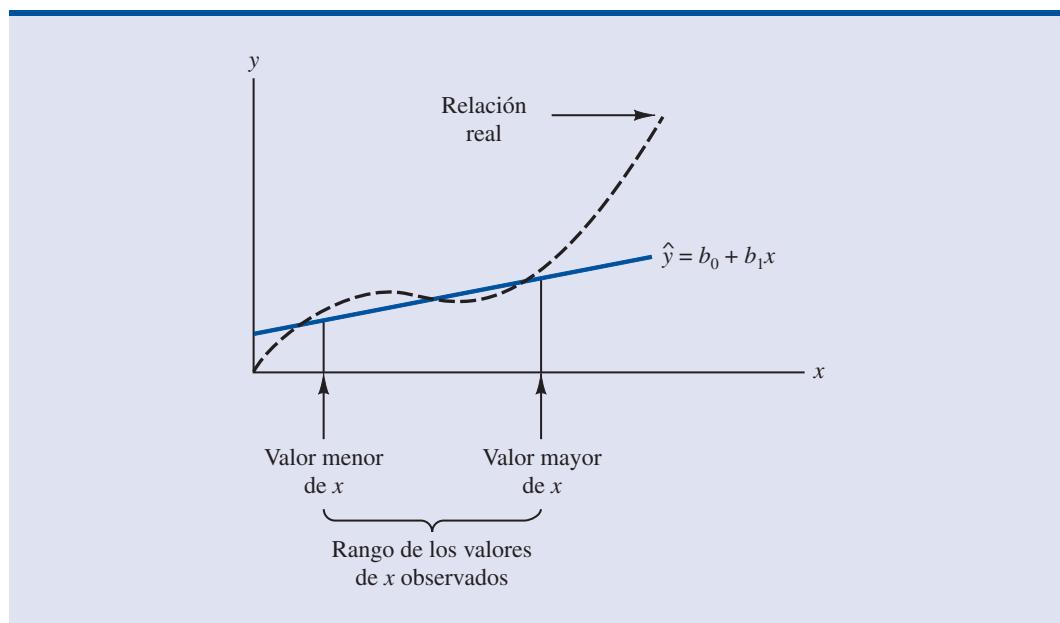
Cuando se rechaza la hipótesis nula  $H_0: \beta_1 = 0$ , concluir que la relación que existe entre  $x$  y  $y$  es significativa no permite que se concluya que existe una relación de causa y efecto entre  $x$  y  $y$ . Que exista una relación de causa y efecto sólo puede concluirse cuando el analista pueda dar justificaciones teóricas de que en efecto la relación es causal. En el ejemplo de Armand's Pizza Parlors, se concluye que existe una relación significante entre el tamaño de la población de estudiantes  $x$  y las ventas trimestrales  $y$ ; aún más, la ecuación de regresión estimada  $\hat{y} = 60 + 5x$  da una estimación de la relación obtenida por el método de mínimos cuadrados. Sin embargo, por el solo hecho de que se haya encontrado que hay una relación estadísticamente significativa entre  $x$  y  $y$ , no se puede concluir que cambios en la población de estudiantes  $x$  causen cambios en las ventas trimestrales  $y$ . Si es posible concluir que haya una relación de causa y efecto se deja a las justificaciones teóricas y a la opinión de los analistas. Los administradores de Armand's creían que el aumento en la población de estudiantes probablemente fuera una causa del aumento de las ventas trimestrales. Por lo tanto, el resultado de la prueba de significancia les permite concluir que hay una relación de causa y efecto.

Además, el hecho de que se pueda rechazar  $H_0: \beta_1 = 0$  y demostrar que hay significancia estadística no permite concluir que la relación entre  $x$  y  $y$  sea lineal. Lo único que se puede decir es que  $x$  y  $y$  están relacionadas y que la relación lineal explica una porción significante de la variabilidad de  $y$  sobre el rango de los valores de  $x$  observados en la muestra. En la figura 14.7 se ilustra esta relación. La prueba de significancia lleva al rechazo de la hipótesis nula  $H_0: \beta_1 = 0$  y a la hipótesis de que  $x$  y  $y$  están significantemente relacionadas, pero en la figura se observa que la verdadera relación entre  $x$  y  $y$  no es lineal. Aunque la aproximación lineal proporcionada

**TABLA 14.6** TABLA ANOVA PARA EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	14 200	1	$\frac{14\ 200}{1} = 14\ 200$	$\frac{14\ 200}{191.25} = 74.25$	0.000
Error	1 530	8	$\frac{1530}{8} = 191.25$		
Total	15 730	9			

**FIGURA 14.7 EJEMPLO DE UNA APROXIMACIÓN LINEAL PARA UNA RELACIÓN QUE NO ES LINEAL**



por  $\hat{y} = b_0 + b_1x$  es buena en el rango de los valores observados de  $x$  en la muestra, se vuelve deficiente fuera de ese rango.

Dada una relación significante, la ecuación de regresión estimada se puede usar con confianza para predicciones correspondientes a valores de  $x$  dentro del rango de los valores de  $x$  observados en la muestra. En el ejemplo de Armand's Pizza Parlors, este rango corresponde a los valores de  $x$  entre 2 y 26. A menos que haya otras razones que indiquen que el modelo es válido más allá de este rango, las predicciones fuera del rango de la variable independiente deben hacerse con cuidado. En el ejemplo de Armand's Pizza Parlors, como se ha encontrado que la relación de regresión es significante al nivel de significancia de 0.01, se puede tener confianza para usar esta relación para predecir las ventas de restaurantes en los que la población de estudiantes correspondiente esté en el intervalo de 2000 a 26 000.

### NOTAS Y COMENTARIOS

1. Las suposiciones hechas acerca del término del error (sección 14.4) son las que permiten las pruebas de significancia estadística de esta sección. Las propiedades de la distribución muestral de  $b_1$  y las subsiguientes pruebas  $t$  y  $F$  siguen directamente de estas suposiciones.
2. No se debe confundir la significancia estadística con la significancia práctica. Con tamaños de muestra muy grandes, se pueden obtener resultados estadísticamente significantes para valores pequeños de  $b_1$ ; en tales casos hay que tener cuidado al concluir que la relación tiene significancia práctica.
3. Una prueba de significancia para la relación lineal entre  $x$  y  $y$  también se puede realizar usando el coeficiente de correlación muestral  $r_{xy}$ .

Empleando  $r_{xy}$  para denotar el coeficiente de correlación poblacional, las hipótesis son las siguientes.

$$\begin{aligned} H_0: \rho_{xy} &= 0 \\ H_a: \rho_{xy} &\neq 0 \end{aligned}$$

Si se rechaza  $H_0$ , se puede concluir que existe una relación significante. En el apéndice 14.2 se proporcionan los detalles de esta prueba. Sin embargo, las pruebas  $t$  y  $F$  presentadas en esta sección dan el mismo resultado que la prueba de significancia usando el coeficiente de correlación. Por lo tanto, si ya se ha realizado una prueba  $t$  o una prueba  $F$  no es necesario realizar una prueba de significancia usando el coeficiente de correlación.

## Ejercicios

### Métodos

23. A continuación se presentan los datos del ejercicio 1.



$x_i$	1	2	3	4	5
$y_i$	3	7	5	11	14

- a. Usando la ecuación (14.15) calcule el error cuadrado medio.
- b. Usando la ecuación (14.16) calcule el error estándar de estimación.
- c. Usando la ecuación (14.18) calcule la desviación estándar estimada de  $b_1$ .
- d. Use la prueba  $t$  para probar las hipótesis siguientes ( $\alpha = 0.05$ )

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

- e. Use la prueba  $F$  para probar las hipótesis del inciso d) empleando como nivel de significancia 0.05. Presente los resultados en el formato de tabla de análisis de varianza.
- 24. A continuación se presentan los datos del ejercicio 2.

$x_i$	3	12	6	20	14
$y_i$	55	40	55	10	15

- a. Usando la ecuación (14.15) calcule el error cuadrado medio.
- b. Usando la ecuación (14.16) calcule el error estándar de estimación.
- c. Usando la ecuación (14.18) calcule la desviación estándar estimada de  $b_1$ .
- d. Use la prueba  $t$  para probar las hipótesis siguientes ( $\alpha = 0.05$ ).

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

- e. Use la prueba  $F$  para probar las hipótesis del inciso d) empleando como nivel de significancia 0.05. Presente los resultados en el formato de tabla de análisis de varianza.
- 25. A continuación se presentan los datos del ejercicio 3.

$x_i$	2	6	9	13	20
$y_i$	7	18	9	26	23

- a. ¿Cuál es el valor del error estándar de estimación?
- b. Pruebe si existe una relación significante usando la prueba  $t$ . Use  $\alpha = 0.05$ .
- c. Emplee la prueba  $F$  para ver si existe una relación significante. Use  $\alpha = 0.05$ . ¿Cuál es la conclusión?

### Aplicaciones



26. En el ejercicio 18 los datos sobre promedio obtenido en la licenciatura y salarios mensuales fueron los siguientes.

Promedio	Salario mensual (\$)	Promedio	Salario mensual (\$)
2.6	3300	3.2	3500
3.4	3600	3.5	3900
3.6	4000	2.9	3600

- a. ¿Indica la prueba  $t$  que haya una relación significante entre promedio y salario mensual?
- b. Pruebe si la relación es significante usando la prueba  $F$ . ¿Cuál es la conclusión? Use  $\alpha = 0.05$ .
- c. Dé la tabla ANOVA.
27. La revista *Outside Magazine* probó 10 modelos de mochilas y botas para excursionismo. En la tabla siguiente se presentan los datos de soporte superior y precio de cada modelo. El soporte superior se midió con una escala del 1 al 5 en la que 1 significa aceptable y 5 denota excelente soporte superior (*Outside Magazine Buyer's Guide 2001*).



Fabricante y modelo	Soporte superior	Precio (\$)
Salomon Super Raid	2	120
Merrell Chameleon Prime	3	125
Teva Challenger	3	130
Vasque Fusion GTX	3	135
Boreal Maigmo	3	150
L.L. Bean GTX Super Guide	5	189
Lowa Kibo	5	190
Asolo AFX 520 GTX	4	195
Raichle Mt. Trail GTX	4	200
Scarpa Delta SL M3	5	220

- a. Use estos datos para obtener la ecuación de regresión estimada para estimar el precio de las mochilas y las botas para excursionismo con base en el soporte superior.
- b. Empleando  $\alpha = 0.05$ , determine si hay relación entre soporte superior y precio.
- c. Confiaría en usar la ecuación de regresión estimada obtenida en el inciso a) para estimar el precio de las mochilas y botas para excursión con base en la evaluación del soporte superior.
- d. Estime el precio de una mochila que tiene un 4 como evaluación del soporte superior
28. En el ejercicio 10, con los datos de temperatura ( $^{\circ}\text{F}$ ) y precio (\$) de 11 sacos de dormir de Bergans de Norway se obtuvo la ecuación de regresión estimada  $\hat{y} = 359.2668 - 5.2772x$ . Empleando 0.05 como nivel de significancia, determine si temperatura y precio están relacionados. Dé la tabla de ANOVA. ¿Cuál es la conclusión?
29. Vuelva al ejercicio 21, en el que se usaron los datos sobre volumen de producción y costos para obtener una ecuación de regresión estimada que relacionaba el volumen de producción y los costos de una determinada operación de producción. Use  $\alpha = 0.05$  para determinar si el volumen de producción está relacionado de manera significativa con los costos totales. Dé la tabla ANOVA. ¿Cuál es la conclusión?
30. Vuelva al ejercicio 22, en el que se emplearon los datos siguientes para determinar si el precio de una impresora estaba relacionado con su velocidad para imprimir un texto (*PC World*, febrero 2003).



Nombre	Tipo	Velocidad (ppm)	Precio (\$)
Minolta-QMS PagePro 1250W	Oficina	12	199
Brother HL-1850	Oficina	10	499
Lexmark E320	Oficina	12.2	299
Minolta-QMS PagePro 1250E	Oficina	10.3	299
HP Laserjet 1200	Oficina	11.7	399
Xerox Phaser 4400/N	Corporativa	17.8	1850
Brother HL-2460N	Corporativa	16.1	1000

(continúa)

Nombre	Tipo	Velocidad (ppm)	Precio (\$)
IBM Infoprint 1120n	Corporativa	11.8	1387
Lexmark W812	Corporativa	19.8	2089
Oki Data B8300n	Corporativa	28.2	2200

¿Indican las evidencias que haya una relación significante entre velocidad de impresión y precio? Realice la prueba estadística apropiada y dé su conclusión. Use  $\alpha = 0.05$ .

31. En el ejercicio 20 con los datos sobre  $x$  = precio (\$) y  $y$  = evaluación general de 10 televisores de plasma, de 42 pulgadas probadas por *Consumer Reports* se obtuvo la ecuación de regresión estimada  $\hat{y} = 12.0169 + 0.0127x$ . Con estos datos se obtuvieron SCE = 540.04 y STC = 982.40. Use la prueba  $F$  para determinar si el precio de los televisores de plasma, de 42 pulgadas y la evaluación general están relacionados. Use  $\alpha = 0.05$ .

14.6

## Uso de la ecuación de regresión estimada para estimaciones y predicciones

Al usar el modelo de regresión lineal simple se hace una suposición acerca de la relación entre  $x$  y  $y$ . Despues se usa el método de mínimos cuadrados para obtener una ecuación de regresión lineal simple estimada. Si existe una relación significante entre  $x$  y  $y$  y si el coeficiente de determinación indica que el ajuste es bueno, la ecuación de regresión estimada es útil para estimaciones y predicciones.

### Estimación puntual

En el ejemplo de Armand's Pizza Parlors, la ecuación de regresión estimada  $\hat{y} = 60 + 5x$  proporciona una estimación de la relación entre  $x$  el tamaño de la población de estudiantes y  $y$  las ventas trimestrales. Con la ecuación de regresión estimada se puede obtener una estimación puntual del valor medio de  $y$  correspondiente a un determinado valor de  $x$  o se puede predecir el valor de  $y$  que corresponde a un valor de  $x$ . Por ejemplo, supóngase que los gerentes de Armand's desean una estimación puntual de la media de las ventas trimestrales de todos los restaurantes que se encuentren cerca de campus de 10 000 estudiantes. Usando la ecuación de regresión estimada  $\hat{y} = 60 + 5x$ , con  $x = 10$  (o 10 000 estudiantes) se obtiene  $\hat{y} = 60 + 5(10) = 110$ . Por lo tanto, una estimación puntual de la media de las ventas trimestrales de todos los restaurantes ubicados cerca de campus de 10 000 estudiantes es \$110 000.

Ahora supóngase que los administradores de Armand's desean predecir las ventas de un determinado restaurante ubicado cerca de Talbot College, una escuela de 10 000 estudiantes. En este caso lo que interesa no es la media correspondiente a todos los restaurantes que están cerca de campus de 10 000 estudiantes, sino únicamente predecir las ventas trimestrales de un determinado restaurante. En realidad, la estimación puntual de un solo valor de  $y$  es igual a la estimación puntual de la media de los valores de  $y$ . Por lo tanto, la predicción de las ventas trimestrales de este restaurante serán  $\hat{y} = 60 + 5(10) = 110$  o \$110 000.

### Estimación por intervalo

*Los intervalos de confianza y los intervalos de predicción indican la precisión de los resultados de la regresión. Los intervalos más estrechos proporcionan mayor precisión.*

Las estimaciones puntuales no proporcionan información alguna acerca de la precisión de una estimación. Para eso es necesario obtener estimaciones por intervalo que son muy parecidas a las de los capítulos 8, 10 y 11. El primer tipo de estimación por intervalo, el **intervalo de confianza** es una estimación por intervalo del *valor medio de las*  $y$  que corresponden a un valor dado de  $x$ . El segundo tipo de estimación por intervalo, el **intervalo de predicción**, se usa cuando se necesita una estimación por intervalo de un *solo valor de*  $y$  para un valor dado de  $x$ . La estimación puntual del valor medio de  $y$  es igual a la estimación puntual de un solo valor de  $y$ . Pero las estimaciones por intervalo que se obtienen para estos dos casos son diferentes. En un intervalo de predicción el margen de error es mayor.

## Intervalo de confianza para el valor medio de $y$

Con la ecuación de regresión estimada se obtiene una estimación puntual del valor medio de  $y$  que corresponde a un valor dado de  $x$ . Para obtener un intervalo de confianza se usa la notación siguiente.

$x_p$  = valor dado de la variable independiente  $x$

$y_p$  = valor de la variable dependiente  $y$  que corresponde al valor dado  $x_p$

$E(y_p)$  = valor medio o valor esperado de la variable dependiente  $y$  que corresponde al valor dado  $x_p$

$\hat{y}_p = b_0 + b_1 x_p$  = estimación puntual de  $E(y_p)$  cuando  $x = x_p$

Empleando esta notación para estimar la media de las ventas de los restaurantes Armand's que se encuentran cerca de un campus de 10 000 estudiantes, se tiene que  $x_p = 10$  y  $E(y_p)$  denota el valor medio desconocido de las ventas de todos los restaurantes para los que  $x_p = 10$ . La estimación puntual de  $E(y_p)$  está dada por  $\hat{y}_p = 60 + 5(10) = 110$ .

En general, no se puede esperar que  $\hat{y}_p$  sea exactamente igual a  $E(y_p)$ . Para hacer una inferencia acerca de qué tan cerca está  $\hat{y}_p$  de la media  $E(y_p)$ , es necesario estimar la varianza de  $\hat{y}_p$ . La fórmula para estimar la varianza de  $\hat{y}_p$  para un  $x_p$  dado se denota  $s_{\hat{y}_p}^2$ , y es

$$s_{\hat{y}_p}^2 = s^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \quad (14.22)$$

Una estimación de la desviación estándar de  $\hat{y}_p$  está dada por la raíz cuadrada de la ecuación (14.22).

$$s_{\hat{y}_p} = s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (14.23)$$

En los resultados obtenidos en la sección 14.5 para el ejemplo de Armand's Pizza Parlors se tiene  $s = 13.829$ . Como  $x_p = 10$ ,  $\bar{x} = 14$  y  $\sum(x_i - \bar{x})^2 = 568$ , usando la ecuación (14.23) se obtiene

$$\begin{aligned} s_{\hat{y}_p} &= 13.829 \sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829 \sqrt{0.1282} = 4.95 \end{aligned}$$

A continuación se presenta la fórmula general para obtener un intervalo de confianza.

### INTERVALO DE CONFIANZA PARA $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \quad (14.24)$$

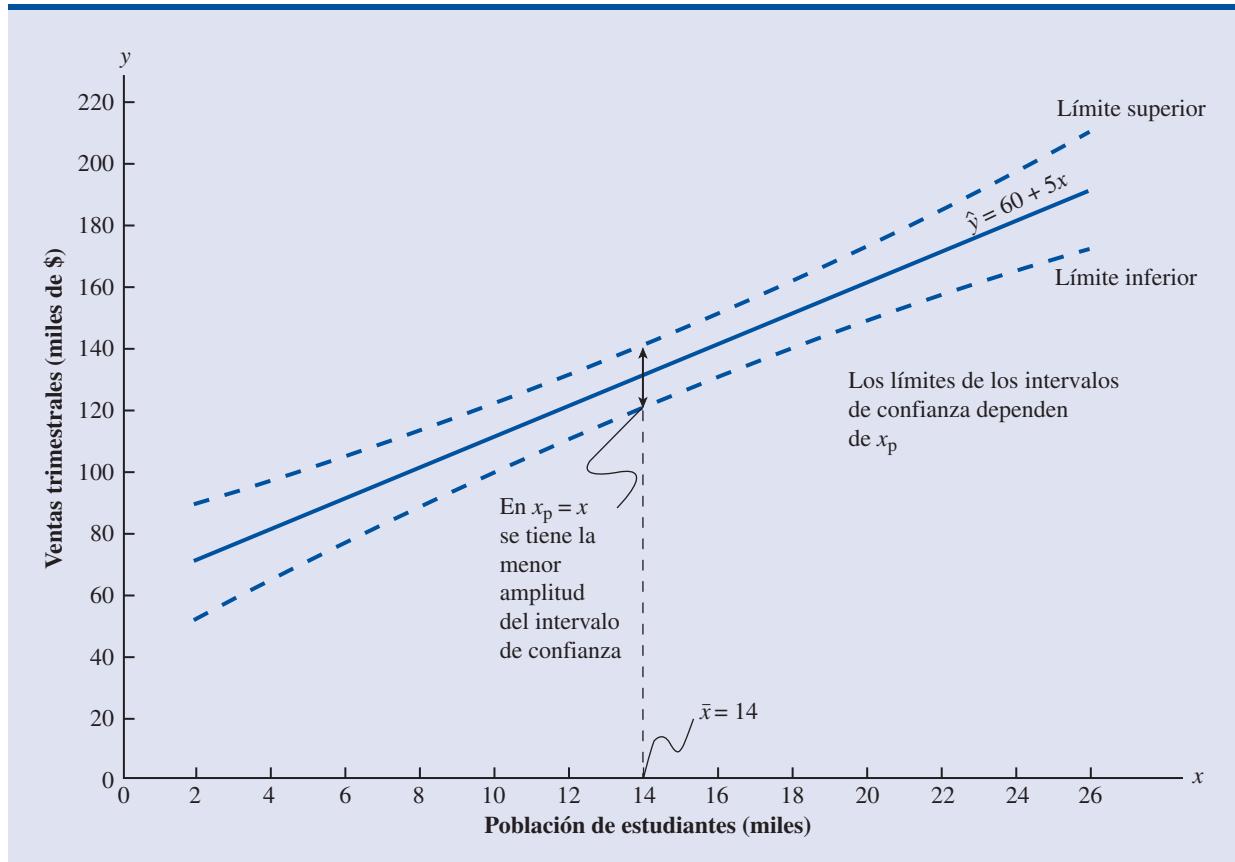
donde el coeficiente de confianza es  $1 - \alpha$  y  $t_{\alpha/2}$  es un valor de la distribución  $t$  con  $n - 2$  grados de libertad.

El margen de error en esta estimación por intervalo (este intervalo de estimación) es  $t_{\alpha/2} s_{\hat{y}_p}$ .

Para obtener, con la fórmula (14.24), un intervalo de confianza de 95% para la media de las ventas trimestrales de los restaurantes Armand's que se encuentran cerca de campus de 10 000 estudiantes, se necesita el valor de  $t$  para  $\alpha/2 = 0.025$  y  $n - 2 = 10 - 2 = 8$  grados de libertad. En la tabla 2 del apéndice B, se encuentra  $t_{0.025} = 2.306$ . Por lo tanto, como  $\hat{y}_p = 110$  y el margen de error es  $t_{\alpha/2} s_{\hat{y}_p} = 2.306(4.95) = 11.415$ , la estimación por intervalo de 95% de confianza es

$$110 \pm 11.415$$

**FIGURA 14.8** INTERVALOS DE CONFIANZA PARA LA MEDIA DE LAS VENTAS  $y$  CORRESPONDIENTES A VALORES DADOS DEL TAMAÑO DE LA POBLACIÓN DE ESTUDIANTES  $x$



En dólares, el intervalo de 95% de confianza para la media de las ventas trimestrales de todos los restaurantes que se encuentran cerca de un campus de 10 000 estudiantes es  $110\,000 \pm \$11\,415$ . Por lo tanto, si el tamaño de la población de estudiantes es 10 000, el intervalo de 95% de confianza para la media de las ventas trimestrales en los restaurantes cercanos a un campus de 10 000 estudiantes es el intervalo que va de \$98 585 a \$121 415.

Obsérvese que la desviación estándar estimada de  $\hat{y}_p$ , dada por la ecuación (14.23), es menor cuando  $x_p = \bar{x}$  y la cantidad  $x_p - \bar{x} = 0$ . En este caso, la desviación estándar estimada de  $\hat{y}_p$  se convierte en

$$s_{\hat{y}_p} = s \sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s \sqrt{\frac{1}{n}}$$

Esto significa que cuando  $x_p = \bar{x}$  se obtiene la mejor estimación o la estimación más precisa del valor medio de  $y$ . Entre más alejada esté  $x_p$  de  $\bar{x}$ , mayor será  $x_p - \bar{x}$ . El resultado es que los intervalos de confianza para el valor medio de  $y$  son más amplios a medida que  $x_p$  se aleja de  $\bar{x}$ . En la figura 14.8 se muestra esto gráficamente.

### Intervalo de predicción para un solo valor de $y$

Supóngase que en lugar de que lo interese sea estimar el valor medio de las ventas de todos los restaurantes Armand's que se encuentran cerca de campus de 10 000 estudiantes, se deseen estimar las ventas de un solo restaurante que se encuentra cerca de Talbot College, una escuela

de 10 000 estudiantes. Como ya se indicó, la estimación puntual de  $y_p$ , el valor de  $y$  que corresponde a un valor dado  $x_p$ , se obtiene mediante la ecuación de regresión  $\hat{y}_p = b_0 + b_1 x_p$ . En el caso del restaurante cerca de Talbot College, como  $x_p = 10$ , las ventas trimestrales pronosticadas serán  $\hat{y}_p = 60 + 5(10) = 110$  o \$110 000. Obsérvese que este valor es el mismo que el obtenido como estimación puntual de la media de las ventas en los restaurantes que se encuentran cerca de campus de 10 000 estudiantes.

Para obtener un intervalo de predicción, es necesario determinar primero la varianza correspondiente al uso de  $\hat{y}_p$  como estimación de un valor individual de  $y$  cuando a  $x = x_p$ . Esta varianza está formada por la suma de los dos componentes siguientes.

1. La varianza de los valores individuales de  $y$  respecto a la media  $E(y_p)$ , para la cual una estimación está dada por  $s^2$
2. La varianza correspondiente al uso de  $\hat{y}_p$  para estimar  $E(y_p)$ , para la cual una estimación está dada por  $s_{\hat{y}_p}^2$

La fórmula para estimar la varianza de un valor individual de  $y_p$  que se denota  $s_{\text{ind}}^2$ , es

$$\begin{aligned}s_{\text{ind}}^2 &= s^2 + s_{\hat{y}_p}^2 \\&= s^2 + s^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \\&= s^2 \left[ 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]\end{aligned}\tag{14.25}$$

Por lo tanto, una estimación de la desviación estándar de un solo valor de  $y_p$  es la dada por

$$s_{\text{ind}} = s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\tag{14.26}$$

En el ejemplo de Armand's Pizza Parlors, la desviación estándar estimada que corresponde a la predicción de las ventas de un determinado restaurante que esté cerca de un campus de 10 000 estudiantes se calcula como sigue.

$$\begin{aligned}s_{\text{ind}} &= 13.829 \sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\&= 13.829 \sqrt{1.1282} \\&= 14.69\end{aligned}$$

La fórmula general para un intervalo de predicción es como sigue

#### INTERVALO DE PREDICCIÓN PARA $y_p$

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}\tag{14.27}$$

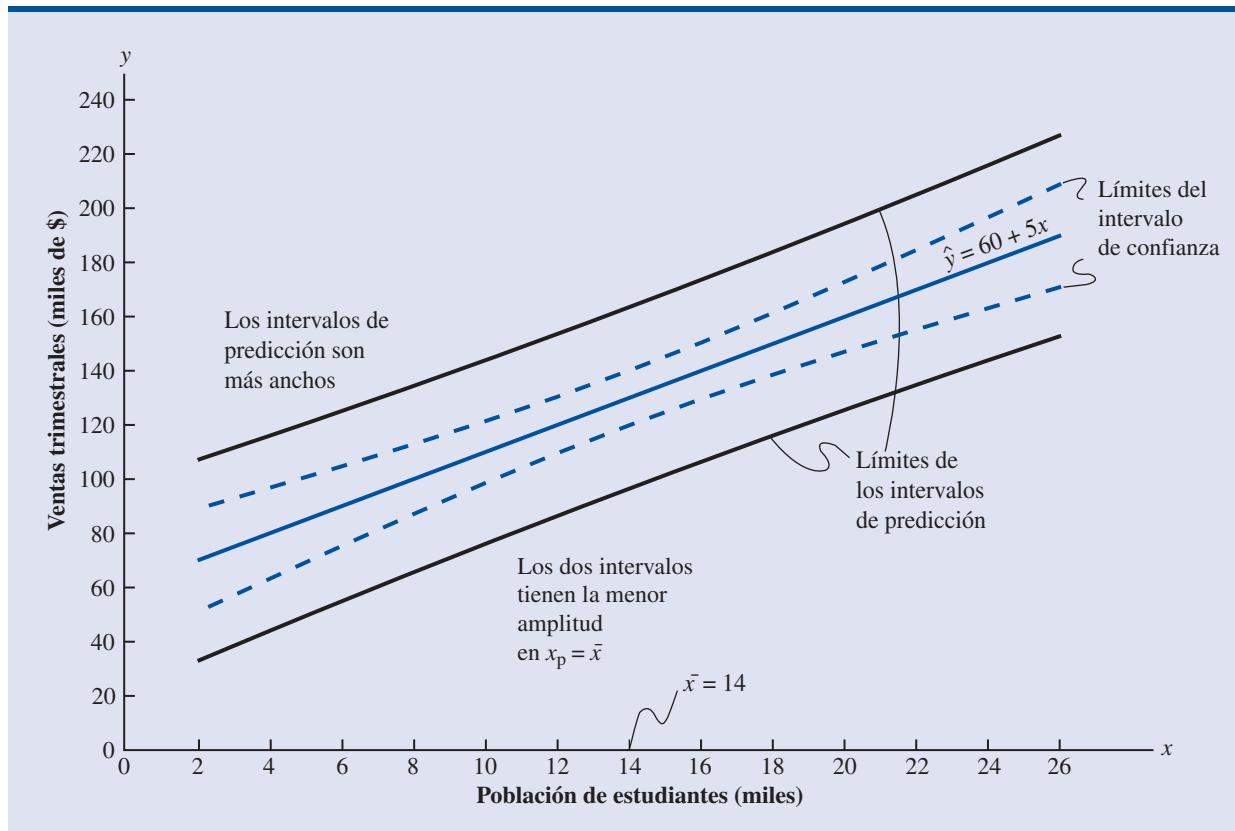
donde el coeficiente de confianza es  $1 - \alpha$  y  $t_{\alpha/2}$  es un valor de la distribución  $t$  para  $n - 2$  grados de libertad

*El margen de error de este intervalo de estimación es  $t_{\alpha/2} s_{\text{ind}}$ .*

El intervalo de predicción de las ventas trimestrales del restaurante situado cerca de Talbot College se encuentra empleando  $t_{0.025} = 2.306$  y  $s_{\text{ind}} = 14.69$ . Por lo tanto, como  $\hat{y}_p = 110$  y el margen de error es  $t_{\alpha/2} s_{\text{ind}} = 2.306(14.69) = 33.875$ , el intervalo de predicción de 95% de confianza es

$$110 \pm 33.875$$

**FIGURA 14.9** INTERVALOS DE CONFIANZA Y DE PREDICCIÓN PARA LAS VENTAS  $y$  QUE CORRESPONDEN A VALORES DADOS  $x$  DEL TAMAÑO DE LA POBLACIÓN DE ESTUDIANTES



En dólares, el intervalo de predicción es  $\$110\,000 \pm \$33\,875$  o el intervalo que va de  $\$76\,125$  a  $\$143\,875$ . Obsérvese que el intervalo de predicción para un solo restaurante que se encuentre cerca de un campo de 10 000 estudiantes es más amplio que el intervalo de confianza para la media de las ventas de todos los restaurantes que se encuentran cerca de campus de 10 000 estudiantes. Esta diferencia refleja el hecho de que se puede estimar con más precisión la media de  $y$  que un solo valor individual de  $y$ .

*En general, tanto las líneas de los límites para los intervalos de confianza como las de los límites para los intervalos de predicción tienen cierta curvatura.*

Tanto las estimaciones mediante un intervalo de confianza como las estimaciones mediante un intervalo de predicción son más precisas cuando el valor de la variable independiente es  $x_p = \bar{x}$ . En la figura 14.9 se muestra la forma general de los intervalos de confianza y de los intervalos de predicción que son más anchos.

## Ejercicios

# Métodos

32. Los datos siguientes son los del ejercicio 1.

$x_i$	1	2	3	4	5
$y_i$	3	7	5	11	14

- a. Use la ecuación (14.23) para estimar la desviación estándar de  $\hat{y}_p$  cuando  $x = 4$ .
  - b. Use la expresión (14.24) para obtener un intervalo de confianza de 95% para el valor esperado de  $y$  cuando  $x = 4$ .

- c. Use la ecuación (14.26) para estimar la desviación estándar de un valor de  $y$  cuando  $x = 4$ .  
 d. Use la expresión (14.27) para obtener un intervalo de predicción de 95% para  $y$  cuando  $x = 4$
33. Los datos siguientes son los del ejercicio 2.
- |       |    |    |    |    |    |
|-------|----|----|----|----|----|
| $x_i$ | 3  | 12 | 6  | 20 | 14 |
| $y_i$ | 55 | 40 | 55 | 10 | 15 |
- a. Estime la desviación estándar de  $\hat{y}_p$  cuando  $x = 8$ .  
 b. Obtenga un intervalo de 95% de confianza para el valor esperado de  $y$  cuando  $x = 8$ .  
 c. Estime la desviación estándar de un valor individual de  $y$  cuando  $x = 8$ .  
 d. Obtenga un intervalo de predicción de 95% para  $y$  cuando  $x = 8$ .
34. Los datos siguientes son los del ejercicio 3.
- |       |   |    |   |    |    |
|-------|---|----|---|----|----|
| $x_i$ | 2 | 6  | 9 | 13 | 20 |
| $y_i$ | 7 | 18 | 9 | 26 | 23 |

Obtenga los intervalos de confianza y de predicción del 95% para  $x = 12$ . Explique por qué son diferentes estos dos intervalos.

## Aplicaciones

**Autoexamen**



35. En el ejercicio 18, con los datos de los promedios de calificaciones  $x$  y los salarios mensuales  $y$  se obtuvo la ecuación de regresión estimada  $\hat{y} = 1790.5 + 581.1x$ .
- Dé un intervalo de 95% de confianza para el salario medio inicial de todos los estudiantes cuyo promedio fue 3.0.
  - Dé un intervalo de 95% de predicción para el salario medio inicial de Joe Heller cuyo promedio fue 3.0.
36. En el ejercicio 10, a partir de los datos de temperatura ( $^{\circ}\text{F}$ ) =  $x$  y precio (\$) =  $y$  de 12 sacos de dormir, fabricados por Bergans of Noway, se obtuvo la ecuación de regresión  $\hat{y} = 359.2668 - 5.2772x$ . Para estos datos  $s = 37.9372$ .
- Dé una estimación puntual del precio de un saco de dormir cuya temperatura sea 30.
  - Dé un intervalo de 95% de confianza para el precio medio de todos los sacos de dormir cuya temperatura sea 30.
  - Suponga que Bergans elabora un nuevo modelo cuya temperatura es 30. Dé un intervalo de predicción de 95% para el precio de este nuevo modelo.
  - Explique la diferencia entre sus respuestas a los incisos b) y c).
37. En el ejercicio 13 se proporcionaron datos sobre el ingreso bruto ajustado y el monto de las deducciones en las declaraciones de impuestos. Los datos se dieron en miles de dólares. Como la ecuación de regresión estimada es  $\hat{y} = 4.68 + 0.16x$ , el monto razonable de las deducciones, para un contribuyente cuyo ingreso bruto ajustado sea \$52 500, es \$13 080.
- Dé un intervalo de 95% de confianza para el monto medio de las deducciones de todos los contribuyentes cuyo ingreso bruto ajustado sea \$52 500.
  - Dé un intervalo de predicción de 95% para el monto total de deducciones de un contribuyente cuyo ingreso bruto ajustado sea \$52 500.
  - Si el contribuyente del inciso b) solicita deducciones de \$20 400, ¿se justificaría que se le quiera hacer una auditoría?
  - Emplee su respuesta al inciso b) para indicar el monto de las deducciones que puede solicitar un contribuyente cuyo ingreso bruto ajustado sea \$52 500 sin que se le haga una auditoría.
38. Retome el ejercicio 21, en el que la ecuación de regresión estimada  $\hat{y} = 1246.67 + 7.6x$  se obtuvo empleando los datos de volumen de producción  $x$  y costos totales  $y$  de una determinada operación de fabricación.
- En el plan de producción de la empresa se ve que el mes próximo deberán producirse 500 unidades. Dé la estimación puntual de los costos totales.

- b. Dé un intervalo de predicción de 99% para el costo total de producción de las 500 unidades, el mes próximo.
  - c. Si al final del mes próximo, el informe de costos de un contador indica que en ese mes los costos reales de producción fueron \$6000, ¿debería preocupar a los gerentes el haber incurrido ese mes en costos totales tan altos? Analice.
39. En Estados Unidos casi todo el sistema de tranvías usa vagones eléctricos que corren sobre vías a nivel de la calle. La Administración de Tránsito Federal afirma que el tranvía es uno de los medios de transporte más seguros, ya que la tasa de accidentes es 0.99 accidentes por millón de millas-pasajero en comparación con 2.29 en los autobuses. En los datos siguientes se dan las millas de vía y la cantidad de pasajeros transportados en los días laborables, en miles, de seis sistemas de tranvías (*USA Today*, 7 de enero 2003).

Ciudad	Millas de vías	Pasajeros transportados (miles)
Cleveland	15	15
Denver	17	35
Portland	38	81
Sacramento	21	31
San Diego	47	75
San Jose	31	30
St. Louis	34	42

- a. Use estos datos para obtener la ecuación de regresión estimada que podría emplearse para predecir la cantidad de pasajeros dadas las millas de vías.
- b. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
- c. Obtenga un intervalo de 95% de confianza para la media de la cantidad de pasajeros transportados en los días laborables en los sistemas de tranvías que tienen 30 millas de vías.
- d. Suponga que Charlotte está considerando la construcción de un sistema de tranvía de 30 millas de vías. Dé un intervalo de predicción de 95% para la cantidad de pasajeros transportada en un día laborable por el sistema Charlotte. ¿Cree usted que el intervalo de predicción que desarrolló pueda ser útil a los que están planeando Charlotte para anticipar la cantidad de pasajeros en un día laborable en su sistema de tranvía? Explique

## 14.7 Solución por computadoras

Realizar los cálculos del análisis de regresión sin la ayuda de una computadora puede costar mucho tiempo. En esta sección se verá cómo mediante el uso de paquetes de software como Minitab puede minimizarse la complicación de hacer tantos cálculos.

Los datos sobre población de estudiantes y ventas se han ingresado en la hoja de cálculo de Minitab. A la variable independiente se le ha llamado Pop y a la variable independiente se le ha llamado Ventas para facilitar la interpretación de los resultados que proporciona la computadora. Usando Minitab para el ejemplo de Armand's Pizza Parlors se obtuvieron los resultados que se muestran en la figura 14.10\*. A continuación se explica cómo interpretar estos resultados.

1. Minitab da la ecuación de regresión estimada como  $Ventas = 60.0 + 5.00 \text{Pop}$ .
2. Da también una tabla en la que indica el valor de los coeficientes  $b_0$  y  $b_1$ , la desviación estándar de cada coeficiente, el valor  $t$  obtenido al dividir cada coeficiente entre su desviación estándar y el valor- $p$  correspondiente a la prueba  $t$ . Como el valor- $p$  es cero (a tres cifras decimales), los resultados muestrales indican que debe rechazarse la hipótesis nula ( $H_0: \beta_1 = 0$ ). O bien, se puede comparar 8.62 (que aparece en la columna  $t$ ) con el valor crítico apropiado. Este procedimiento para la prueba  $t$  se describió en la sección 14.5.

\*En el apéndice 14.3 se dan los pasos que hay que seguir con Minitab para obtener estos resultados.

**FIGURA 14.10** RESULTADOS DADOS POR MINITAB PARA EL PROBLEMA DE ARMAND'S PIZZA PARLORS

The regression equation is Sales = 60.0 + 5.00 Pop	Estimated regression equation
Predictor Coef SE Coef T p	
Constant 60.000 9.226 6.50 0.000	
Pop 5.0000 0.5803 8.62 0.000	
S = 13.83 R-sq = 90.3% R-sq(adj) = 89.1%	
Analysis of Variance	
SOURCE DF SS MS F p	
Regression 1 14200 14200 74.25 0.000	
Residual Error 8 1530 191	
Total 9 15730	ANOVA table
Predicted Values for New Observations	
New Obs Fit SE Fit 95% C.I. 95% P.I.	
1 110.00 4.95 (98.58, 121.42) (76.12, 143.88)	Interval estimates

3. Minitab da el error estándar de estimación,  $s = 13.83$ , así como información acerca de la bondad de ajuste. Observe que “R-sq = 90.3%” es el coeficiente de determinación expresado como porcentaje. El valor “R-sq (adj) = 89.1%” se verá en el capítulo 15.
4. La tabla ANOVA se presenta bajo el encabezado Analysis of Variance. Minitab usa el rótulo Residual Error para la fuente de variación del error. Obsérvese que DF es la abreviación de degrees of freedom (= grados de libertad) y que CMR está dado como 14 200 y ECM como 191. El cociente de estos dos valores da el valor  $F$  que es 74.25 y el correspondiente valor- $p$  0.000. como el valor- $p$  es cero (a tres lugares decimales), la relación entre ventas (Sales) y población (Pop) se considera estadísticamente significante.
5. La estimación de las ventas esperadas mediante un intervalo de confianza de 95% y la estimación de las ventas de un determinado restaurante cercano a un campus de 10 000 estudiantes mediante un intervalo de estimación de 95% se dan abajo de la tabla ANOVA. El intervalo de confianza es (98.58, 121.42) y el intervalo de predicción es (76.12, 143.88) como se indicó en la sección 14.6.

## Ejercicios

### Aplicaciones

40. La división comercial de una empresa inmobiliaria realiza un análisis de regresión de la relación entre  $x$ , rentas brutas anuales (en miles de dólares) y  $y$ , precio de venta (en miles de dólares) de edificios de departamentos. Se obtuvieron datos sobre varias propiedades vendidas últimamente y con la computadora se obtuvieron los resultados siguientes.

The regression equation is  
 $Y = 20.0 + 7.21 X$

Predictor	Coef	SE Coef	T
Constant	20.000	3.2213	6.21
X	7.210	1.3626	5.29

#### Analysis of Variance

SOURCE	DF	SS
Regression	1	41587.3
Residual Error	7	
Total	8	51984.1

- a. ¿Cuántos edificios de departamentos había en la muestra?
- b. Dé la ecuación de regresión estimada
- c. ¿Cuál es el valor de  $s_{b_1}$ ?
- d. Use el estadístico  $F$  para probar la significancia de la relación empleando 0.05 como nivel de significancia.
- e. Estime el precio de venta de un edificio de departamentos cuyas rentas anuales brutas son \$50 000.
41. A continuación se presenta una parte de los resultados por computadora de un análisis de regresión en el que se relaciona  $y$  = gastos de mantenimiento (dólares por mes) con  $x$  uso (horas por semana) para una marca determinada de terminal de computadora.

The regression equation is  
 $Y = 6.1092 + .8951 X$

Predictor	Coef	SE Coef
Constant	6.1092	0.9361
X	0.8951	0.1490

#### Analysis of Variance

SOURCE	DF	SS	MS
Regression	1	1575.76	1575.76
Residual Error	8	349.14	43.64
Total	9	1924.90	

- a. Dé la ecuación de regresión estimada.
- b. Use una prueba  $t$  para determinar si los gastos mensuales de mantenimiento están relacionados con el uso, empleando 0.05 como nivel de significancia.
- c. Utilice la ecuación de regresión estimada para predecir los gastos mensuales de mantenimiento de una terminal que se usa 25 hora por semana.
42. Un modelo de regresión que relaciona  $x$ , el número de vendedores en una sucursal, con  $y$ , las ventas anuales en esa sucursal (en miles de dólares), proporcionó el siguiente resultado de computadora empleando análisis de regresión de los datos.

The regression equation is $Y = 20.0 + 7.21 X$
Predictor      Coef      SE Coef      T
Constant      20.000      3.2213      6.21
X              7.210      1.3626      5.29
Analysis of Variance
SOURCE      DF      SS
Regression      1      41587.3
Residual Error      7
Total      8      51984.1

- a. Dé la ecuación de regresión estimada.
- b. ¿Cuántas sucursales participaron en el estudio?
- c. Calcule el estadístico  $F$  y pruebe la significancia de la relación empleando 0.05 como nivel de significancia.
- d. Pronostique las ventas anuales de la sucursal Memphis. En esta sucursal hay 12 vendedores.
43. Los expertos en salud recomiendan que los corredores beban 4 onzas de agua por cada 15 minutos que corran. Aunque las botellas de plástico son una buena alternativa para la mayoría de los corredores, cuando se corre todo un día a campo traviesa se requieren sistemas de hidratación que se llevan sobre la cintura o sobre la espalda. Estos sistemas de hidratación además de permitir llevar más agua permiten llevar también alimento o ropa. Por supuesto, a medida que aumenta la capacidad de estos sistemas, aumenta también su peso y su precio. En la lista siguiente se da peso y precio de 26 de estos sistemas de hidratación (*Trail Runner Gear Guide*, 2003).

Modelo	Peso (onzas)	Precio (\$)
Fastdraw	3	10
Fastdraw Plus	4	12
Fitness	5	12
Access	7	20
Access Plus	8	25
Solo	9	25
Serenade	9	35
Solitaire	11	35
Gemini	21	45
Shadow	15	40
SipStream	18	60
Express	9	30
Lightning	12	40
Elite	14	60
Extender	16	65
Stinger	16	65
GelFlask Belt	3	20
GelDraw	1	7
GelFlask Clip-on Holster	2	10
GelFlask Holster SS	1	10
Strider (W)	8	30

Modelo	Peso (onzas)	Precio (\$)
Walkabout (W)	14	40
Solitude I.C.E.	9	35
Getaway I.C.E.	19	55
Profile I.C.E.	14	50
Traverse I.C.E.	13	60

- a. Con estos datos obtenga una ecuación de regresión estimada que pueda ser empleada para predecir el precio de un sistema de hidratación en función de su peso.
- b. Pruebe la significancia de la relación empleando 0.05 como nivel de significancia.
- c. ¿Proporciona un buen ajuste la ecuación de regresión estimada?
- d. Suponga que la ecuación de regresión estimada obtenida en el inciso a) también pueda usarse para sistemas de hidratación elaborados por otras empresas. Obtenga un intervalo de confianza de 95% para estimar el precio de todos los sistemas de hidratación que pesan 10 onzas.
- e. Suponga que la ecuación de regresión estimada obtenida en el inciso a) también pueda usarse para sistemas de hidratación elaborados por otras empresas. Obtenga un intervalo de predicción de 95% para estimar el precio del sistema Back Draft elaborado por Eastern Mountain Sports; este sistema de hidratación pesa 10 onzas.
44. Cushman Wakefield, Inc. recoge datos sobre la tasa de desocupación en edificios de oficinas y las tasas de las rentas en mercados de Estados Unidos. Los datos siguientes dan la tasa de desocupación (%) y las tasas de rentas promedio (por pie cuadrado) en las zonas comerciales centrales de 18 mercados.

Mercado	Tasa de desocupación (%)	Tasa promedio (\$)
Atlanta	21.9	18.54
Boston	6.0	33.70
Hartford	22.8	19.67
Baltimore	18.1	21.01
Washington	12.7	35.09
Philadelphia	14.5	19.41
Miami	20.0	25.28
Tampa	19.2	17.02
Chicago	16.0	24.04
San Francisco	6.6	31.42
Phoenix	15.9	18.74
San Jose	9.2	26.76
West Palm Beach	19.7	27.72
Detroit	20.0	18.20
Brooklyn	8.3	25.00
Downtown, NY	17.1	29.78
Midtown, NY	10.8	37.03
Midtown South, NY	11.1	28.64

- a. Con estos datos trace un diagrama de dispersión; en el eje horizontal grafique la tasa de desocupación.
- b. ¿Parece haber alguna relación entre las tasas de desocupación y las tasas de rentas?
- c. Dé la ecuación de regresión para predecir la tasa promedio de renta en función de una tasa de desocupación dada.
- d. Empleando como nivel de significancia 0.05 pruebe la significancia de esta relación.

- e. ¿Proporciona la ecuación de regresión estimada, un buen ajuste? Explique.
- f. Pronostique la tasa de renta esperada en los mercados en los que la tasa de desocupación en zonas comerciales centrales es 25%.
- g. La tasa de desocupación general en la zona comercial central de Ft. Lauderdale es 11.3%. Pronostique la tasa de renta esperada en Ft. Lauderdale.

### 14.8

## Análisis residual: confirmación de las suposiciones del modelo

*El análisis residual es la herramienta principal para determinar si el modelo de regresión empleado es apropiado.*

Como ya se indicó, el *residual* de la observación  $i$  es la diferencia entre el valor observado de la variable dependiente ( $y_i$ ) y el valor estimado de la variable dependiente ( $\hat{y}_i$ )

RESIDUAL DE LA OBSERVACIÓN  $i$

$$y_i - \hat{y}_i \quad (14.28)$$

donde

$y_i$  es el valor observado de la variable dependiente

$\hat{y}_i$  es el valor estimado de la variable dependiente

En otras palabras, el residual  $i$  es el error que resulta de usar la ecuación de regresión estimada para predecir el valor de la variable dependiente. En la tabla 14.7 se calculan estos residuales correspondientes a los datos del ejemplo de Armand's Pizza Parlors. En la segunda columna de la tabla se presentan los valores observados de la variable dependiente y en la tercera columna, los valores estimados de la variable dependiente obtenidos usando la ecuación de regresión estimada  $\hat{y} = 60 + 5x$ . Un análisis de los residuales correspondientes, que se encuentran en la cuarta columna de la tabla, ayuda a determinar si las suposiciones hechas acerca del modelo de regresión son adecuadas.

A continuación se revisan las suposiciones de regresión en el ejemplo de Armand's Pizza Parlors. Se supuso un modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon \quad (14.29)$$

**TABLA 14.7** RESIDUALES EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Población de estudiantes $x_i$	Ventas $y_i$	Ventas estimadas $\hat{y}_i = 60 + 5x_i$	Residuales $y_i - \hat{y}_i$
2	58	70	-12
6	105	90	15
8	88	100	-12
8	118	100	18
12	117	120	-3
16	137	140	-3
20	157	160	-3
20	169	160	9
22	149	170	-21
26	202	190	12

Este modelo indica que se supone que las ventas trimestrales ( $y$ ) son función lineal del tamaño de la población de estudiantes ( $x$ ), más un término del error  $\epsilon$ . En la sección 14.4, para el término del error  $\epsilon$  se hicieron las siguientes suposiciones

1.  $E(\epsilon) = 0$ .
2. La varianza de  $\epsilon$ , que se denota  $\sigma^2$ , es la misma para todos los valores de  $x$ .
3. Los valores de  $\epsilon$  son independientes.
4. El término del error  $\epsilon$  tiene distribución normal.

Estas suposiciones son la base teórica para las pruebas  $t$  y  $F$  que se usan para determinar si la relación entre  $x$  y  $y$  es significativa y para las estimaciones, mediante intervalos de confianza y de predicción, presentadas en la sección 14.6. Si las suposiciones acerca del término del error  $\epsilon$  son dudosas, puede ser que las pruebas de hipótesis acerca de la significancia de la relación de regresión y los resultados de la estimación por intervalo no sean correctos.

Los residuales proporcionan la mejor información acerca de  $\epsilon$ ; por lo tanto, el análisis de los residuales es muy importante para determinar si las suposiciones hechas acerca de  $\epsilon$  son apropiadas. Gran parte del análisis residual se basa en examinar gráficas. En esta sección se estudiarán las siguientes gráficas de residuales.

1. La gráfica de residuales contra los valores de la variable independiente  $x$
2. La gráfica de residuales contra los valores pronosticados para la variable dependiente  $\hat{y}$
3. La gráfica de residuales estandarizados
4. La gráfica de probabilidad normal.

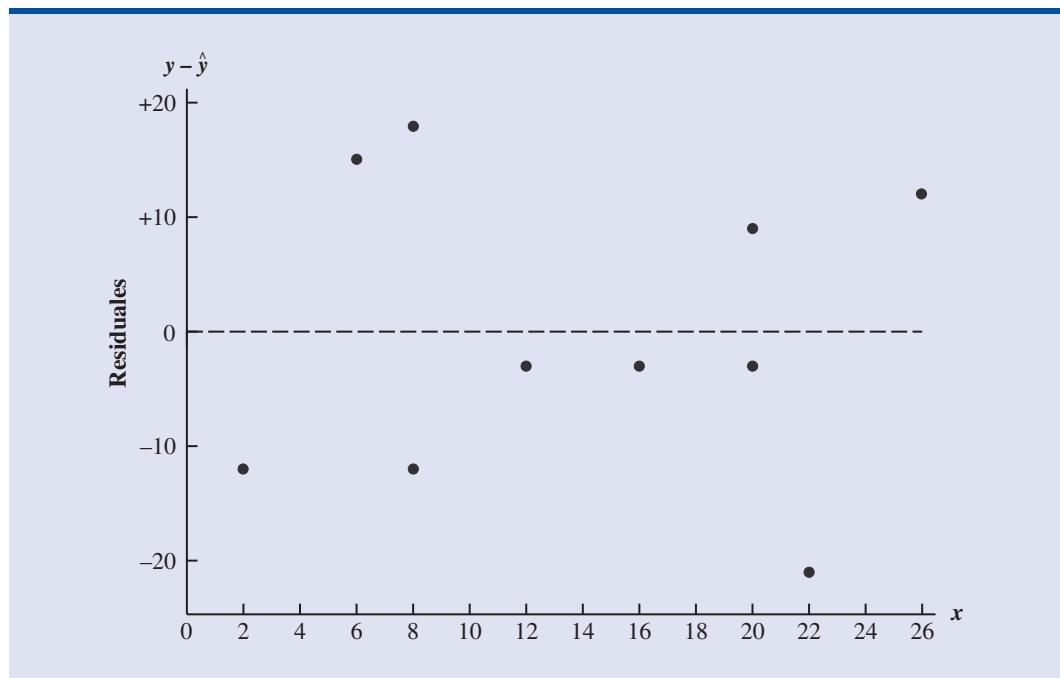
## Gráfica de residuales contra $x$

La **gráfica de residuales** contra la variable independiente  $x$  es una gráfica en la que los valores de la variable independiente se representan en el eje horizontal y los valores de los residuales correspondientes se representan en el eje vertical. Para cada residual se grafica un punto. La primera coordenada de cada punto está dada por el valor  $x_i$  y la segunda coordenada está dada por el correspondiente valor del residual  $y_i - \hat{y}_i$ . En la gráfica de residuales contra  $x$  obtenida con los datos de Armand's Pizza Parlors de la tabla 14.7, las coordenadas del primer punto son  $(2, -12)$ , que corresponden a  $x_1 = 2$  y  $y_1 - \hat{y}_1 = -12$ ; las coordenadas del segundo punto son  $(6, 15)$ , que corresponden a  $x_2 = 6$  y  $y_2 - \hat{y}_2 = 15$ ; etc. En la figura 14.11 se muestra la gráfica de residuales obtenida.

Antes de interpretar los resultados de esta gráfica de residuales, se considerarán algunas de las formas generales que pueden tener las gráficas de residuales. En la figura 14.12 se muestran tres ejemplos. Si la suposición de que la varianza de  $\epsilon$  es la misma para todos los valores de  $x$  y si el modelo de regresión empleado representa adecuadamente la relación entre las variables, el aspecto general de la gráfica de residuales será el de una banda horizontal de puntos como en la gráfica A de la figura 14.12. Pero si la varianza de  $\epsilon$  no es la misma para todos los valores  $x$  —por ejemplo, si la variabilidad respecto a la línea de regresión es mayor para valores de  $x$  mayores— el aspecto de la gráfica puede ser como el de la gráfica B de la figura 14.12. En este caso, se viola la suposición de que  $\epsilon$  tiene una varianza constante. En la gráfica C se muestra otra forma que puede tomar la gráfica de residuales. En este caso, se puede concluir que el modelo de regresión empleado no representa adecuadamente la relación entre las variables, y deberá considerarse un modelo de regresión curvilíneo o múltiple.

Volviendo, ahora, a la gráfica de los residuales del ejemplo de Armand's Pizza Parlors, figura 14.11. Estos residuales parecen tener una forma que se aproxima a la forma de banda horizontal de la gráfica A de la figura 14.12. Por lo tanto, se concluye que esta gráfica de residuales no muestra evidencias de que las suposiciones hechas para el modelo de regresión de Armand's puedan ser dudosas. Se concluye que el modelo de regresión lineal simple empleado para el ejemplo de Armand's, es válido.

**FIGURA 14.11** GRÁFICA DE RESIDUALES CONTRA LA VARIABLE INDEPENDIENTE  $x$  OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



Para la adecuada interpretación de las gráficas de residuales experiencia y criterio son muy importantes. Es raro que una gráfica de residuales tenga exactamente la forma de una de las gráficas presentadas en la figura 14.12. Sin embargo, los analistas que realizan frecuentemente estudios de regresión y gráficas de residuales se vuelven expertos en reconocer las diferencias entre las formas razonables y las que indican que se pude dudar de las suposiciones del modelo. Una gráfica de residuales proporciona una técnica para evaluar la validez de las suposiciones en un modelo de regresión.

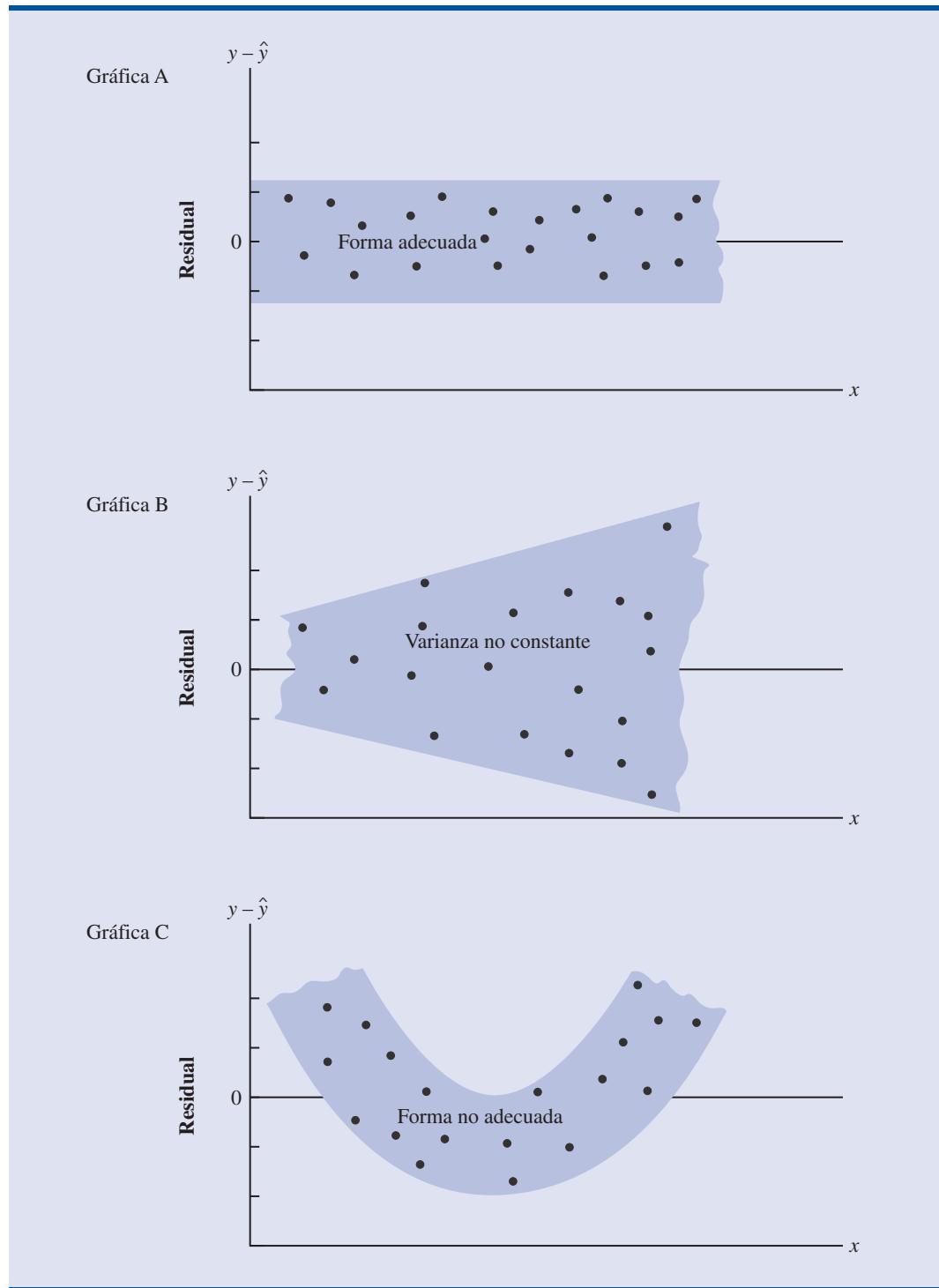
### Gráfica de residuales contra $\hat{y}$

En otra gráfica de residuales los valores pronosticados para la variable dependiente  $\hat{y}$  se representan en el eje horizontal y los valores de los residuales en el eje vertical. A cada residual corresponde un punto en la gráfica. La primera coordenada de cada uno de los puntos es  $\hat{y}_i$  y la segunda coordenada es el valor correspondiente del residual  $i, y_i - \hat{y}_i$ . Empleando los datos de Armand's, tabla 14.7, las coordenadas del primer punto son (70, -12), que corresponden a  $\hat{y}_1 = 70$  y  $y_1 - \hat{y}_1 = -12$ ; las coordenadas del segundo punto son (90, 15), etc. En la figura 14.13 se presenta esta gráfica de residuales. Obsérvese que la forma de esta gráfica de residuales es igual a la forma de la gráfica de residuales contra la variable independiente  $x$ . Esta no es una forma que pudiera llevar a dudar de las suposiciones del modelo. En la regresión lineal simple, tanto la gráfica de residuales contra  $x$  como la gráfica de residuales contra  $\hat{y}$  tienen la misma forma. En el análisis de regresión múltiple, la gráfica de residuales contra  $\hat{y}$  se usa más debido a que se tiene más de una variable independiente.

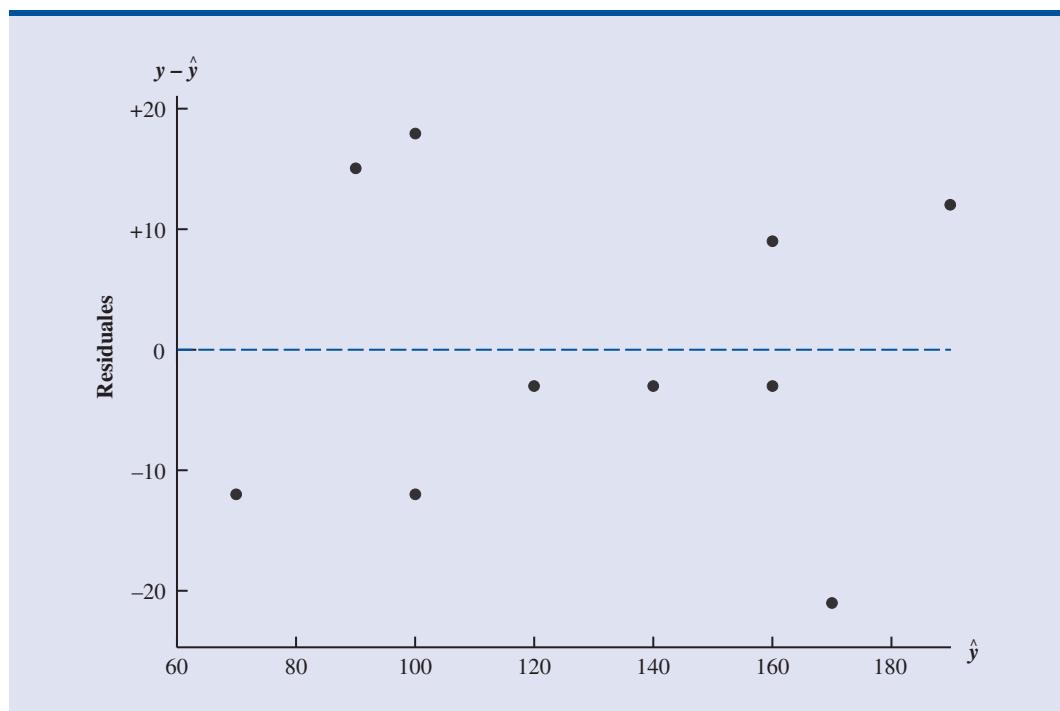
### Residuales estandarizados

Muchas de las gráficas de residuales que se obtienen con los paquetes de software utilizan una versión estandarizada de los residuales. Como se demostró en el capítulo anterior, una variable aleatoria se estandariza sustrayéndole su media y dividiendo el resultado entre su desviación es-

**FIGURA 14.12** GRÁFICAS DE LOS RESIDUALES CORRESPONDIENTES A TRES ESTUDIOS DE REGRESIONES



**FIGURA 14.13** GRÁFICA DE RESIDUALES CONTRA EL VALOR PRONOSTICADO  $\hat{y}$  OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



tándar. Cuando se emplea el método de mínimos cuadrados, la media de los residuales es cero. Por lo tanto, para obtener el **residual estandarizado** sólo es necesario dividir cada residual entre su desviación estándar.

Se puede demostrar que la desviación estándar del residual  $i$  depende del error estándar de estimación  $s$  y del valor correspondiente de la variable independiente  $x_i$ .

DESVIACIÓN ESTÁNDAR DEL RESIDUAL  $i^*$

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad (14.30)$$

donde

$$\begin{aligned} s_{y_i - \hat{y}_i} &= \text{desviación estándar del residual } i \\ s &= \text{error estándar de estimación} \\ h_i &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \end{aligned} \quad (14.31)$$

Obsérvese que la ecuación (14.30) indica que la desviación estándar del residual  $i$  depende de  $x_i$ , debido a la presencia de  $h_i$  en la fórmula.\*\* Una vez calculada la desviación estándar de cada uno de los residuales, se pueden calcular los residuales estandarizados dividiendo cada residual entre sus desviaciones estándar correspondientes.

\*En realidad, esta ecuación proporciona una estimación de la desviación estándar del residual  $i$  ya que se usa  $s$  en lugar de  $\sigma$ .

\*\* A  $h_i$  se le conoce como el influencial de la observación  $i$ . El influencial se verá en la sección 14.9 cuando se consideren las observaciones influyentes.

**TABLA 14.8** CÁLCULO DE LOS RESIDUALES ESTANDARIZADOS DEL EJEMPLO DE ARMAND'S PIZZA PARLORS

Restaurantes				$\frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$	$h_i$	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Residuales estandarizados
<i>i</i>	$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$					
1	2	-12	144	0.2535	0.3535	11.1193	-12	-1.0792
2	6	-8	64	0.1127	0.2127	12.2709	15	1.2224
3	8	-6	36	0.0634	0.1634	12.6493	-12	-0.9487
4	8	-6	36	0.0634	0.1634	12.6493	18	1.4230
5	12	-2	4	0.0070	0.1070	13.0682	-3	-0.2296
6	16	2	4	0.0070	0.1070	13.0682	-3	-0.2296
7	20	6	36	0.0634	0.1634	12.6493	-3	-0.2372
8	20	6	36	0.0634	0.1634	12.6493	9	0.7115
9	22	8	64	0.1127	0.2127	12.2709	-21	-1.7114
10	26	12	144	0.2535	0.3535	11.1193	12	1.0792
		Total	568					

Nota: En la tabla 14.7 se calculó el valor de los residuales.

#### RESIDUAL ESTANDARIZADO DE LA OBSERVACIÓN *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \quad (14.32)$$

En la tabla 14.8 se presentan los cálculos de los residuales estandarizados utilizando el ejemplo de Armand's Pizza Parlors. Recuérdese que ya en cálculos previos se obtuvo  $s = 13.829$ . La figura 14.14 es la gráfica de los residuales estandarizados contra la variable independiente  $x$ .

La gráfica de los residuales estandarizados permite ver si la suposición de que el término del error  $\epsilon$  tiene distribución normal es correcta. Si esta suposición se satisface debe parecer que la distribución de los residuales estandarizados, proviene de una distribución de probabilidad normal estándar.\* Por lo tanto, al observar la gráfica de los residuales estandarizados, se espera encontrar que aproximadamente 95% de los residuales estandarizados están entre -2 y +2. En la figura 14.14 se ve que en el ejemplo de Armand's todos los residuales estandarizados se encuentran entre -2 y +2. Por lo tanto, de acuerdo con los residuales estandarizados, esta gráfica no da razones para dudar de la suposición de que  $\epsilon$  tiene una distribución normal.

Debido al trabajo que significa calcular los valores estimados de  $\hat{y}$ , los residuales y los residuales estandarizados, la mayoría de los paquetes de software para estadística proporcionan, de manera opcional, estos datos como parte de los resultados de la regresión.

### Gráfica de probabilidad normal

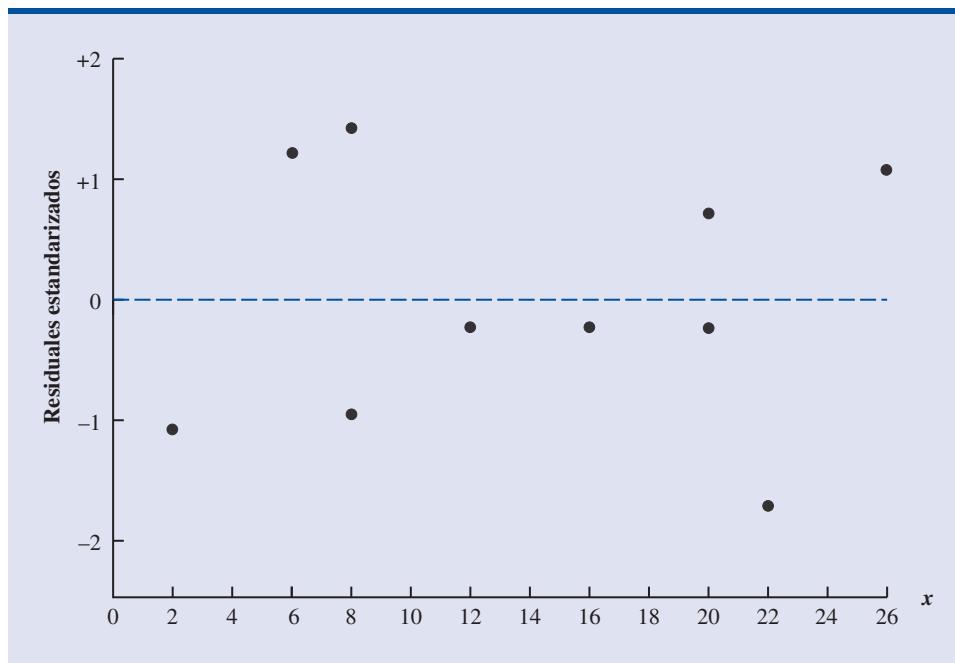
Otro manera de determinar la validez de la suposición de que el término del error tiene una distribución normal es la **gráfica de probabilidad normal**. Para mostrar cómo se elabora una gráfica de probabilidad normal, se introduce el concepto de *puntos normales*.

Supóngase que, de una distribución de probabilidad normal en la que la media es cero y la desviación estándar es uno, se toman aleatoriamente 10 valores; supóngase que este proceso de muestreo se repite una y otra vez y que los 10 valores de cada muestra se ordenan de menor a mayor. Por ahora, considérese únicamente el valor menor de cada muestra. A la variable aleato-

\*Como en la fórmula (14.30) se usa  $s$  en lugar de  $\sigma$ , la distribución de probabilidad de los residuales estandarizados no es técnicamente normal. Sin embargo, en la mayoría de los estudios de regresión, el tamaño de la muestra es suficientemente grande para que una aproximación normal sea muy buena.

Desviaciones pequeñas de la normalidad no tienen un efecto grande en las pruebas estadísticas empleadas en el análisis de regresión.

**FIGURA 14.14** GRÁFICA DE RESIDUALES ESTANDARIZADOS CONTRA LA VARIABLE INDEPENDIENTE  $x$ , OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS.



**TABLA 14.9**

PUNTOS NORMALES  
PARA  $n = 10$

Estadístico de orden	Punto normal
1	-1.55
2	-1.00
3	-0.65
4	-0.37
5	-0.12
6	0.12
7	0.37
8	0.65
9	1.00
10	1.55

**TABLA 14.10**

PUNTOS NORMALES Y  
RESIDUALES ORDE-  
NADOS DE ARMAND'S  
PIZZA PARLORS

Puntos normales	Residuales estandarizados ordenados
-1.55	-1.7114
-1.00	-1.0792
-0.65	-0.9487
-0.37	-0.2372
-0.12	-0.2296
0.12	-0.2296
0.37	0.7115
0.65	1.0792
1.00	1.2224
1.55	1.4230

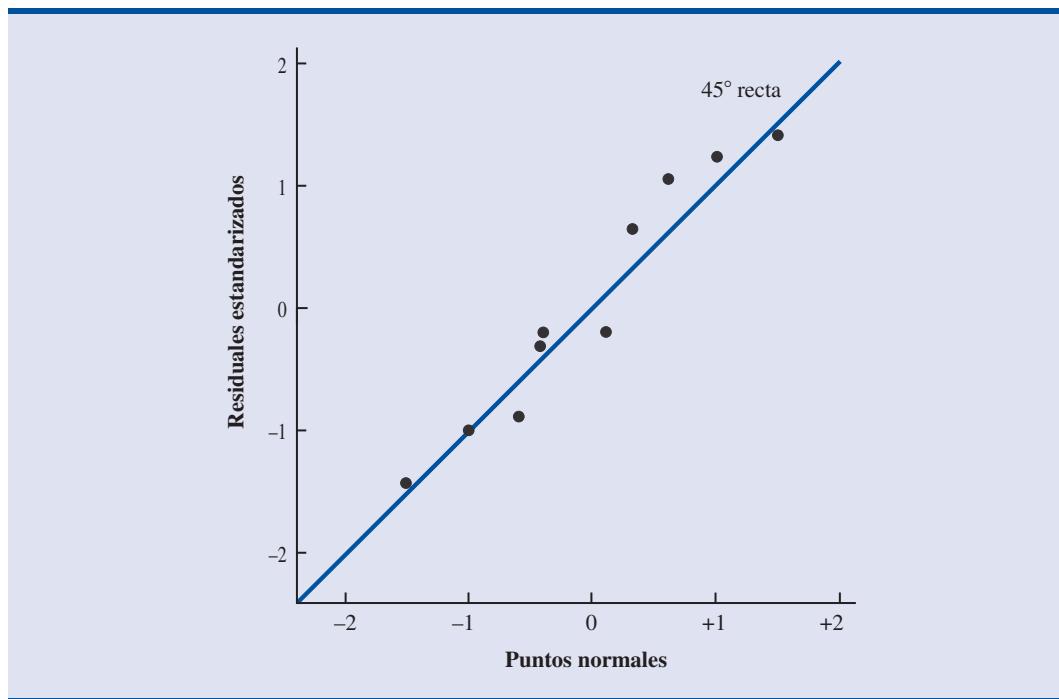
ria que representa el valor menor de estos varios muestreos se le conoce como el estadístico de primer orden.

En la ciencia de la estadística se ha demostrado que en muestras de tamaño 10 tomadas de una distribución de probabilidad normal estándar, el valor esperado del estadístico de primer orden es  $-1.55$ . A este valor esperado se le conoce como punto normal. En el caso de una muestra de tamaño  $n = 10$ , hay 10 estadísticos de orden y 10 puntos normales (ver tabla 14.9). En general, un conjunto de datos que conste de  $n$  observaciones tendrá  $n$  estadísticos de orden y por lo tanto  $n$  puntos normales.

A continuación se ve el uso de estos 10 puntos normales para determinar si parece ser que los residuales estandarizados de Armand's Pizza Parlors provengan de una distribución de probabilidad normal. Para empezar los 10 residuales estandarizados de la tabla 14.8 se ordenan. En la tabla 14.10 se presentan los 10 puntos normales y los residuales estandarizados normales. Si se satisface la suposición de normalidad, el menor residual estandarizado deberá tener un valor parecido al del menor punto normal, el siguiente residual estandarizado deberá tener un valor parecido al del siguiente punto normal, y así sucesivamente. En el caso de que los residuales estandarizados se encuentren distribuidos de una manera aproximadamente normal, en una gráfica en la que los puntos normales correspondan al eje horizontal y los correspondientes residuales estandarizados al eje vertical, los puntos de la gráfica estarán situados cercanos a una línea recta a 45 grados que pase por el origen. A esta gráfica es a lo que se le conoce como *gráfica de probabilidad normal*.

La figura 14.15 es la gráfica de probabilidad normal del ejemplo de Armand's Pizza Parlors. Para determinar si el patrón observado se desvía lo suficiente de la recta como para concluir que los residuales estandarizados no provienen de una distribución de probabilidad normal habrá que emplear el propio criterio. En la figura 14.15, todos los puntos se encuentran cerca de esta recta. Se concluye, por lo tanto, que la suposición de que los términos del error tienen una distribución de probabilidad normal es razonable. En general, entre más cerca de la recta a 45 grados se encuentren los puntos, más fuerte es la evidencia a favor de la suposición de normalidad. Cualquier curvatura sustancial en la gráfica de probabilidad normal es evidencia de que los residuales no provienen de una distribución de probabilidad normal. Tanto los puntos normales como la correspondiente gráfica de probabilidad normal pueden obtenerse fácilmente empleando paquetes como Minitab.

**FIGURA 14.15** GRÁFICA DE PROBABILIDAD NORMAL OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



### NOTAS Y COMENTARIOS

1. Las gráficas de residuales y de probabilidad normal se usan para confirmar las suposiciones de un modelo de regresión. Si en esta revisión se encuentra que una o más de las suposiciones son dudosas, habrá que considerar otro modelo o una transformación de los datos. Cuando se violan las suposiciones, las medidas a tomar deben basarse en un criterio adecuado; las recomendaciones de una persona con experiencia en estadística pueden ser útiles.
2. El análisis de residuales es el principal método estadístico para verificar si las suposiciones del

modelo de regresión son válidas. Aun cuando no se encuentre ninguna violación, esto no necesariamente implica que el modelo vaya a proporcionar buenas predicciones. Pero, si además existen otras pruebas estadísticas que favorezcan la conclusión de significancia y si el coeficiente de determinación es grande, deberá ser posible obtener buenas estimaciones y predicciones empleando la ecuación de regresión estimada.

### Ejercicios

#### Métodos

45. Dados los datos de las dos variables  $x$  y  $y$ .

$x_i$	6	11	15	18	20
$y_i$	6	8	12	20	30

- a. A partir de estos datos obtenga una ecuación de regresión estimada.
- b. Calcule los residuales.
- c. Trace una gráfica de residuales contra la variable independiente  $x$ . ¿Parecen satisfacerse las suposiciones acerca de los términos del error?

- d. Calcule los residuales estandarizados.  
 e. Elabore una gráfica de residuales estandarizados contra  $\hat{y}$ . ¿Qué conclusión puede sacar de esta gráfica?
46. En un estudio de regresión se emplearon los datos siguientes.

Observación	$x_i$	$y_i$	Observación	$x_i$	$y_i$
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

- a. A partir de estos datos obtenga una ecuación de regresión estimada.  
 b. Trace una gráfica de residuales. ¿Parecen satisfacerse las suposiciones del término del error?

## Aplicaciones

**Autoexamen**

47. A continuación se presentan datos sobre los gastos en publicidad y los ingresos (en miles de dólares) del restaurante Cuatro Estaciones.

Gastos en publicidad	Ingresos
1	19
2	32
4	44
6	40
10	52
14	53
20	54

- a. Sea  $x$  igual gastos en publicidad y  $y$  igual a ingresos. Utilice el método de mínimos cuadrados para obtener una línea recta que aproxime la relación entre las dos variables.  
 b. Empleando como nivel de significancia 0.05, pruebe si los ingresos y los gastos en publicidad están relacionados.  
 c. Elabore una gráfica de residuales de  $y - \hat{y}$  contra  $\hat{y}$ . Use el resultado del inciso a) para obtener los valores de  $\hat{y}$ .  
 d. ¿Qué conclusiones se pueden sacar del análisis de residuales? ¿Se puede usar este modelo o se debe buscar uno mejor?
48. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relaciona los años de experiencia con las ventas anuales.
- Calcule los residuales y trace una gráfica de residuales para este problema.
  - A la luz de la gráfica de residuales, ¿parecen razonables las suposiciones acerca de los términos del error?
49. American Depository Receipts (ADR) son certificados que cotizan en la bolsa de Nueva York y que representan acciones de empresas extranjeras que mantienen un depósito en un banco de su propio país. En la tabla siguiente se presenta la relación precio/ganancia (P/G) y el porcentaje de rendimiento de la inversión (ROE, por sus siglas en inglés), de 10 empresas hindúes que es probable que sean nuevos (*Bloomberg Personal Finance*, abril 2000).

	ROE	P/G
Bharti Televentures	6.43	36.88
Gujarat Ambuja Cements	13.49	27.03
Hindalco Industries	14.04	10.83
ICICI	20.67	5.15
Mahanagar Telephone Nigam	22.74	13.35
NIIT	46.23	95.59
Pentamedia Graphics	28.90	54.85
Satyam Computer Services	54.01	189.21
Silverline Technologies	28.02	75.86
Videsh Sanchar Nigam	27.04	13.17

- Emplee un paquete de software para obtener una ecuación de regresión estimada que relacione  $y = P/G$  y  $x = ROE$ .
- Construya una gráfica de residuales contra la variable independiente.
- A la luz de la gráfica de residuales, ¿parecen razonables las suposiciones acerca de los términos del error y de la forma del modelo?

14.9

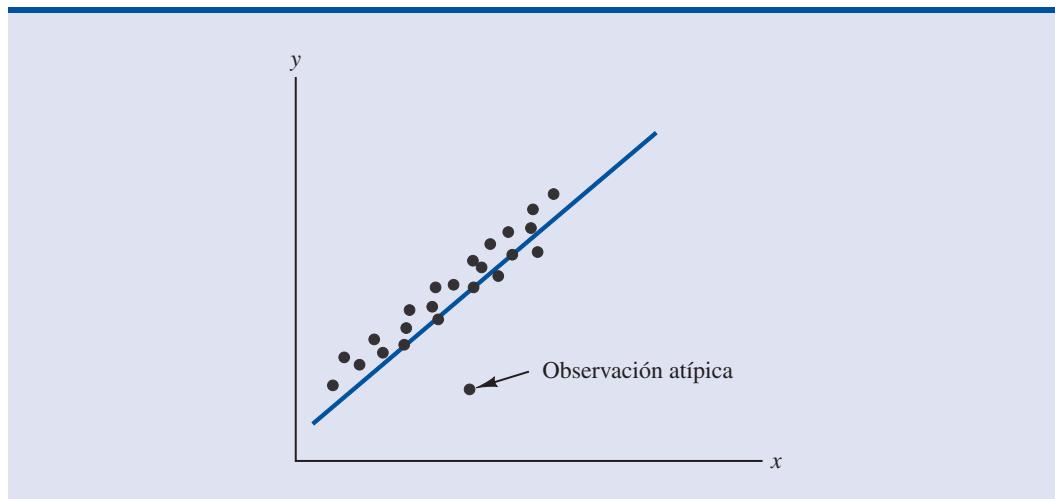
## Análisis de residuales: observaciones atípicas y observaciones influyentes

En la sección 14.8 se mostró cómo emplear el análisis de residuales para determinar violaciones a las suposiciones del modelo de regresión. En esta sección se ve el uso del análisis de residuales para identificar observaciones que se pueden clasificar como observaciones atípicas o como observaciones especialmente influyentes sobre la ecuación de regresión estimada. También se discuten algunas de las medidas que han de tomarse cuando se presentan tales observaciones.

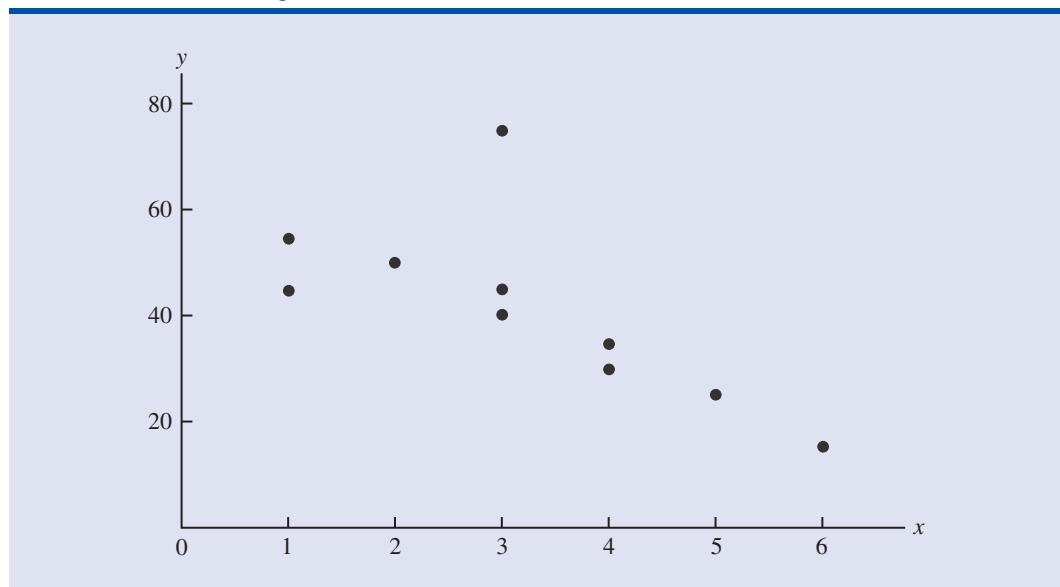
### Detección de observaciones atípicas

La figura 14.16 es un diagrama de dispersión de un conjunto de datos que contiene una **observación atípica**, un dato (una observación) que no sigue la tendencia del resto de los datos. Las observaciones atípicas son observaciones que son sospechosas y que requieren un análisis cuidado.

**FIGURA 14.16 UN CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA**



**FIGURA 14.17** DIAGRAMA DE DISPERSIÓN DE UN CONJUNTO DE DATOS EN EL QUE HAY UNA OBSERVACIÓN ATÍPICA



**TABLA 14.11**

CONJUNTO  
DE DATOS PARA  
ILUSTRAR EL  
EFFECTO DE UNA  
OBSERVACIÓN ATÍPICA

$x_i$	$y_i$
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

doso. Puede tratarse de datos erróneos; si es así, esos datos deben ser corregidos. Puede tratarse de una violación a las suposiciones del modelo; si es así, habrá que considerar otro modelo. Por último, puede tratarse, simplemente, de valores inusuales que se presenten por casualidad. En ese caso, esos valores deberán conservarse.

Para mostrar cómo se detectan las observaciones atípicas, considérense los datos de la tabla 14.11; la figura 14.17 muestra el diagrama de dispersión de estos datos. Con excepción de la observación 4 ( $x_4 = 3, y_4 = 75$ ), estos datos parecen seguir un patrón que indicar una relación lineal negativa. En efecto, dado el patrón que parece seguir el resto de los datos, se esperaría que  $y_4$  fuera mucho más pequeño, por lo que a esta observación se le considera como un dato atípico. En el caso de la regresión lineal simple, las observaciones atípicas pueden detectarse mediante un simple examen del diagrama de dispersión.

Para detectar observaciones atípicas también se pueden usar los residuales estandarizados. Si una observación se aleja mucho del patrón del resto de los datos (por ejemplo, la observación atípica de la figura 14.16), el valor absoluto del correspondiente residual estandarizado será grande. Muchos paquetes de software identifican de manera automática las observaciones cuyos residuales tienen un valor absoluto grande. En la figura 14.18 se presentan los resultados dados por Minitab para el análisis de regresión de los datos de la tabla 14.11. En el penúltimo renglón de los resultados dados por Minitab se lee que el residual estandarizado de la observación 4 es 2.67. Minitab identifica como una observación inusual toda observación cuyo residual estandarizado sea menor a  $-2$  o mayor a  $+2$ ; en tales casos la observación aparece en un renglón aparte con una R al lado del residual estandarizado, como se observa en la figura 14.18. Si los errores están distribuidos normalmente, sólo 5% de los residuales estandarizados se encontrarán fuera de estos límites.

Para decidir qué hacer con una observación atípica, primero hay que verificar si es una observación correcta. Puede ser que se trate de un error al anotar los datos o al ingresarlos a la computadora. Supóngase, por ejemplo, que al verificar la observación atípica de la tabla 14.17, se encuentra que hubo un error; el valor correcto de la observación 4 era  $x_4 = 3, y_4 = 30$ . En la figura 14.19 se presenta el resultado que proporciona Minitab una vez corregido el valor de  $y_4$ . Se observa que el dato incorrecto afecta sustancialmente la bondad de ajuste. Con el dato correcto, el valor de R-sq aumenta de 49.7% a 83.8% y el valor de  $b_0$  disminuye de 64.958 a 59.237. La pendiente de la recta cambia de  $-7.33$  a  $-6.949$ . La identificación de los datos atípicos permite corregir errores en los datos y mejora los resultados de la regresión.

**FIGURA 14.18** RESULTADOS QUE DA MINITAB PARA EL ANÁLISIS DE REGRESIÓN DEL CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA

```
The regression equation is
y = 65.0 - 7.33 x

Predictor      Coef    SE Coef      T      p
Constant      64.958    9.258     7.02  0.000
X             -7.331    2.608    -2.81  0.023

S = 12.67    R-sq = 49.7%   R-sq(adj) = 43.4%

Analysis of Variance

SOURCE        DF        SS        MS        F      p
Regression    1  1268.2  1268.2  7.90  0.023
Residual Error 8  1284.3   160.5
Total         9  2552.5

Unusual Observations
Obs      x      y      Fit    SE Fit  Residual  St Resid
4     3.00  75.00  42.97     4.04     32.03     2.67R

R denotes an observation with a large standardized residual.
```

**FIGURA 14.19** RESULTADOS QUE DA MINITAB PARA EL CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA YA CORREGIDA

```
The regression equation is
Y = 59.2 - 6.95 X

Predictor      Coef    SE Coef      T      p
Constant      59.237   3.835    15.45  0.000
X             -6.949   1.080    -6.43  0.000

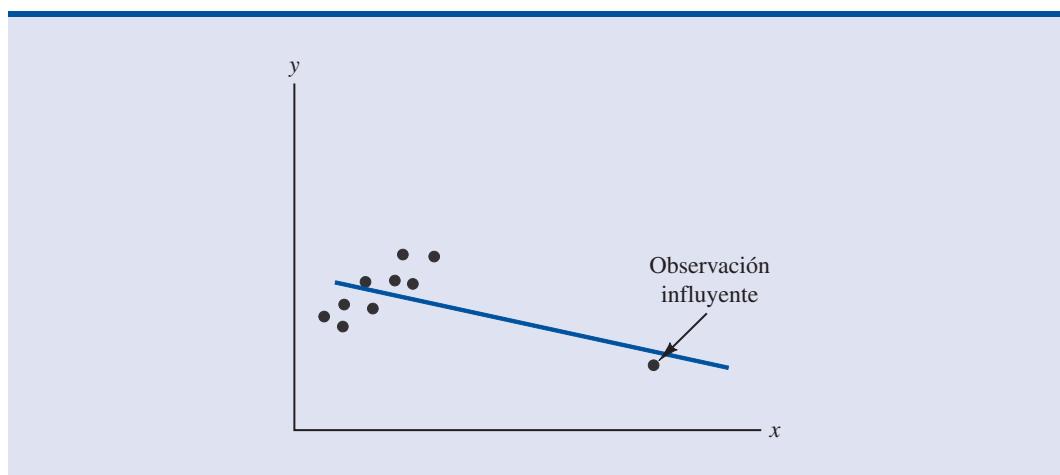
S = 5.248    R-sq = 83.8%   R-sq(adj) = 81.8%

Analysis of Variance

SOURCE        DF        SS        MS        F      p
Regression    1  1139.7  1139.7  41.38  0.000
Residual Error 8   220.3   27.5
Total         9  1360.0
```

## Detección de observaciones influyentes

Algunas veces una o más de las observaciones tienen una influencia fuerte sobre los resultados que se obtienen. En la figura 14.20 se muestra un ejemplo de una **observación influyente** en una regresión lineal simple. La recta de regresión estimada tiene pendiente negativa, pero si la observación influyente se elimina del conjunto de datos, la pendiente de la recta de regresión estimada cambia de negativa a positiva y la intersección con el eje y es menor. Es claro que esta sola

**FIGURA 14.20 CONJUNTO DE DATOS CON UNA OBSERVACIÓN INFLUYENTE**

observación tiene mucha más influencia sobre la recta de regresión estimada que cualquiera otra observación; el efecto que tiene la eliminación de cualquiera de las otras observaciones sobre la ecuación de regresión estimada es muy pequeño.

Cuando sólo se tiene una variable independiente, las observaciones influyentes pueden identificarse mediante un diagrama de dispersión. Una observación influyente puede ser una observación atípica (una observación cuyo valor de  $y$  se desvía sustancialmente de la tendencia general), puede ser un valor de  $x$  muy alejado de la media (por ejemplo, ver la figura 14.20) o puede tratarse de la combinación de estas dos cosas (un valor de  $y$  algo fuera de la tendencia y un valor de  $x$  un poco extremo).

Las observaciones influyentes deben examinarse cuidadosamente dado el gran efecto que tienen sobre la ecuación de regresión estimada. Lo primero que hay que hacer es verificar que no se haya cometido algún error al recolectar los datos. Si se cometió algún error, se corrige y se obtiene una nueva ecuación de regresión estimada. Si la observación es correcta, puede uno considerarse afortunado de tenerla. Tal dato, cuando es correcto, contribuye a una mejor comprensión del modelo adecuado y conduce a una mejor ecuación de regresión estimada. En la figura 14.20, la presencia de la observación influyente, si es correcta, llevará a tratar de obtener datos con valores  $x$  intermedios, que permitan comprender mejor la relación entre  $x$  y  $y$ .

Las observaciones en las que la variable independiente toma valores extremos se denominan **datos (puntos, observaciones) de gran influencia**. La observación influyente de la figura 14.20 es un punto de gran influencia. La influencia de una observación depende de qué tan lejos está el valor de la variable independiente de su media. En el caso de una sola variable independiente, la influencia (*leverage*) de la observación  $i$ , que se denota  $h_i$  se calcula mediante la ecuación (14.33).

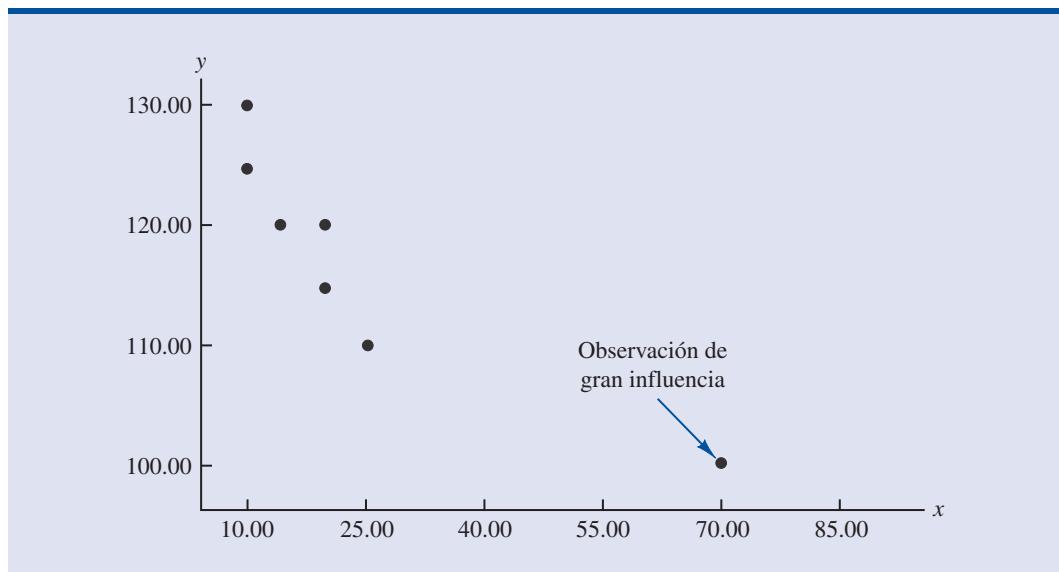
#### INFLUENCIA DE LA OBSERVACIÓN $i$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \quad (14.33)$$

De acuerdo con esta fórmula, es claro que entre más alejada se encuentre  $x_i$  de su media  $\bar{x}$  mayor será la influencia (leverage) de la observación  $i$ .

Muchos de los paquetes para estadística identifican automáticamente, como parte de los resultados estándar de regresión, los puntos de gran influencia. Para ilustrar cómo identifica Minitab los puntos de gran influencia, se considerará el conjunto de datos de la tabla 14.12.

**FIGURA 14.21** DIAGRAMA DE DISPERSIÓN DEL CONJUNTO DE DATOS CON UN DATO DE GRAN INFLUENCIA



**TABLA 14.12**

CONJUNTO  
DE DATOS CON UNA  
OBSERVACIÓN DE  
GRAN INFLUENCIA

$x_i$	$y_i$
10	125
10	130
15	120
20	115
20	120
25	110
70	100

Los paquetes de software son esenciales para hacer los cálculos que permiten determinar las observaciones influyentes. Aquí se discute la regla de selección que emplea Minitab.

Observando la figura 14.21, que es el diagrama de dispersión del conjunto de datos presentado en la tabla 14.12, se ve que la observación 7 ( $x = 70, y = 100$ ) es una observación en la que el valor de  $x$  es un valor extremo. Por lo tanto, es de esperarse que sea identificado como un punto de gran influencia. La influencia de esta observación se calcula usando la ecuación (14.33).

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = 0.94$$

En el caso de la regresión lineal simple, Minitab identifica como observaciones de gran influencia las observaciones para las que  $h_i > 6/n$  o  $h_i = 0.99$ , lo que sea menor. En el conjunto de datos de la tabla 14.12,  $6/n = 6/7 = 0.86$ . Como  $h_7 = 0.94 > 0.86$ . Minitab identificará la observación 7 como una observación cuyo valor  $x$  tiene una gran influencia. En la figura 14.22 se presenta el resultado que da Minitab del análisis de regresión de este conjunto de datos. A la observación 7 ( $x = 70, y = 100$ ) la identifica como una observación de gran influencia; esta observación la presenta en un renglón aparte en la parte inferior de los resultados con una X en el margen derecho.

Las observaciones influyentes debidas a la interacción de una observación de gran influencia y de residuales grandes, suelen ser difíciles de detectar. Existen procedimientos de diagnóstico que para determinar si una observación es influyente toman en cuenta ambas cosas. En el capítulo 15 se estudiará uno de estos procedimientos, el estadístico  $D$  de Cook.

### NOTAS Y COMENTARIOS

Una vez identificada una observación como potencialmente influyente, debido a que tiene un residual grande o por ser de gran influencia, su impacto sobre la ecuación de regresión estimada debe ser evaluado. En textos más avanzados se presentan los métodos de diagnóstico apropiados.

Pero, cuando no se está familiarizado con el material más avanzado, una manera sencilla de hacer este diagnóstico es realizar el análisis de regresión con y sin esa observación. Este método permite apreciar la influencia que tiene la observación potencialmente influyente sobre el resultado.

**FIGURA 14.22** RESULTADO DE MINITAB EMPLEANDO EL CONJUNTO DE DATOS CON UNA OBSERVACIÓN DE GRAN INFLUENCIA

The regression equation is $y = 127 - 0.425 x$
Predictor      Coef    SE Coef      T      p
Constant      127.466    2.961    43.04    0.000
X              -0.42507    0.09537    -4.46    0.007
S = 4.883    R-sq = 79.9%    R-sq(adj) = 75.9%
Analysis of Variance
SOURCE      DF      SS      MS      F      p
Regression      1      473.65    473.65    19.87    0.007
Residual Error    5      119.21    23.84
Total          6      592.86
Unusual Observations
Obs      x      y      Fit    SE Fit    Residual    St Resid
7      70.0    100.00    97.71    4.73      2.29      1.91 X
X denotes an observation whose X value gives it large influence.

## Ejercicios

### Métodos

**Autoexamen**

50. Considérense los datos siguientes para las variables  $x$  y  $y$ .

$x_i$	135	110	130	145	175	160	120
$y_i$	145	100	120	120	130	130	110

- a. Calcule los residuales estandarizados de estos datos. ¿Hay entre los datos alguna observación atípica? Explique.
  - b. Haga una gráfica de residuales estandarizados contra  $\hat{y}$ . ¿Se observa en esta gráfica la presencia de alguna observación atípica?
  - c. Con estos datos elabore un diagrama de dispersión. ¿Se observa en el diagrama de dispersión la presencia de alguna observación atípica? En general, ¿qué consecuencias tienen, para la regresión lineal simple, estos hallazgos?
51. Considérense los datos siguientes para las variables  $x$  y  $y$ .

$x_i$	4	5	7	8	10	12	12	22
$y_i$	12	14	16	15	18	20	24	19

- a. Calcule los residuales estandarizados de estos datos. ¿Hay entre los datos alguna observación atípica? Explique.
- b. Calcule las observaciones de influencia que haya en estos datos. Entre estos datos, ¿parece haber alguna observación influyente? Explique.
- c. Con estos datos elabore un diagrama de dispersión. ¿Se observa en el diagrama de dispersión la presencia de alguna observación atípica? Explique.

## Aplicaciones

52. Los datos siguientes muestran los gastos (en millones de \$) y los envíos en bbls. (millones) de 10 importantes marcas de cerveza.

Marca	Gastos medios (millones de \$)	Envío
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

- a. Con estos datos obtenga una ecuación de regresión estimada.  
 b. Emplee el análisis residual para hallar observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.
53. Los especialistas en salud recomiendan que las personas que corren tomen unos 200 ml de agua cada 15 minutos mientras están corriendo. Las personas que corren de tres a ocho horas, requieren sistemas de hidratación que se llevan sobre la cintura o sobre la espalda. En los datos a continuación se da el volumen (en onzas fluidas, 1 oz. flu = 30 ml aprox.) y el precio de 26 sistemas de hidratación que se llevan sobre la cintura o sobre la espalda (*Trail Runner Gear Guide*, 2003).

Modelo	Volumen (oz fl)	Precio (\$)
Fastdraw	20	10
Fastdraw Plus	20	12
Fitness	20	12
Access	20	20
Access Plus	24	25
Solo	20	25
Serenade	20	35
Solitaire	20	35
Gemini	40	45
Shadow	64	40
SipStream	96	60
Express	20	30
Lightning	28	40
Elite	40	60
Extender	40	65
Stinger	32	65
GelFlask Belt	4	20
GelDraw	4	7
GelFlask Clip-on Holster	4	10
GelFlask Holster SS	4	10
Strider (W)	20	30
Walkabout (W)	230	40
Solitude I.C.E.	20	35
Getaway I.C.E.	40	55
Profile I.C.E.	64	50
Traverse I.C.E.	64	60

archivo en CD  
Beer

archivo en CD  
Hydration2

- Obtenga la ecuación de regresión estimada que sirva para predecir el precio de un sistema de hidratación, dado su volumen.
  - Use el análisis residual para determinar si hay observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.
54. En la tabla siguiente se presenta la capitalización de mercado y los salarios del presidente del consejo de administración (CEO, por sus siglas en inglés) de 20 empresas (*The Wall Street Journal*, 24 de febrero de 2000 y 6 de abril de 2000).

archivo  
en  
CD  
CEO

	Capitalización de mercado (millones de \$)	Salario del CEO (miles de \$)
Anheuser-Busch	32 977.4	1130
AT&T	162 365.1	1400
Charles Schwab	31 363.8	800
Chevron	56 849.0	1350
DuPont	68 848.0	1000
General Electric	507 216.8	3325
Gillette	44 180.1	978
IBM	194 455.9	2000
Johnson & Johnson	143 131.0	1365
Kimberly-Clark	35 377.5	950
Merrill Lynch	31 062.1	700
Motorola	92 923.7	1275
Philip Morris	54 421.2	1625
Procter & Gamble	144 152.9	1318.3
Qualcomm	116 840.8	773
Schering-Plough	62 259.4	1200
Sun Microsystems	120 966.5	116
Texaco	30 040.7	950
USWest	36 450.8	897
Walt Disney	61 288.1	750

- Obtenga la ecuación de regresión estimada para predecir el salario del CEO dada la capitalización de mercado.
- Use el análisis de residuales para determinar si hay observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.

## Resumen

En este capítulo se mostró el uso del análisis de regresión para determinar cómo es la relación entre una variable dependiente  $y$  y una variable independiente  $x$ . En la regresión lineal simple, el modelo de regresión es  $y = \beta_0 + \beta_1 x + \epsilon$ . La ecuación de regresión lineal simple  $E(y) = \beta_0 + \beta_1 x$  describe la relación de la media o valor esperado de  $y$  con  $x$ . Para obtener la ecuación de regresión estimada  $\hat{y} = b_0 + b_1 x$  se emplearon datos muestrales y el método de mínimos cuadrados. En efecto,  $b_0$  y  $b_1$  son estadísticos muestrales que se usan para estimar los parámetros desconocidos del modelo,  $\beta_0$  y  $\beta_1$ .

El coeficiente de determinación se presentó como una medida de la bondad de ajuste de la ecuación de regresión estimada; el coeficiente de determinación se puede interpretar como la proporción de la variación en la variable dependiente que puede ser explicada por la ecuación de regresión estimada. Se volvió a ver la correlación como una medida descriptiva de la intensidad de la relación lineal entre las dos variables.

Se discutieron las suposiciones acerca del modelo de regresión y del correspondiente término del error, y se presentaron las pruebas  $t$  y  $F$ , basadas en esas suposiciones, como un medio para determinar si la relación entre las dos variables es estadísticamente significativa. Se mostró

cómo usar la ecuación de regresión estimada para obtener estimaciones por medio de intervalos de confianza para el valor medio de  $y$  y estimaciones por medio de intervalos de predicción para valores individuales de  $y$ .

El capítulo concluyó con una sección sobre soluciones por computadora de los problemas de regresión y dos secciones sobre el uso del análisis residual para verificar las suposiciones del modelo e identificar las observaciones atípicas e influyentes.

## Glosario

**Variable dependiente** La variable que se predice o explicada. Se denota  $y$ .

**Variable independiente** Variable que predice o explica. Se denota  $x$ .

**Regresión lineal simple** Análisis de regresión en el que participan una variable independiente y una variable dependiente, y en el que la relación entre estas variables se aproxima mediante una línea recta.

**Modelo de regresión** Ecuación que describe cómo están relacionadas  $y$  y  $x$ , más un término del error; en la regresión lineal simple, el modelo de regresión es  $y = \beta_0 + \beta_1x + \epsilon$ .

**Ecuación de regresión** Ecuación que describe cómo está relacionada la media o valor esperado de la variable dependiente con la variable independiente; en la regresión lineal simple,  $E(y) = \beta_0 + \beta_1x$ .

**Ecuación de regresión estimada** Estimación de la ecuación de regresión obtenida a partir de datos muestrales, empleando el método de mínimos cuadrados. En la regresión lineal simple, la ecuación de regresión estimada es  $\hat{y} = b_0 + b_1x$ .

**Método de mínimos cuadrados** Procedimiento empleado para obtener la ecuación de regresión estimada. El objetivo es minimizar  $\sum(y_i - \hat{y}_i)^2$ .

**Diagrama de dispersión** Gráfica de datos bivariados en la que la variable independiente va en el eje horizontal y la variable dependiente va en el eje vertical.

**Coefficiente de determinación** Medida de la bondad de ajuste de la ecuación de regresión estimada. Se puede interpretar como la proporción de la variabilidad de la variable dependiente y que es explicada por la ecuación de regresión estimada.

**Residual  $i$**  Diferencia que existe entre el valor observado de la variable dependiente y el valor pronosticado empleando la ecuación de regresión estimada; para la observación  $i$ , el residual  $i$  es  $y_i - \hat{y}_i$ .

**Coefficiente de correlación** Medida de la intensidad de la relación lineal entre dos variables (ya visto en el capítulo 3).

**Error cuadrado medio** Estimación insesgada de la varianza del término del error  $\sigma^2$ . Se denota ECM o  $s^2$ .

**Error estándar de estimación** Raíz cuadrada del error cuadrado medio, se denota  $s$ . Es una estimación de  $\sigma$ , la desviación estándar del error.

**Tabla ANOVA** En el análisis de varianza, tabla que se usa para resumir los cálculos necesarios en la prueba  $F$  de significancia.

**Intervalo de confianza** Estimación por intervalo del valor medio de  $y$  para un valor dado de  $x$ .

**Intervalo de predicción** Estimación por intervalo de un solo valor de  $y$  para un valor dado de  $x$ .

**Análisis residual** Análisis de los residuales que se usa para determinar si parecen ser válidas las suposiciones hechas acerca del modelo de regresión. El análisis de residuales también se usa para identificar observaciones atípicas y observaciones influyentes.

**Gráfica de residuales** Representación gráfica de los residuales, se usa para determinar si parecen ser válidas las suposiciones hechas acerca del modelo de regresión.

**Residual estandarizado** Valor obtenido al dividir un residual entre su desviación estándar.

**Gráfica de probabilidad normal** Gráfica en la que los residuales estandarizados se grafican contra los puntos normales. Esta gráfica ayuda a determinar si parece ser válida la suposición de que los términos del error tienen una distribución de probabilidad normal.

**Observación atípica** Dato u observación que no sigue la tendencia del resto de los datos.

**Observación influyente** Observación en la que la variable independiente tiene un valor extremo.

**Puntos de gran influencia** Observaciones en las que la variable independiente tiene valores extremos.

## Fórmulas clave

### Modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon \quad (14.1)$$

### Ecuación de regresión lineal simple

$$E(y) = \beta_0 + \beta_1 x \quad (14.2)$$

### Ecuación de regresión lineal simple estimada

$$\hat{y} = b_0 + b_1 x \quad (14.3)$$

### Criterio de mínimos cuadrados

$$\min \sum (y_i - \hat{y}_i)^2 \quad (14.5)$$

### Intersección con el eje y y pendiente de la ecuación de regresión lineal simple

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.7)$$

### Suma de cuadrados debidos al error

$$SCE = \sum (y_i - \hat{y}_i)^2 \quad (14.8)$$

### Suma de cuadrados total

$$STC = \sum (y_i - \bar{y})^2 \quad (14.9)$$

### Suma de cuadrados debida a la regresión

$$SCR = \sum (\hat{y}_i - \bar{y})^2 \quad (14.10)$$

### Relación entre STC, SCR y SCE

$$STC = SCR + SCE \quad (14.11)$$

### Coeficiente de determinación

$$r^2 = \frac{SCR}{STC} \quad (14.12)$$

### Coeficiente de correlación muestral

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1) \sqrt{\text{Coeficiente de determinación}} \\ &= (\text{signo de } b_1) \sqrt{r^2} \end{aligned} \quad (14.13)$$

**Error cuadrado medio (estimación de  $\sigma^2$ )**

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2} \quad (14.15)$$

**Error estándar de estimación**

$$s = \sqrt{\text{CME}} = \sqrt{\frac{\text{SCE}}{n - 2}} \quad (14.16)$$

**Desviación estándar de  $b_1$**

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} \quad (14.17)$$

**Desviación estándar estimada de  $b_1$**

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \quad (14.18)$$

**Estadístico de prueba  $t$**

$$t = \frac{b_1}{s_{b_1}} \quad (14.19)$$

**Regresión cuadrática media**

$$\text{CMR} = \frac{\text{SCR}}{\text{Número de variables independientes}} \quad (14.20)$$

**Estadístico de prueba  $F$**

$$F = \frac{\text{CMR}}{\text{CME}} \quad (14.21)$$

**Desviación estándar estimada de  $\hat{y}_p$**

$$s_{\hat{y}_p} = s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (14.23)$$

**Intervalo de confianza para  $E(y_p)$**

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \quad (14.24)$$

**Desviación estándar estimada para un solo valor**

$$s_{\text{ind}} = s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (14.26)$$

**Intervalo de predicción para  $y_p$**

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \quad (14.27)$$

**Residual de la observación  $i$**

$$y_i - \hat{y}_i \quad (14.28)$$

**Desviación estándar del residual  $i$** 

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad (14.30)$$

**Residual estandarizado de la observación  $i$** 

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \quad (14.32)$$

**Influencia de la observación**

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \quad (14.33)$$

**Ejercicios complementarios**

55. Si el valor de  $r^2$  es elevado, ¿implica eso que entre las dos variables hay una relación de causa y efecto?
56. Explique con sus propias palabras la diferencia entre estimación por intervalo del valor medio de las  $y$  para un valor dado de  $x$  y estimación por intervalo de un valor de  $y$  para una  $x$  dada.
57. ¿Qué objeto tiene probar si  $\beta_1 = 0$ ? Si se rechaza que  $\beta_1 = 0$ , ¿significa eso un buen ajuste?
58. En la tabla siguiente se da el número de acciones vendidas (en millones) y el precio esperado (el promedio del precio mínimo y del precio máximo) de 10 acciones de oferta pública inicial.

Empresa	Acciones vendidas	Precio esperado (\$)
American Physician	5.0	15
Apex Silver Mines	9.0	14
Dan River	6.7	15
Franchise Mortgage	8.75	17
Gene Logic	3.0	11
International Home Foods	13.6	19
PRT Group	4.6	13
Rayovac	6.7	14
RealNetworks	3.0	10
Software AG Systems	7.7	13

- a. Obtenga la ecuación de regresión estimada en la que la cantidad de acciones vendidas sea la variable independiente y el precio la variable dependiente.
- b. Empleando 0.05 como nivel de significancia, ¿existe una relación significativa entre las dos variables?
- c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
- d. Empleando la ecuación de regresión estimada, estime el precio esperado en una empresa que considera una oferta pública inicial de 6 millones de acciones.
59. Los programas de recompra de acciones corporativas, suelen promoverse como un beneficio para los accionistas. Pero Robert Gabele, director de investigación interna de First Call/Thomson Financial, hizo notar que muchos de estos programas se realizan únicamente con el objetivo de obtener acciones que se emplean como opciones como incentivo para los altos directivos de la empresa. En todas las empresas, las opciones de acciones existentes en 1998 representaban el 6.2 por ciento de todas las acciones comunes en circulación. En los datos siguientes se da la cantidad de opciones otorgadas y la cantidad de acciones en circulación de 13 empresas (*Bloomberg Personal Finance*, enero/febrero, 2000)

	Opciones otorgadas en circulación (en millones)	Acciones comunes en circulación (en millones)
Adobe Systems	20.3	61.8
Apple Computer	52.7	160.9
Applied Materials	109.1	375.4
Autodesk	15.7	58.9
Best Buy	44.2	203.8
Fruit of the Loom	14.2	66.9
ITT Industries	18.0	87.9
Merrill Lynch	89.9	365.5
Novell	120.2	335.0
Parametric Technology	78.3	269.3
Reebok International	12.8	56.1
Silicon Graphics	52.6	188.8
Toys "R" Us	54.8	247.6

- Obtenga una ecuación de regresión estimada que sirva para estimar la cantidad en circulación de opciones otorgadas dada la cantidad de acciones comunes en circulación.
  - Emplee la ecuación de regresión estimada para estimar la cantidad en circulación de opciones otorgadas por una empresa que tiene 150 millones de acciones comunes en circulación.
  - ¿Cree que la ecuación de regresión estimada proporcione una buena predicción de la cantidad en circulación de opciones otorgadas? Emplee  $r^2$  para justificar su respuesta.
60. El promedio industrial Dow Jones (DJIA) y el Estándar & Poor's 500 (S & P) son índices que se emplean como una medida del movimiento general del mercado de valores. El DJIA se basa en los movimientos de los precios de 30 empresas grandes; el S&P 500 es un índice compuesto de 500 acciones. Algunos dicen que el S&P 500 es una mejor medida de la actividad del mercado de valores porque tiene una base más amplia. A continuación se presenta el precio de cierre del DJIA y del S&P 500 durante 20 semanas a partir del 9 de septiembre del 2005 (*Borron's*, 30 de enero de 2006).

Fecha	DJIA	S&P 500
9 de septiembre	10 679	1241
16 de septiembre	10 642	1238
23 de septiembre	10 420	1215
30 de septiembre	10 569	1229
7 de octubre	10 292	1196
14 de octubre	10 287	1187
21 de octubre	10 215	1180
28 de octubre	10 403	1198
4 de noviembre	10 531	1220
11 de noviembre	10 686	1235
18 de noviembre	10 766	1248
25 de noviembre	10 932	1268
2 de diciembre	10 878	1265
9 de diciembre	10 779	1259
16 de diciembre	10 876	1267
23 de diciembre	10 883	1269
30 de diciembre	10 718	1248
6 de enero	10 959	1285
13 de enero	10 960	1288
20 de enero	10 667	1261

- Dé el diagrama de dispersión de estos datos empleando DJIA como variable independiente.
  - Obtenga la ecuación de regresión estimada.
  - Pruebe la significancia de la relación. Use  $\alpha = 0.05$ .
  - ¿Proporciona un buen ajuste la ecuación de regresión estimada? Explique.
  - Suponga que el precio de cierre del DJIA es 11 000. Estime el precio de cierre del S&P 500.
  - ¿Debe preocupar que el valor de 11 000 del DJIA empleado en el inciso e) para predecir el del S&P 500 se encuentre fuera del intervalo de los datos empleado para obtener la ecuación de regresión estimada?
61. Jensen Tire & Auto está por decidir si firma un contrato de mantenimiento para su nueva máquina de alineamiento y balanceo de neumáticos. Los gerentes piensan que los gastos de mantenimiento deberán estar relacionados con el uso y recolectan los datos siguientes sobre uso semanal (horas) y gastos anuales de mantenimiento (en cientos de dólares).

archivo  
en  
**CD**  
Jensen

Uso semanal (horas)	Gastos anuales de mantenimiento
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

- Obtenga la ecuación de regresión estimada que relaciona gastos anuales de mantenimiento con el uso semanal.
  - Pruebe la significancia de la relación del inciso a) con 0.05 como nivel de significancia.
  - Jensen piensa que usará la nueva máquina 30 horas a la semana. Obtenga un intervalo de predicción de 95% para los gastos anuales de mantenimiento de la empresa.
  - Si el precio del contrato de mantenimiento es \$3000 anuales, ¿recomendaría firmar el contrato de mantenimiento? ¿Por qué sí o por qué no?
62. En un determinado proceso de fabricación se cree que la velocidad (pies por minuto) de la línea de ensamblado afectaba al número de partes defectuosas halladas en el proceso de inspección. Para probar esto, los administradores idearon un procedimiento en el que la misma cantidad de partes por lote se examinaba visualmente a diferentes velocidades de la línea. Se recolectaron los datos siguientes.

Velocidad de la línea	Número de partes defectuosas halladas
20	21
20	19
40	15
30	16
60	14
40	17

- Obtenga la ecuación de regresión estimada que relaciona velocidad de la línea de producción con el número de partes defectuosas encontradas.

- b. Empleando el nivel de significancia 0.05, determine si la velocidad de la línea y el número de partes defectuosas halladas están relacionadas.
- c. ¿Se ajusta bien a los datos la ecuación de regresión estimada?
- d. Dé un intervalo de confianza de 95% para predecir el número medio de partes defectuosas si la velocidad de la línea es 50 pies por minuto.
63. Un hospital grande de una ciudad contrató a un sociólogo para que investigara la relación entre el número de días por año de ausencia con autorización, y la distancia (en millas) entre la casa y el trabajo del empleado. Se tomó una muestra de 10 empleados y se obtuvieron los datos siguientes.

archivo  
en CD  
Absent

Distancia al trabajo	Número de días de ausencia
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

- a. Elabore, con estos datos, un diagrama de dispersión.
- b. Obtenga la ecuación de regresión de mínimos cuadrados.
- c. ¿Existe una relación significativa entre las dos variables? Explique.
- d. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique
- e. Emplee la ecuación de regresión estimada obtenida en el inciso b) para calcular un intervalo de confianza de 95% para el número esperado de ausencias (días) de los empleados que vivan a 5 millas de la empresa.
64. La autoridad de tránsito de una zona metropolitana importante desea determinar si hay relación entre la antigüedad de un autobús y los gastos de mantenimiento del mismo. En una muestra de 10 autobuses se obtuvieron los datos siguientes.

archivo  
en CD  
AgeCost

Antigüedad del autobús (años)	Costo de mantenimiento (\$)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

- a. Empleando el método de mínimos cuadrados obtenga la ecuación de regresión estimada.
- b. Haga una prueba para determinar si las dos variables están relacionadas de manera significativa con  $\alpha = 0.05$ .
- c. ¿Proporciona la recta de mínimos cuadrados una buena aproximación a los datos observados? Explique.
- d. Calcule un intervalo de predicción de 95% para los gastos de mantenimiento de un determinado autobús cuya antigüedad es de 4 años.

65. Un profesor de mercadotecnia de una universidad desea saber cuál es la relación entre las horas de estudio y la calificación en un curso. A continuación se presentan los datos obtenidos de 10 estudiantes que tomaron el curso el trimestre pasado.

archivo  
en CD  
HoursPts

	Horas de estudio	Calificación total
	45	40
	30	35
	90	75
	60	65
	105	90
	65	50
	90	90
	80	80
	55	45
	75	65

- a. Obtenga la ecuación de regresión estimada que indica la relación entre calificación y horas de estudio.
- b. Empleando  $\alpha = 0.05$  pruebe la significancia del modelo.
- c. Pronostique la calificación que obtendrá Mark Sweeney. Él estudió 95 horas.
- d. Calcule un intervalo de predicción de 95% para la calificación de Mark Sweeney.
66. *Bloomberg Personal Finance* (julio/agosto 2001) publicó que la beta del mercado de Texas Instrument era 1.46. La beta del mercado de cada acción se determina mediante regresión lineal simple. En cada caso, la variable dependiente es la rentabilidad porcentual trimestral (revalorización más dividendos) menos el rendimiento porcentual que se hubiera obtenido en una inversión libre de riesgos (como tasa libre de riesgo se empleó la tasa Treasury Bill). La variable independiente es la rentabilidad porcentual trimestral (revalorización de capital más dividendos) para el mercado de valores (S&P 500) menos la rentabilidad porcentual de una inversión libre de riesgos. A partir de los datos trimestrales se desarrolla la ecuación de regresión estimada; la beta del mercado de la acción en cuestión es la pendiente de la ecuación de regresión estimada ( $b_1$ ). La beta del mercado suele interpretarse como una medida de lo riesgoso de la acción. Si la beta del mercado es mayor a 1, la volatilidad de la acción es mayor al promedio en el mercado; si la beta del mercado es menor a 1, la volatilidad de la acción es menor al promedio en el mercado. Supóngase que las cifras siguientes son diferencias entre rentabilidad porcentual y rentabilidad libre de riesgos a lo largo de 10 trimestres de S&P 500 y Horizon Technology.

archivo  
en CD  
MktBeta

	S&P 500	Horizon
	1.2	-0.7
	-2.5	-2.0
	-3.0	-5.5
	2.0	4.7
	5.0	1.8
	1.2	4.1
	3.0	2.6
	-1.0	2.0
	0.5	-1.3
	2.5	5.5

- a. Obtenga la ecuación de regresión estimada que sirve para determinar la beta del mercado de Horizon Technology. ¿Cuál es la beta del mercado de Horizon Technology?
  - b. Empleando 0.05 como nivel de significancia, pruebe la significancia de la relación.
  - c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
  - d. Utilice las betas del mercado de Horizon Techology y de Texas Instrument para comparar los riesgos de estas dos acciones.
67. La Transactional Record Access Clearinghouse de la Universidad de Syracuse publica datos que muestran las posibilidades de una auditoría del Departamento de Tesorería de los Estados Unidos. En la tabla siguiente se muestra la media del ingreso bruto ajustado y el porcentaje de declaraciones que fueron auditadas en 20 municipios

Municipio	Ingreso bruto ajustado	Porcentaje auditado
Los Ángeles	36 664	1.3
Sacramento	38 845	1.1
Atlanta	34 886	1.1
Boise	32 512	1.1
Dallas	34 531	1.0
Providence	35 995	1.0
San José	37 799	0.9
Cheyenne	33 876	0.9
Fargo	30 513	0.9
Nueva Orleans	30 174	0.9
Oklahoma City	30 060	0.8
Houston	37 153	0.8
Portland	34 918	0.7
Phoenix	33 291	0.7
Augusta	31 504	0.7
Albuquerque	29 199	0.6
Greensboro	33 072	0.6
Columbia	30 859	0.5
Nashville	32 566	0.5
Buffalo	34 296	0.5

- a. Obtenga la ecuación de regresión estimada que sirve para pronosticar el porcentaje de auditorías dado un ingreso bruto ajustado.
  - b. Empleando como nivel de significancia 0.05, determine si hay relación entre el ingreso bruto ajustado y el porcentaje de auditorías.
  - c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
  - d. Emplee la ecuación de regresión estimada del inciso a) para calcular un intervalo de 95% de confianza para el porcentaje de auditorías en un municipio en el que el promedio del ingreso bruto ajustado es \$35 000.
68. Una institución de un determinado país publicó evaluaciones sobre la satisfacción con el trabajo. Una de las cosas que se pedían en la encuesta era elegir (de una lista de factores) los cinco factores principales para la satisfacción en el trabajo. Después se pedía a los encuestados que indicaran su nivel de satisfacción con cada uno de esos cinco factores. En la tabla siguiente se presentan los porcentajes de personas para los que el factor indicado fue uno de los cinco factores principales, junto con una evaluación obtenida empleando el porcentaje de personas que consideraron al factor como uno de los principales y que estaban “muy satisfechos” o “satisfechos” con ese factor. ([www.apse.gov.au/stateoftheservice](http://www.apse.gov.au/stateoftheservice)).

Factor	Cinco principales (%)	Evaluación (%)
Carga de trabajo adecuada	30	49
Posibilidad de ser creativo o de hacer innovaciones	38	64
Posibilidad de hacer contribuciones útiles a la sociedad	40	67
Obligaciones y expectativas claramente planteadas	40	69
Condiciones flexibles de trabajo	55	86
Buena relación de trabajo	60	85
Trabajo interesante	48	74
Oportunidad de hacer carrera	33	43
Oportunidad de desarrollar sus habilidades	46	66
Oportunidad de utilizar sus habilidades	50	70
Retroalimentación y reconocimiento al esfuerzo realizado	42	53
Salario	47	62
Poder ver resultados tangibles del trabajo	42	69

- Elabore un diagrama de dispersión colocando en el eje horizontal los porcentajes de los factores principales y en el eje vertical la evaluación correspondiente.
- ¿Qué indica, respecto a la relación entre las dos variables, el diagrama de dispersión elaborado en el inciso a)?
- Obtenga la ecuación de regresión estimada que sirva para pronosticar la evaluación (%) dado el porcentaje del factor (%).
- Empleando como nivel de significancia 0.05 realice una prueba para determinar la significancia de la relación.
- ¿Proporciona la ecuación de regresión estimada un buen ajuste?
- Dé el valor del coeficiente de correlación muestral.

## Caso problema 1 Medición del riesgo en el mercado bursátil

Una medida del riesgo o volatilidad de una acción es la desviación estándar del rendimiento durante un lapso de tiempo. Aunque la desviación estándar es fácil de calcular, no toma en cuenta la variación del precio de una acción en función de un índice estándar del mercado, como el S&P 500. Por esta razón, muchos analistas financieros prefieren emplear otra medida, conocida como *beta*, para medir el riesgo.

La beta de una acción se determina mediante regresión lineal simple. La variable independiente es la rentabilidad total de la acción de que se trate y la variable independiente es la rentabilidad total del mercado de valores.\* En este caso problema se usará el índice S&P 500 como medida de la rentabilidad total del mercado de valores y se obtendrá una ecuación de regresión estimada usando datos mensuales. La beta de una acción es la pendiente en la ecuación de regresión estimada ( $b_1$ ). En el archivo Beta del disco compacto que se distribuye con el libro se proporciona la rentabilidad total de ocho acciones comunes muy conocidas y la del S&P 500 a lo largo de 36 meses.

El valor beta del mercado de valores siempre será 1; por lo tanto, una acción que tienda a subir o a bajar con el mercado de valores tendrá también una beta cercana a 1. Betas mayores a 1 corresponden a acciones que son más volátiles que el mercado y betas menores a 1 corresponden a acciones menos volátiles que el mercado. Por ejemplo, si la beta de una acción es 1.4, esta acción es 40% más volátil que el mercado, y si la beta de una acción es 0.4, la acción es 60% menos volátil que el mercado.

\*Diversas fuentes emplean diferentes métodos para calcular las betas. Por ejemplo, algunas fuentes, antes de calcular la ecuación de regresión estimada, restan, de la variable independiente, la rentabilidad que podría haberse obtenido con una inversión libre de riesgos [por ejemplo, letras del tesoro (Estados Unidos)(T-bills)]. Otras, para la rentabilidad total del mercado de valores emplean diversos índices; por ejemplo, *Value line* calcula las betas usando el índice compuesto de la bolsa de Nueva York.

## Reporte administrativo

Se le ha encomendado la tarea de analizar las características del riesgo de estas acciones. Elabore un informe que comprenda los puntos siguientes, sin limitarse sólo a ellos.

- Calcular los estadísticos descriptivos de cada una de las acciones y del S&P 500. Hacer comentarios sobre los resultados. ¿Qué acción es la más volátil?
- Calcular la beta de cada acción. ¿Cuál de estas acciones se esperaría que se comportara mejor en un mercado de alta calidad? ¿Cuál conservaría mejor su valor en un mercado para el sector popular?
- Haga un comentario sobre qué tanto de la rentabilidad de cada una de las acciones es explicado por el mercado.

## Caso problema 2 Departamento de Transporte de Estados Unidos

Como parte de un estudio sobre seguridad en el transporte, el Departamento de Transporte de Estados Unidos, de una muestra de 42 ciudades, recogió datos sobre el número de accidentes fatales por cada 1000 licencias y sobre el porcentaje de licencia de conductores menores de 21 años. A continuación se presentan los datos recogidos en el lapso de un año. Estos datos se encuentran también en el archivo titulado Safety del disco compacto que se distribuye con el libro.



Porcentaje de menores de 21 años	Accidentes fatales por 1000 licencias	Porcentaje de menores de 21 años	Accidentes fatales por 1000 licencias
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

## Informe administrativo

- Presente resúmenes numéricos y gráficos de los datos.
- Emplee el análisis de regresión para investigar la relación entre el número de accidentes fatales y el porcentaje de conductores menores de 21 años. Analice sus hallazgos.
- ¿Qué conclusión y qué recomendaciones puede deducir de su análisis?

## Caso problema 3 Donaciones de los ex alumnos

Las donaciones de los ex alumnos son una importante fuente de ingresos para las universidades. Si los gerentes pudieran determinar los factores que influyen sobre el aumento del porcentaje de alumnos que hace donaciones, podrían poner en marcha políticas que llevarían a ganancias mayores. Las investigaciones indican que estudiantes más satisfechos con la relación con sus profesores tienen más probabilidad de titularse, lo que a su vez puede llevar al aumento del porcentaje de alumnos que haga donaciones. En la tabla 14.13 se muestran datos de 48 universidades de Estados Unidos (*American's Best Collage*, edición del año 2000). La columna titulada “% de grupos con menos de 20” muestra el porcentaje de grupos con menos de 20 alumnos. La columna que tiene como título “Tasa de estudiantes/facultad” da el número de estudiantes inscritos, dividido entre el número total de facultades. Por último, la columna que tiene como título “Tasa de alumnos que donan” da el porcentaje de alumnos que han hecho alguna donación a la universidad.

### Reporte administrativo

1. Presente resúmenes numéricos y gráficos de los datos.
2. Emplee el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de los estudiantes que hacen donaciones dado el porcentaje de grupos con menos de 20 estudiantes.
3. Use el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de los alumnos que hacen donaciones dada la proporción de estudiantes por facultad.
4. ¿Cuál de las dos ecuaciones de regresión estimada muestra un mejor ajuste? Con esa ecuación de regresión estimada realice un análisis de residuales y discuta sus hallazgos y conclusiones.
5. ¿Qué conclusiones y recomendaciones puede obtener de este análisis?

## Caso problema 4 Valor de los equipos de béisbol de la liga mayor

Un grupo encabezado por John Henry pagó \$700 millones por la adquisición del equipo Boston Red Sox (Medias Rojas de Boston) en 2002, aun cuando el Boston Red Sox no había ganado la serie mundial desde 1918 y tenía una pérdida de operación de \$11.4 millones de 2001. Es más, la revista *Forbes* estima que el valor actual del equipo es en realidad \$426 millones. *Forbes* atribuye la diferencia entre valor actual del equipo y precio que los inversionistas están dispuestos a pagar, al hecho de que la compra de un equipo suele incluir la adquisición de una red de cable exageradamente subvaluada. Por ejemplo, con la compra del equipo, los nuevos propietarios obtuvieron también la New England Sports Network. En la tabla 14.14 se presentan los datos de 30 equipos de la liga mayor (*Forbes*, 15 de abril de 2002). En la columna titulada Valor se da el valor de los equipos con base en las actuales negociaciones con los estadios, sin deducción de deudas. En la columna titulada Ingreso se presentan las ganancias sin intereses, impuestos y depreciación.

### Informe administrativo

1. Presente resúmenes numéricos y gráficos de los datos.
2. Use el análisis de regresión para investigar la relación entre valor e ingreso. Discuta sus hallazgos.
3. Use el análisis de regresión para investigar la relación entre valor y ganancias. Discuta sus hallazgos.
4. ¿Qué conclusiones y recomendaciones puede sacar de este análisis?

**TABLA 14.13** DATOS DE 48 UNIVERSIDADES NACIONALES

	% de grupos con menos de 20	Tasa de estudiantes/facultad	Tasa de alumnos que donan
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve Univ.	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State Univ.	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California–Berkeley	58	17	18
U. of California–Davis	32	19	7
U. of California–Irvine	42	20	9
U. of California–Los Angeles	41	18	13
U. of California–San Diego	48	19	8
U. of California–Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois–Urbana Champaign	29	15	23
U. of Michigan–Ann Arbor	51	15	13
U. of North Carolina–Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22
U. of Texas–Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin–Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University–St. Louis	73	7	33
Yale University	77	7	50

**TABLA 14.14** DATOS DE LOS EQUIPOS DE LA LIGA MAYOR DE BÉISBOL

Equipo	Valor	Ganancia	Ingreso
New York Yankees	730	215	18.7
New York Mets	482	169	14.3
Los Angeles Dodgers	435	143	-29.6
Boston Red Sox	426	152	-11.4
Atlanta Braves	424	160	9.5
Seattle Mariners	373	166	14.1
Cleveland Indians	360	150	-3.6
Texas Rangers	356	134	-6.5
San Francisco Giants	355	142	16.8
Colorado Rockies	347	129	6.7
Houston Astros	337	125	4.1
Baltimore Orioles	319	133	3.2
Chicago Cubs	287	131	7.9
Arizona Diamondbacks	280	127	-3.9
St. Louis Cardinals	271	123	-5.1
Detroit Tigers	262	114	12.3
Pittsburgh Pirates	242	108	9.5
Milwaukee Brewers	238	108	18.8
Philadelphia Phillies	231	94	2.6
Chicago White Sox	223	101	-3.8
San Diego Padres	207	92	5.7
Cincinnati Reds	204	87	4.3
Anaheim Angels	195	103	5.7
Toronto Blue Jays	182	91	-20.6
Oakland Athletics	157	90	6.8
Kansas City Royals	152	85	2.2
Tampa Bay Devil Rays	142	92	-6.1
Florida Marlins	137	81	1.4
Minnesota Twins	127	75	3.6
Montreal Expos	108	63	-3.4

## Apéndice 14.1 Deducción de la fórmula de mínimos cuadrados empleando el cálculo

Como ya se indicó en este capítulo, el método de mínimos cuadrados se usa para determinar los valores de  $b_0$  y  $b_1$  que minimicen la suma de los cuadrados de los residuales. La suma de los cuadrados de los residuales está dada por

$$\Sigma(y_i - \hat{y}_i)^2$$

Sustituyendo  $\hat{y}_i = b_0 + b_1x_i$ , se obtiene

$$\Sigma(y_i - b_0 - b_1x_i)^2 \quad (14.34)$$

como expresión que hay que minimizar.

Para minimizar la expresión (14.14), se sacan las derivadas parciales respecto a  $b_0$  y  $b_1$ , se igualan a cero y despeja. Haciendo esto se obtiene

$$\frac{\partial \Sigma(y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2\Sigma(y_i - b_0 - b_1 x_i) = 0 \quad (14.35)$$

$$\frac{\partial \Sigma(y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\Sigma x_i(y_i - b_0 - b_1 x_i) = 0 \quad (14.36)$$

Dividiendo la ecuación (14.35) entre dos y haciendo las sumas por separado, se obtiene

$$-\Sigma y_i + \Sigma b_0 + \Sigma b_1 x_i = 0$$

Llevando  $\Sigma y_i$  al otro lado del signo igual y observando que  $\Sigma b_0 = nb_0$ , se obtiene

$$nb_0 + (\Sigma x_i)b_1 = \Sigma y_i \quad (14.37)$$

Simplificaciones algebraicas similares aplicadas a la ecuación (14.36) producen

$$(\Sigma x_i)b_0 + (\Sigma x_i^2)b_1 = \Sigma x_i y_i \quad (14.38)$$

A las ecuaciones (14.37) y (14.38) se les conoce como *ecuaciones normales*. Despejando  $b_0$  en la ecuación (14.37) se obtiene

$$b_0 = \frac{\Sigma y_i}{n} - b_1 \frac{\Sigma x_i}{n} \quad (14.39)$$

Usando la ecuación (14.39) para sustituir a  $b_0$  en la ecuación (14.38) da

$$\frac{\Sigma x_i \Sigma y_i}{n} - \frac{(\Sigma x_i)^2}{n} b_1 + (\Sigma x_i^2)b_1 = \Sigma x_i y_i \quad (14.40)$$

Reordenando los términos de la ecuación (14.40), se obtiene

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} \quad (14.41)$$

Como  $\bar{y} = \Sigma y_i/n$  y  $\bar{x} = \Sigma x_i/n$ , la ecuación (14.39) se puede escribir como

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.42)$$

Las ecuaciones (14.41) y (14.42) son las fórmulas (14.6) y (14.7) usadas en este capítulo para calcular los coeficientes de la ecuación de regresión estimada.

## Apéndice 14.2 Una prueba de significancia usando correlación

Empleando el coeficiente de correlación muestral  $r_{xy}$ , también se puede determinar si la relación lineal entre  $x$  y  $y$  es significativa mediante la siguiente prueba de hipótesis acerca del coeficiente de correlación muestral.

$$\begin{aligned} H_0: \rho_{xy} &= 0 \\ H_a: \rho_{xy} &\neq 0 \end{aligned}$$

Si  $H_0$  es rechazada, se concluye que el coeficiente de correlación no es igual a cero y que la relación entre las dos variables no es significativa. A continuación se presenta esta prueba de significancia.

#### PRUEBA DE SIGNIFICANCIA USANDO CORRELACIÓN

$$H_0: \rho_{xy} = 0$$

$$H_a: \rho_{xy} \neq 0$$

#### ESTADÍSTICO DE PRUEBA

$$t = r_{xy} \sqrt{\frac{n - 2}{1 - r_{xy}^2}} \quad (14.43)$$

#### REGLA DE RECHAZO

Método del valor- $p$ : Rechazar  $H_0$  si valor- $p \leq \alpha$

Método del valor crítico: Rechazar  $H_0$  si  $t \leq -t_{\alpha/2}$  o si  $t \geq t_{\alpha/2}$

donde  $t_{\alpha/2}$  pertenece a la distribución  $t$  con  $n - 2$  grados de libertad.

En la sección 14.4 con una muestra  $n = 10$  se encontró que el coeficiente de correlación muestral para la población de estudiantes y las ventas trimestrales era  $r_{xy} = 0.9501$ . El estadístico de prueba es

$$t = r_{xy} \sqrt{\frac{n - 2}{1 - r_{xy}^2}} = 0.9501 \sqrt{\frac{10 - 2}{1 - (0.9501)^2}} = 8.61$$

En la tabla de la distribución  $t$  se encuentra que para  $n - 2 = 10 - 2 = 8$  grados de libertad,  $t = 3.355$  proporciona un área de 0.005 en la cola superior. Por lo tanto, al área en la cola superior que corresponde al estadístico de prueba  $t = 8.61$  debe ser menor a 0.005. Como esta prueba es una prueba de dos colas, se duplica este valor y se concluye que el valor  $t$  que corresponde a  $t = 8.62$  debe ser menor a  $2(0.005) = 0.01$ . Con Excel o con Minitab se obtiene valor- $p = 0.000$ . Como el valor- $p$  es menor a  $\alpha = 0.01$ , se rechaza  $H_0$  y se concluye que  $r_{xy}$  no es igual a cero. Esta evidencia es suficiente para concluir que entre la población de estudiantes y las ventas trimestrales existe una relación lineal significativa.

Obsérvese que el valor del estadístico de prueba  $t$  y la conclusión sobre la significancia de la relación son idénticos con los resultados obtenidos en la prueba  $t$  de la sección 14.5, en donde se usó la ecuación de regresión estimada  $\hat{y} = 60 + 5x$ . El análisis de regresión permite obtener una conclusión sobre la relación entre las variables  $x$  y  $y$ ; además, permite obtener la ecuación que indica cuál es la relación entre las variables. Por consiguiente, la mayoría de los analistas emplean paquetes modernos de software para realizar el análisis de regresión y encuentran que el empleo de la correlación como prueba de significancia es innecesario.

## Apéndice 14.3 Análisis de regresión con Minitab



En la sección 14.7 mostrando los resultados que da Minitab para el problema de Armand's Pizza Parlors se estudió la solución de los problemas de regresión mediante el empleo de paquetes de software. En este apéndice se describen los pasos necesarios al emplear Minitab para generar esos resultados. Primero, en una hoja de cálculo de Minitab se ingresan los datos. Los datos de las poblaciones de estudiantes se ingresan en la columna C1 y los datos de las ventas trimestrales se ingresan en la columna C2. Los nombres de las variables Pop y Sales (Ventas) se ingresan como encabezados de esas columnas. En la descripción de los pasos a seguir, para referirse a los datos se emplearán los nombres de las variables o los indicadores de las columnas C1 y C2. Los

pasos siguientes describen cómo usar Minitab para obtener los resultados del análisis de regresión que se muestran en la figura 14.10.

**Paso 1.** Seleccionar el menú **Stat**

**Paso 2.** Seleccionar el menú **Regression**

**Paso 3.** Elegir **Regression**

**Paso 4.** Cuando aparezca el cuadro de diálogo Regresión:

    Ingresar Sales en el cuadro **Response**

    Ingresar Pop en el cuadro **Predictors**

    Clic en el botón **Options**

Cuando aparezca el cuadro de diálogo Regression-Options:

    Ingresar 10 en el cuadro **Prediction intervals for new observations**

    Clic en **OK**

Cuando aparezca el cuadro de diálogo Regression:

    Clic en **OK**

El cuadro de diálogo de Minitab tiene otras posibilidades más que se pueden aprovechar seleccionando las opciones deseadas. Por ejemplo, para obtener una gráfica de residuales, en la que los valores pronosticados  $\hat{y}$  aparezcan en el eje horizontal y los valores de los residuales estandarizados en el eje vertical, el paso 4 deberá ser como sigue:

**Paso 4** Cuando aparezca el cuadro de diálogo Regression:

    Ingresar Sales en el cuadro **Response**

    Ingresar Pop en el cuadro **Predictors**

    Clic en el botón **Graphs**

Cuando aparezca el cuadro de diálogo Regression-Graphs:

    Seleccionar **Standardized** en Residuals for Plots

    Seleccionar **Residuals versus fits** en Residual Plots

    Clic en **OK**

Cuando aparezca el cuadro de diálogo Regression:

    Clic en **OK**

## Apéndice 14.4 Análisis de regresión con Excel



En este apéndice se ilustra el uso de la herramienta de Excel para realizar los cálculos del análisis de regresión empleando el problema de Armand's Pizza Parlors. Consultese la figura 14.23, para seguir la descripción de los pasos. En las celdas A1:C1 de la hoja de cálculo se ingresan los rótulos Restaurante, Población y Ventas. Para identificar cada una de las 10 observaciones, se ingresan los números del 1 al 10 en las celdas A2:A11. Los datos muestrales se ingresan en las celdas B2:C11. Los pasos siguientes indican cómo obtener los resultados del análisis de regresión.

**Paso 1.** Seleccionar el menú **Herramientas**

**Paso 2.** Elegir el menú **Análisis de datos**

**Paso 3.** Elegir **Regresión** en el menú de Funciones para análisis

**Paso 4.** Clic en **OK**

**Paso 5.** Cuando aparezca el cuadro de diálogo Regresión:

    Ingresar C1:C11 en el cuadro **Rango Y de entrada**

    Ingresar B1:B11 en el cuadro **Rango X de entrada**

    Seleccionar **Rótulos**

    Seleccionar **Nivel de confianza**

    Ingresar **99** en el cuadro Nivel de confianza

    Seleccionar **Rango de salida**

    Ingresar A13 en el cuadro **Rango de salida**

    (También se puede ingresar cualquier celda, de la esquina superior izquierda, para indicar dónde deberán empezar los resultados.)

    Clic en **OK**

**FIGURA 14.23** SOLUCIÓN CON EXCEL AL PROBLEMA DE ARMAND'S PIZZA PARLORS

	A	B	C	D	E	F	G	H	I	J
1	Restaurant	Population	Sales							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	SUMMARY OUTPUT									
14										
15	Regression Statistics									
16	Multiple R	0.9501								
17	R Square	0.9027								
18	Adjusted R Square	0.8906								
19	Standard Error	13.8293								
20	Observations	10								
21										
22	ANOVA									
23		df	SS	MS	F	Significance F				
24	Regression	1	14200	14200	74.2484	2.55E-05				
25	Residual	8	1530	191.25						
26	Total	9	15730							
27										
28		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%	
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										
32										
33										
34										

La primera sección de los resultados, titulada *Estadísticas de regresión*, contiene resúmenes estadísticos como el coeficiente de determinación ( $R^2$ ). La segunda sección de los resultados, titulada Análisis de varianza, contiene la tabla del análisis de varianza. La última sección de los resultados, que no tiene ningún título, contiene los coeficientes de regresión estimados e información relacionada con ellos. A continuación se da la interpretación de los resultados de la regresión empezando con la información contenida en las celdas A28:I30.

### Interpretación de los resultados de la ecuación de regresión estimada

La intersección de la recta de regresión con el eje  $y$ ,  $b_0 = 60$ , aparece en la celda B29 y la pendiente de la recta de regresión estimada,  $b_1 = 5$ , aparece en la celda B30. El rótulo Intercepción en la celda A29 y el rótulo Población en la celda A30 sirven para identificar estos dos valores.

En la sección 14.5 se mostró que la desviación estándar estimada de  $b_1$  es  $s_{b_1} = 0.5803$ . Obsérvese que el valor de la celda C30 es 0.5803. El rótulo Error típico que aparece en la celda C28, es la manera en que Excel indica que el valor de la celda C30 es el error estándar o la desviación estándar de  $b_1$ . Recuérdese que en la prueba  $t$  de significancia de la relación fue necesario calcular el estadístico  $t$ ,  $t = b_1/s_{b_1}$ . Empleando los datos de Armand's, se obtuvo como valor  $t$ ,  $t = 5/0.5803 = 8.62$ . El rótulo Estadístico  $t$  de la celda D28 sirve para recordar que en la celda D30 se encuentra el valor del estadístico  $t$ .

El valor en la celda E30 es el valor  $-p$  que corresponde a la prueba  $t$  de significancia. El valor- $p$  que da Excel en la celda E30, está en notación científica. Para obtener este valor en notación decimal, se recorre el punto decimal 5 lugares a la izquierda, con lo que se obtiene 0.0000255. Dado que valor- $p = 0.0000255 < \alpha = 0.01$ , se rechaza  $H_0$  y se concluye que entre la población de estudiantes y las ventas trimestrales existe una relación significativa.

La información de las celdas F28:I30 se emplea para obtener estimaciones por Intervalos de confianza para la intersección con el eje  $y$  y la pendiente de la ecuación de regresión estimada. Excel siempre da los límites inferior y superior de un intervalo de 95% de confianza. Como en el paso 4 se seleccionó Intervalo de confianza y se ingresó 99 en el cuadro de Nivel de Confianza, la herramienta de Excel para regresión da también los límites inferior y superior de un intervalo de 99% de confianza. El valor en la celda H30 es el límite inferior de la estimación por intervalo del 99% de confianza de  $b_1$  y el valor en la celda I30 es el límite superior. Por lo tanto, una vez redondeada, el intervalo de 99% de confianza para estimar  $b_1$  va de 3.05 a 6.95. Los valores en las celdas F30 a G30 proporcionan los límites inferior y superior del intervalo de 95% de confianza. El intervalo de 95% de confianza va de 3.66 a 6.34.

## Interpretación de los resultados del ANOVA

La información en las celdas A22:F26 es un resumen de los cálculos del análisis de varianza. Las tres fuentes de variación están rotuladas como Regresión, Residuo y Total. La etiqueta  $df$  en la celda B23 representa los grados de libertad, la etiqueta  $SS$  en la celda C23 representa la suma de los cuadrados y la etiqueta  $MS$  en la celda D23 representa el cuadrado de la media.

En la sección 14.5 se dijo que el error cuadrado medio, que se obtiene dividiendo el error o la suma de cuadrados del residual entre sus grados de libertad, proporciona una estimación de  $\sigma^2$ . El valor en la celda D25, 191.25, es el error cuadrado medio de los resultados de regresión para el problema de Armand's. En la sección 14.5 se mostró que también se puede usar una prueba  $F$  como prueba de significancia en la regresión. El valor en la celda F24, 0.0000255, es el valor- $p$  que corresponde a la prueba  $F$  de significancia. Dado que valor- $p = 0.0000255 < \alpha = 0.01$ , se rechaza  $H_0$  y se concluye que se tiene una relación significante entre la población de estudiantes y las ventas trimestrales. En la celda F23, el rótulo que emplea Excel para identificar el valor- $p$  de la prueba  $F$  de significancia es *Valor crítico de F*.

*El rótulo Valor crítico de F se entiende mejor si se considera el valor en la celda F24 como el nivel de significancia observado en la prueba F.*

## Interpretación de los estadísticos de regresión de los resultados

El coeficiente de determinación, 0.9027, aparece en la celda B17; el rótulo correspondiente, Coeficiente de determinación R\*2, aparece en la celda A17. La raíz cuadrada del coeficiente de determinación es el coeficiente de correlación muestral, 0.9501, que aparece en la celda B16. Obsérvese que para identificar este valor, Excel emplea como rótulo Coeficiente de correlación múltiple. En la celda A19, el rótulo Error Estándar se usa para identificar el valor del error estándar de estimación que aparece en la celda B19. Así que el error estándar de estimación es 13.8293. Hay que tener presente que en los resultados de Excel, el rótulo Error típico aparece en dos lugares. En la sección de los resultados titulada Estadísticas de regresión, el rótulo Error típico se refiere a la estimación de  $\sigma$ ; en la sección de los resultados correspondiente a la Ecación de regresión estimada, el rótulo *Error típico* se refiere a  $s_{b_1}$ , la desviación estándar de la distribución muestral de  $b_1$ .