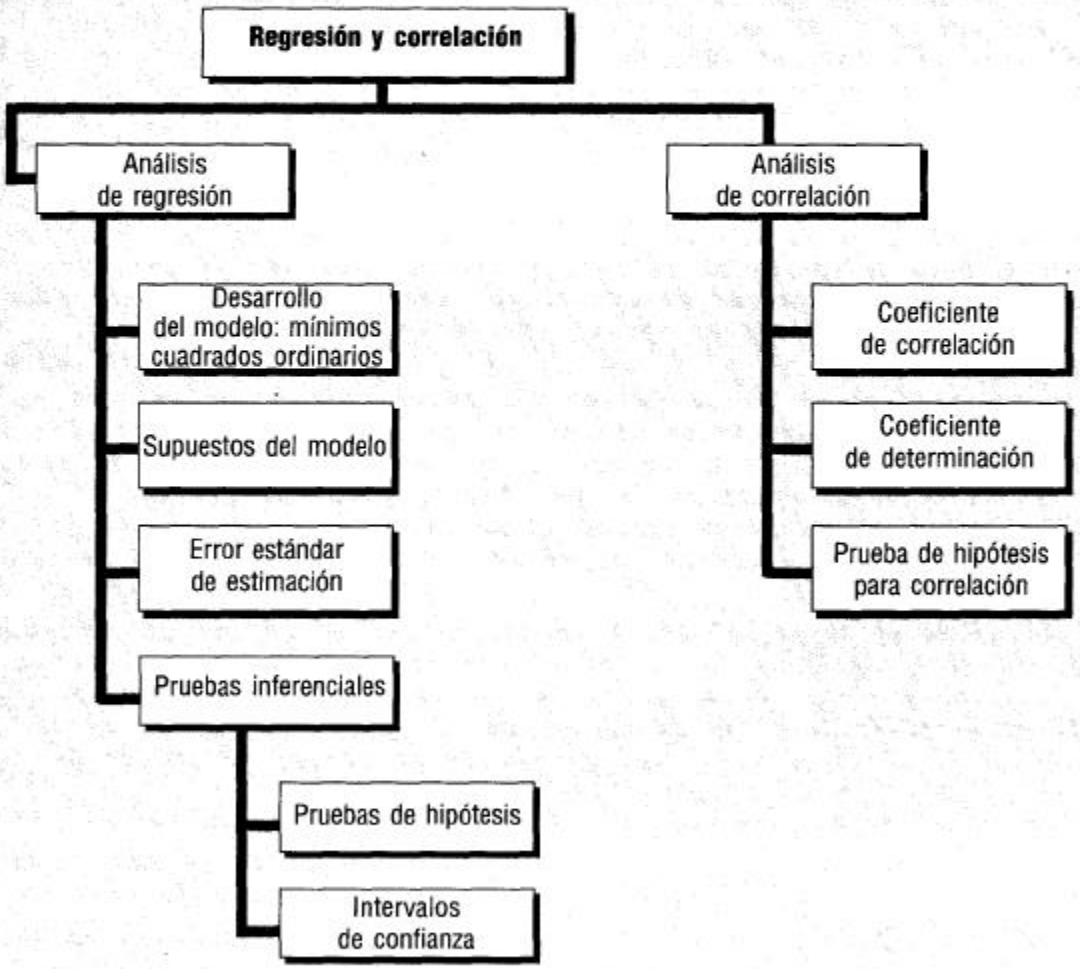
11

Regresión simple y correlación

Plan del capítulo

Este capítulo estudia dos de las más importantes y útiles herramientas del análisis estadístico: la regresión y la correlación. Estas técnicas poderosas ilustran la forma como pueden analizarse las relaciones entre dos variables para predecir eventos futuros.



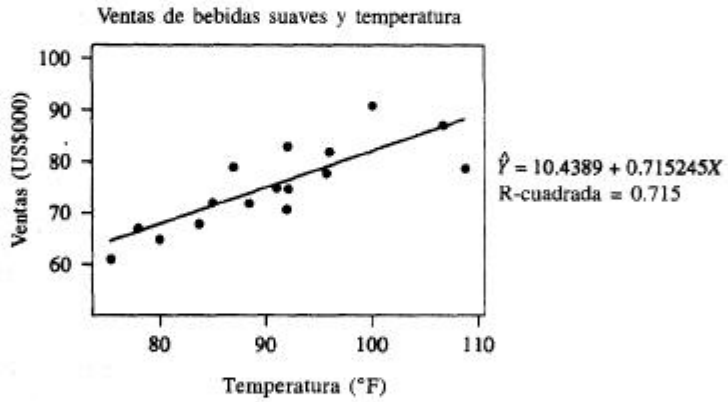
ESCENARIO

La competencia en la industria de las bebidas suaves siempre ha sido intensa. Recientemente, la lucha entre Coca-Cola y Pepsi-Cola se ha puesto álgida por incrementar sus participaciones respectivas de US$27 mil millones en el mercado nacional de bebidas. Cada compañía ha ofrecido su propia marca de *flair* promocional en un esfuerzo continuo por reorganizar la mezcla en su mercadeo y promocionar su respectivo producto. Coca-Cola actualmente goza de un 21.7% de participación en el mercado, seguido de Pepsi al 18.9%.

Sin duda alguna los ejecutivos de mercadeo, los especialistas en gerencia y los estadísticos trabajan duro en ambas compañías intentando sobrepasar a sus contrapartes de mente competitiva. Hasta ahora se han puesto de acuerdo en muy poco, salvo en que las ventas parecen incrementar con las elevadas temperaturas del verano.

Predecir las tendencias en la participación de mercado es una tarea especialmente ardua y difícil. Muchos ejecutivos han echado a perder sus carreras en el frustrado intento de anticipar correctamente el comportamiento de los volubles consumidores.

La regresión y el análisis de correlación son las dos herramientas más poderosas y útiles que los analistas de todo tipo tienen a su disposición para escudriñar el interior del futuro sombrío. En este capítulo se analizarán estos procedimientos y se enseñará cómo ellos pueden orientar a los profesionales en negocios en su búsqueda de una carrera exitosa.



11.1 Introducción

La regresión y la correlación son las dos herramientas estadísticas más poderosas y versátiles que se pueden utilizar para solucionar problemas comunes en los negocios. Muchos estudios se basan en la creencia de que se puede identificar y cuantificar alguna relación funcional entre dos o más variables. Se dice que una variable depende de la otra. Se puede decir que Y depende de X en donde Y y X son dos variables cualquiera. Esto se puede escribir así

Y es una función de X $$Y = f(X)$$ [11.1]

Se lee "Y es función de X".

Debido a que Y depende de X, Y es la **variable dependiente** y X es la **variable independiente**. Es importante identificar cuál es la variable dependiente y cuál es la variable independiente en el modelo de regresión. Esto depende de la lógica y de lo que el estadístico intente medir. El decano de la universidad desea analizar la relación entre las notas de los estudiantes y el tiempo que pasan estudiando. Se recolectaron datos sobre ambas variables. Es lógico presumir que las notas dependen de la cantidad y calidad de tiempo que los estudiantes pasan con sus libros. Por tanto, "notas" es la variable dependiente y "tiempo" es la variable independiente.

Variable dependiente Es la variable que se desea explicar o predecir; también se le denomina *regresando* o *variable de respuesta.*

La variable independiente X se utiliza para explicar Y.

Variable independiente Es la variable independiente, también se le denomina *variable explicativa* o *regresor*

Se dice que "Y está regresando por X".

El primero en desarrollar el análisis de regresión fue el científico inglés Sir Francis Galton (1822 - 1911). Sus primeros experimentos con regresión comenzaron con un intento de analizar los patrones de crecimiento hereditarios de los guisantes. Animado por los resultados, Sir Francis extendió su estudio para incluir los patrones hereditarios en la estatura de las personas adultas. Descubrió que los niños que tienen padres altos o bajos tendían a "regresar" a la estatura promedio de la población adulta. Con este modesto inicio el uso del análisis de regresión se dio a conocer convirtiéndose en una de las herramientas estadísticas más poderosas que se encuentran disponibles actualmente.

Se debe diferenciar entre la regresión simple y la regresión múltiple. En la **regresión simple**, se establece que Y es una función de sólo una variable independiente. Con frecuencia se le denomina regresión *bivariada* porque sólo hay dos variables, una dependiente y una independiente, y la regresión simple se representa con la fórmula (11.1). En un modelo de **regresión múltiple**, Y es una función de dos o más variables independientes. Un modelo de regresión con k variables independientes se puede expresar así:

En un modelo de regresión múltiple Y es una función de dos o más variables independientes

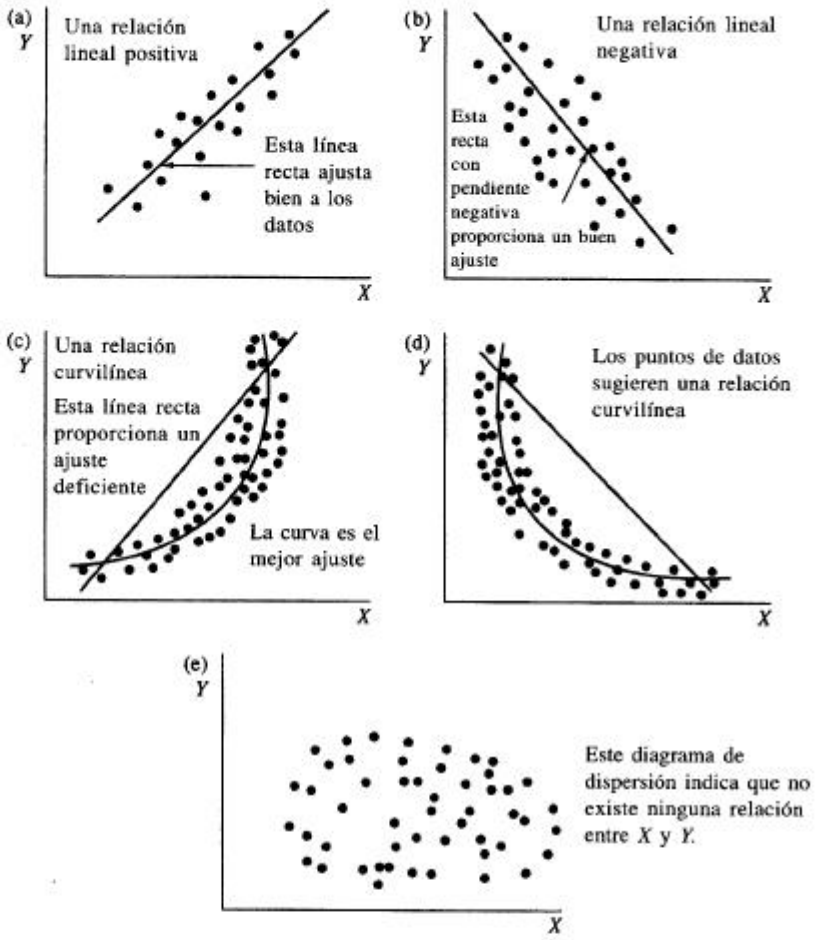
$$Y = f(X_1, X_2, X_3, \ldots, X_k) \qquad [11.2]$$

en donde $X_1, X_2, X_3, \ldots, X_k$ son variables independientes que permiten explicar Y.

También es necesario hacer una distinción entre la regresión lineal y la regresión curvilineal (no lineal) En modelo de **regresión lineal**, la relación entre X y Y puede representarse por medio de una línea recta. Sostiene que a medida que X cambia, Y cambia en una cantidad constante. La **regresión curvilineal** utiliza una curva para expresar la relación entre X y Y. Sostiene que a medida que X cambia, Y cambia en una cantidad *diferente* cada vez.

Algunas de estas relaciones aparecen en la figura 11.1 y muestran **diagramas de dispersión** que representan las observaciones por pares para X y Y. Es habitual colocar la variable independiente en el eje horizontal. La figura 11.1a) sugiere una relación positiva y lineal entre X y Y. Es positiva porque X y Y parecen moverse en la misma dirección. A medida que X aumenta (disminuye), Y aumenta (disminuye). Es lineal porque la relación puede

Figura 11.1
Diagramas de dispersión



identificarse mediante una línea recta que se dibuja entre los puntos. La figura 11.1 b) muestra una relación lineal y negativa entre X y Y, porque las dos variables parecen moverse en direcciones opuestas. Las figuras 11.1 c) y 11.1d) indican relaciones curvilineales. El patrón de los puntos de dispersión no se describe bien con la línea recta, pero se define de manera más exacta con la curva que proporciona un mejor ajuste. Finalmente, es difícil observar toda relación entre X y Y en la figura 11.1 e). La ausencia de todo patrón detectable sugiere que no existe ninguna relación entre X y Y.

Relaciones lineales y curvilineales Si X y Y se relacionan en forma lineal, entonces a medida que X cambia, Y cambia en una cantidad constante. Si existe una relación curvilínea, Y cambiará en una cantidad diferente a medida que X cambia.

11.2 Determinación del modelo de regresión lineal simple

Este capítulo se concentrará en la regresión lineal simple. Sólo son necesarios dos puntos para dibujar la línea recta que representa esta relación lineal. La ecuación de una recta puede expresarse como

Ecuación de la recta	$Y = b_0 + b_1 X$	[11.3]

en donde b_0 es el intercepto y b_1 es la pendiente de la recta. Por ejemplo se tiene que:

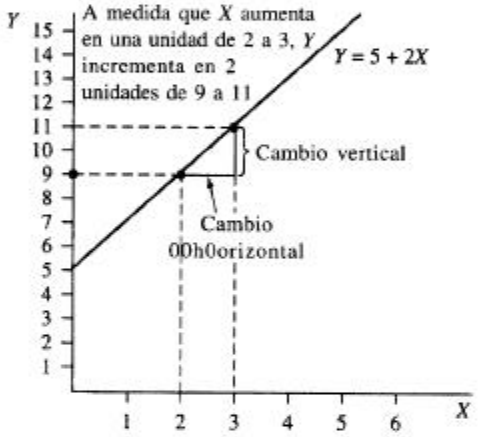
$$Y = 5 + 2X$$

entonces, como se observa en la figura 11.2, la recta intersecta el eje vertical en 5. Además, la pendiente de la recta se halla

$$b_1 = \text{pendiente} = \frac{\text{variación vertical}}{\text{variación horizontal}} = \frac{2}{1} = 2$$

Por cada cambio de una unidad en X, Y cambia en dos unidades.

Figura 11.2
Línea recta con pendiente positiva

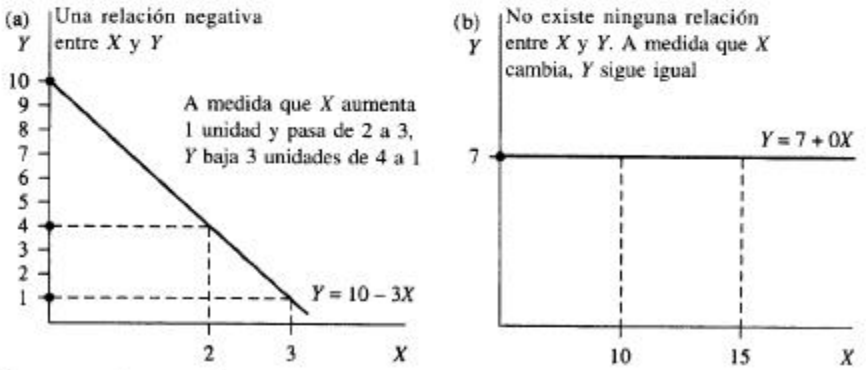


Vale la pena destacar que a medida que X incrementa de 2 a 3 (un incremento de una unidad), Y incrementa de 9 a 11 (un incremento de dos unidades). La figura 11.3 a) muestra una gráfica en la cual $b_1 < 0$, por ejemplo,

$$Y = 10 - 3X$$

Revela que existe una relación negativa tal entre X y Y que por cada incremento (reducción) de una unidad en X, Y reducirá (aumentará) en 3 unidades. Si la pendiente de la recta es $b_1 = 0$ como en la figura 11.3 b), entonces un cambio de X no tiene relación con un cambio en Y. Por tanto, X no puede utilizarse como variable explicativa de Y.

Figura 11.3
Gráfica de líneas rectas



Las relaciones entre variables son o **determinísticas** o **estocásticas** (aleatorias). Una relación determinística puede expresarse mediante la fórmula que convierte la velocidad expresada en millas por hora (mph) a kilómetros

por hora (kph). Ya que 1 milla es aproximadamente igual a 1.6 kilómetros, este modelo es 1 mph = 1.6 kph. Por tanto, una velocidad de 5 mph = 5 (1.6) kph = 8.0 kph. Este es un modelo determinístico porque la relación es exacta y no hay error (salvo la aproximación).

Infortunadamente, muy pocas relaciones en el mundo de los negocios son así de exactas. Con frecuencia se encuentra que al utilizar una variable para explicar otra, existe alguna variación en la relación. Por ejemplo, se supone que la gerencia de Vita + Plus, Inc., distribuidores de productos para la salud, desea desarrollar un modelo de regresión en el cual se utiliza la publicidad para explicar los ingresos por concepto de ventas. Probablemente encontrarán que cuando hacen publicidad y ésta se fija en cierta cantidad X_i, las ventas tendrán algún valor Y_i. Sin embargo, la próxima vez que se fije la publicidad en la misma cantidad, las ventas pueden producir otro valor. La variable dependiente (ventas, en este caso) presenta algún grado de aleatoreidad. Por tanto, habrá algún *error* en el intento por explicar o predecir las ventas. Se dice que un modelo de esta naturaleza es estocástico, por la presencia de la variación aleatoria y puede expresarse como

Un modelo lineal	$Y = \beta_0 + \beta_1 X + \varepsilon$	[11.4]

La fórmula (11.4) es la relación poblacional (o verdadera) según la cual se hace regresión de Y sobre X. Además, $\beta_0 + \beta 1(X)$ es la porción determinística de la relación, mientras que ε (la letra griega epsilon) representa el carácter aleatorio que muestra la variable dependiente y por tanto denota el término del error en la expresión. Los parámetros β_0 y β_1, lo mismo que la mayoría de los parámetros, permanecerán desconocidos y se pueden estimar sólo con los datos muestrales. Esto se expresa así:

Un modelo lineal con base en datos muestrales	$Y = b_0 + b_1 X + e$	[11.5]

en donde los valores b_0 y b_1 son estimaciones de β_0 y β_1, respectivamente, y e es el término aleatorio. Habitualmente se le denomina *residual* cuando se utilizan datos muestrales, e reconoce que no todas las observaciones caen exactamente en una línea recta. Si se supiera el valor exacto de e, se podría calcular de manera precisa Y. Sin embargo, debido a que e es aleatoria, Y sólo puede estimarse. El modelo de regresión por ende toma la forma de:

El modelo de regresión estimada	$\hat{Y} = b_0 + b_1 X$	[11.6]

en donde $\hat{Y}$ (que se lee como Y sombrero) es el valor *estimado* de Y, y b_0 y b_1 son el intercepto y la pendiente de la recta de regresión estimada. Es decir, $\hat{Y}$ simplemente es el valor estimado para las ventas con base en el modelo de regresión.

11.3 Mínimos cuadrados ordinarios: La recta de mejor ajuste

El propósito del análisis de regresión es determinar una recta que se ajuste a los datos muestrales mejor que cualquier otra recta que pueda dibujarse. Para ilustrarlo, se asume que Vita + Plus, Inc., recolecta datos sobre los gastos publicitarios y los ingresos por ventas de 5 meses, como se muestra en la tabla 11.1.

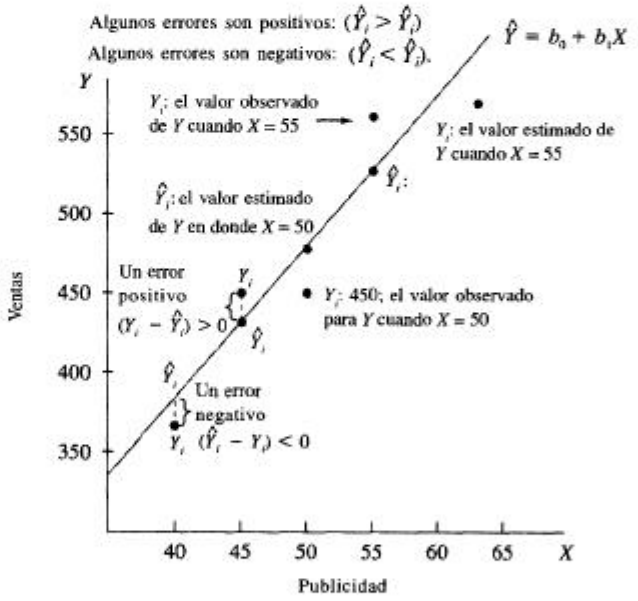
Tabla 11.1
Datos de ventas para Vita + Plus, Inc.

Mes	Ventas (X US$1000)	Publicidad (X US$100)
1	US$450	US$50
2	380	40
3	540	65
4	500	55
5	420	45

Aunque una muestra de sólo 5 datos probablemente sería insuficiente, servirá por el momento para los propósitos de los autores.

Estos cinco datos y la recta que mejor les ajusta aparecen en la figura 11.4. Esta recta está determinada mediante la estimación de b_0 y b_1. Un procedimiento matemático utilizado para estimar esos valores se denomina **mínimos cuadrados ordinarios (MCO)**. *MCO* producirá una recta que se extiende por el centro del diagrama de dispersión aproximándose a todos los puntos de datos más que cualquier otra recta. Regresando a la figura 11.4, para los 5 datos Y_i en el diagrama de dispersión son los valores de los datos observados reales para Y en la tabla 11.1. Los valores $\hat{Y}$ se obtienen mediante la recta de regresión y representan el estimado de las ventas. La diferencia entre lo que Y era realmente, Y_i, y lo que se estima que es $\hat{Y}_i$, es el error.

Figura 11.4
Datos para Vita + Plus, Inc.



El término del error es la diferencia entre los valores reales de $Y(Y_i)$, y el estimado de $Y(\hat{Y}_i)$.

$$\text{Error} = (Y_i - \hat{Y}_i) \qquad [11.7]$$

Si el valor real de Y, Y_i, es mayor que el estimado, entonces $(Y_i - \hat{Y}_i)$ y el error es positivo. Este es el caso en la figura 11.4, en donde la publicidad es 55. Por el contrario, si se sobrestiman las ventas, entonces $(Y_i - \hat{Y}_i)$ el error es negativo. Esto ocurre cuando la publicidad es 50. Debido a que algunos errores son negativos y algunos son positivos, MCO producirá una recta tal que la suma de esos errores sea cero:

$$\Sigma(Y_i - \hat{Y}_i) = 0$$

MCO también asegurará que se minimice la suma de estos errores al cuadrado. Es decir, si se toman cinco diferencias, todas verticales, entre los valores reales de Y y la recta de regresión $(Y_i - \hat{Y}_i)$, se elevan al cuadrado estas diferencias verticales y se suman, el número resultante será menor que el que se obtendría con cualquier otra recta. Es decir, MCO minimizará la *suma de los errores al cuadrado*. Es por esto que se denomina mínimos cuadrados ordinarios; produce una recta tal que la suma de los errores al cuadrado es menor de lo que sería con cualquier otra recta. Ver fórmula [11.8]

La suma de errores al cuadrado se minimiza	$\Sigma(Y_i - \hat{Y}_i)^2 = \min$	[11.8]

en donde $(Y_i - \hat{Y}_i)$, es el error de cada dato y *min* es el valor mínimo.

Para determinar esta *recta de mejor ajuste*, *MCO* requiere que se calcule la *suma de cuadrados* y *productos cruzados*. Es decir, se debe calcular la suma de los valores de X al cuadrado (SCx), la suma de los valores de Y al cuadrado (SCy) y la suma de X multiplicado por Y ($SCxy$). Esto se muestra en las fórmulas (11.9) a (11.11).

Suma de los cuadrados de X	$SCx = \Sigma(X_i - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$	[11.9]

Suma de los cuadrados de Y	$SCy = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$	[11.10]

y

Suma de los productos cruzados de X y Y	$SCxy = \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$	[11.11]

Vale la pena notar que las primeras porciones de cada una de estas fórmulas

$$SCx = \Sigma(X_i - \bar{X})^2$$
$$SCy = \Sigma(Y_i - \bar{Y})^2$$
y
$$SCxy = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})$$

ilustran cómo la recta MCO realmente se basa en las desviaciones de las observaciones a partir de su media. Por ejemplo, SCx se halla 1) calculando la cantidad en la cual cada una de las observaciones de $X(X_i)$ se desvía de su media (X), 2) elevando al cuadrado tales desviaciones, y 3) sumando tales desviaciones al cuadrado. Sin embargo,

estos cálculos son muy tediosos cuando se realizan manualmente. Por tanto, se utilizará generalmente la segunda versión de cada una de estas fórmulas al hacer los cálculos.

Dadas las sumas de cuadrados y los productos cruzados, es un asunto sencillo calcular la pendiente de la recta de regresión, llamada el *coeficiente de regresión* y el intercepto, así:

La pendiente de la recta de regresión	$b_1 = \frac{SCxy}{SCx}$	[11.12]

y

El intercepto de la recta de regresión	$b_0 = \bar{Y} - b_1\bar{X}$	[11.13]

en donde $\bar{Y}$ y $\bar{X}$ son las medias de los valores Y y los valores X.

Cuidado! Estos cálculos son extremadamente sensibles a la aproximación. Esto es especialmente cierto para el cálculo del coeficiente de determinación, lo cual se demostrará más tarde en este capítulo. Por tanto, se aconseja en aras de la exactitud, efectuar los cálculos hasta con cinco o seis cifras decimales.

Ejercicios de la sección

1. ¿Cuál es la diferencia entre la regresión simple y la regresión múltiple?
2. ¿Cuál es la diferencia entre la regresión lineal y la regresión curvilínea? ¿Cómo cambia Y cuando X cambia en cada caso?
3. Diferencie entre los componentes estocásticos y aleatorios de un modelo de regresión.
4. ¿Por qué el método de mínimos cuadrados ordinarios para determinar el modelo de regresión se denomina "mínimos cuadrados ordinarios"? ¿Qué papel juega el error en este análisis?
5. Identifique la variable dependiente y la independiente en cada uno de estos casos

   a. El tiempo que se pasa trabajando en una composición y la nota obtenida.

   b. La estatura del hijo y la estatura del padre.

   c. La edad de una mujer y el costo de su seguro de vida.

   d. El precio de un producto y el número de unidades vendidas.

   e. La demanda de un producto y el número de consumidores en el mercado.

6. Dados los siguientes datos para X y Y:

$$X \quad 28, 54, 67, 37, 41, 69, 76$$

$$Y \quad 14, 21, 36, 39, 18, 54, 52$$

a. Haga un diagrama de dispersión para los datos.

b. ¿Qué sugieren los datos sobre una relación entre X y Y?

c. Haga una recta para aproximar la relación.

7. ¿Cuál es la diferencia entre $\hat{Y}_i$ y Y_i en el análisis de regresión?

8. ¿Qué es el término ε en el modelo de regresión y por qué ocurre?

11.4 Ejemplo utilizando *MCO* (mínimos cuadrados ordinarios)

La gerencia de Hop Scotch Airlines, la aerolínea transportadora más pequeña del mundo, considera que existe una relación directa entre los gastos publicitarios y el número de pasajeros que escogen viajar por Hop Scotch. Para determinar si esta relación existe, y si es así cuál podría ser la naturaleza exacta, los estadísticos empleados por Hop Scotch decidieron utilizar los procedimientos MCO para determinar el modelo de regresión.

Se recolectaron los valores mensuales por gastos de publicidad y número de pasajeros para los $n = 15$ meses más recientes. Los datos aparecen en la tabla 11.2, junto con otros cálculos necesarios para hallar el modelo de regresión. Se observará que los *pasajeros* están representados con la variable Y, ya que se asume que depende de la publicidad.

Tabla 11.2
Datos de regresión para Hop Scotch Airlines

Observación (mes)	Publicidad (en US$1,000's) ($X$)	Pasajeros (en 1,000's) (Y)	XY	X^2	Y^2
1	10	15	150	100	225
2	12	17	204	144	289
3	8	13	104	64	169
4	17	23	391	289	529
5	10	16	160	100	256
6	15	21	315	225	441
7	10	14	140	100	196
8	14	20	280	196	400
9	19	24	456	361	576
10	10	17	170	100	289
11	11	16	176	121	256
12	13	18	234	169	324
13	16	23	368	256	529
14	10	15	150	100	225
15	12	16	192	144	256
	187	268	3,490	2,469	4,960

Con este simple conjunto de datos, y los cálculos subsiguientes para XY, X^2, y Y^2, es tarea fácil determinar el modelo de regresión mediante el cálculo de los valores de la constante de regresión y el coeficiente de regresión de la recta de regresión $\hat{Y} = b_0 + b_1X$. Las sumas de los cuadrados y de los productos cruzados son:

$$SCx = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$
$$= 2{,}469 - \frac{(187)^2}{15}$$
$$= 137.7333333$$
$$SCy = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$
$$= 4{,}960 - \frac{(268)^2}{15}$$
$$= 171.733333$$
$$SCxy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$
$$= 3{,}490 - \frac{(187)(268)}{15}$$
$$= 148.933333$$

Utilizando la fórmula (11.12) se puede establecer el coeficiente de regresión así:

$$b_1 = \frac{SCxy}{SCx}$$
$$= \frac{148.933333}{137.733333}$$
$$= 1.0813166 \text{ o } 1.08$$

Debido a que

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{268}{15} = 17.86667$$

y

$$\bar{X} = \frac{\Sigma X}{n} = \frac{187}{15} = 12.46667$$

La fórmula (11.13) revela que el intercepto es:

$$b_0 = \bar{Y} - b_1\bar{X}$$
$$= 17.866667 - 1.08(12.46667)$$
$$= 4.3865 \text{ o } 4.40$$

El modelo de regresión entonces es:

$$\hat{Y}_i = 4.40 + 1.08X_i$$

en donde $\hat{Y}_i$ es el valor individual pronosticado para los pasajeros. Así, si X_i es igual a 10, tendremos:

$$\hat{Y}_i = 4.40 + 1.08(10) = 15.2$$

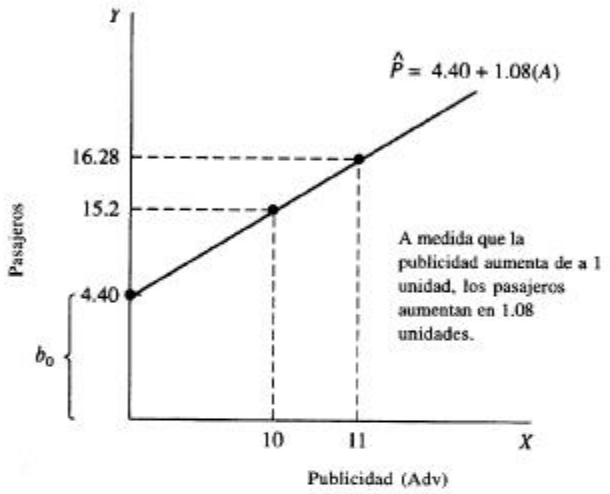
Debido a que tanto X como Y están expresadas en miles, esto significa que si se gastan US$10,000 en publicidad, el modelo predice que 15,200 personas valientes decidirán volar en Hop Scotch Airlines. El coeficiente de 1.08

significa que por cada incremento de una unidad en X, Y aumentará en 1.08 unidades. Por tanto, si se incrementan los gastos publicitarios en US$1.000, entonces 1.080 pasajeros más abordarán aviones de Hop Scotch.

$$\hat{Y}_i = 4.40 + 1.08(11) = 16.28$$

La figura 11.5 muestra la recta de regresión que se ha estimado. El intercepto es 4.40 y se indica una pendiente positiva. La pantalla 11.1 es una impresión parcial en Minitab. El intercepto o constante, como se denomina en Minitab, es 4.3863 y el coeficiente para la publicidad es 1.08. En breve se analizarán algunas de las estadísticas que aparecen impresas. La recta ajustada aparece en la pantalla de Minitab número 11.2. Vale la pena notar que la recta pasa por la mitad del diagrama de dispersión.

Figura 11.5
Recta de regresión para Hop Scotch Airlines



Pantalla de Minitab 11.1

Regression Analysis (Análisis de regresión)

The Regression equation is (La ecuación de regresión es)

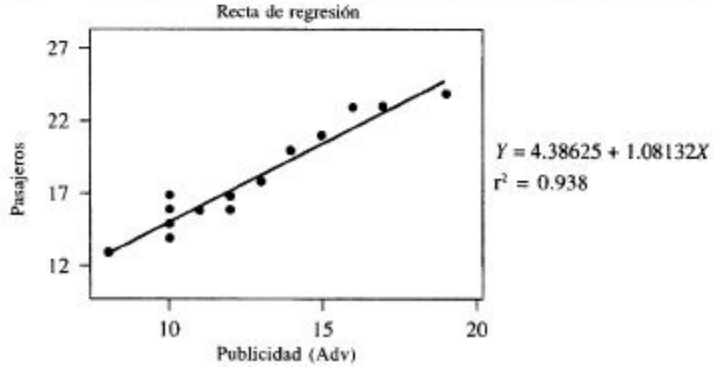
PASS = 4.39 + 1.08 ADV

Predictor	Coef	Stdev	t-ratio	p
Constant	4.3863	0.9913	4.42	0.001
ADV	1.08132	0.07726	13.99	0.000

s = 0.9068 **R-sq = 93.8%** R-sq (adj) = 93.3%

PASS = Pasajeros
Stdev = Desviación estándar
t-radio = Razón t
R-sq = r^2
ADV = Publicidad

Pantalla de Minitab 11.2



Ejercicios de la sección

9. El centro de ubicación laboral en State University desea determinar si los promedios puntuales en notas de los estudiantes (GPAs) puede explicar el número de ofertas laborales que ellos reciben después de graduarse. Los datos siguientes corresponden a los 10 recién graduados.

Estudiante	1	2	3	4	5	6	7	8	9	10
GPA	3.25	2.35	1.02	0.36	3.69	2.65	2.15	1.25	3.88	3.37
Ofertas	3	3	1	0	5	4	2	2	6	2

a. Haga un diagrama de dispersión para los datos.

b. Calcule e interprete el modelo de regresión. ¿Qué le dice este modelo sobre la relación entre GPA y las ofertas de trabajo?

c. Si Steve tiene un GPA de 3.22, ¿cuántas ofertas laborales pronostica usted que él recibirá?

10. Un economista del Departamento de Recursos Humanos de Florida State está preparando un estudio sobre el comportamiento del consumidor. Él recolectó los datos que aparecen en miles de dólares para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo. Determine cuál es la variable dependiente.

Consumidor	1	2	3	4	5	6	7	8	9	10	11	12
Ingreso	24.3	12.5	31.2	28.0	35.1	10.5	23.2	10.0	8.5	15.9	14.7	15
Consumo	16.2	8.5	15	17	24.2	11.2	15	7.1	3.5	11.5	10.7	9.2

a. Haga un diagrama de dispersión para los datos.

b. Calcule e interprete el modelo de regresión. ¿Qué le dice este modelo sobre la relación entre el consumo y el ingreso? ¿Qué proporción de cada dólar adicional que se gana se invierte en consumo?

c. ¿Qué consumo pronosticaría el modelo para alguien que gana US$27,500?

11. Un banco en Atlanta que se especializa en créditos para vivienda intenta analizar el mercado de finca raíz, midiendo el poder explicativo que las tasas de interés tienen sobre el número de casas vendidas en el área. Se compilaron los datos para un período de 10 meses, así:

Mes	1	2	3	4	5	6	7	8	9	10
Interés	12.3	10.5	15.6	9.5	10.5	9.3	8.7	14.2	15.2	12
Casas	196	285	125	225	248	303	265	102	105	114

a. Haga un diagrama de dispersión para los datos

b. Calcule e interprete el modelo de regresión. ¿Qué le dice este modelo sobre la relación entre las tasas de interés y las ventas de vivienda?

c. Si la tasa de interés es del 9.5%, ¿cuántas casas se venderían de acuerdo con el modelo?

12. Overland Group produce partes para camión que se utilizan en los semirremolques. El jefe de contabilidad desea desarrollar un modelo de regresión que pueda utilizarse para predecir los costos. Él selecciona unidades de producción fabricadas como una variable de predicción y recolecta los datos que se observan aquí. Los costos están en miles de dólares y las unidades en cientos.

Unidades	12.3	8.3	6.5	4.8	14.6	14.6	14.6	6.5
Costo	6.2	5.3	4.1	4.4	5.2	4.8	5.9	4.2

a. Haga un diagrama de dispersión para los datos.

b. Calcule e interprete el modelo de regresión. ¿Qué le dice el contador sobre la relación entre producción y costos?

c. Según el modelo, ¿cuánto costaría producir 750 unidades?

13. El profesor Mundane ha notado que muchos de sus estudiantes se han ausentado de clase este semestre. Considera que puede explicar esta falta de asistencia por las distancias a las que sus estudiantes viven del campus. Se practica una encuesta a once estudiantes sobre cuántas millas deben viajar para asistir a clase y el número de clases a las que han faltado.

Millas	5	6	2	0	9	12	16	5	7	0	8
Ausencias	2	2	4	5	4	2	5	2	3	1	4

a. Haga un diagrama de dispersión para los datos.

b. Compare e interprete el modelo de regresión. ¿Qué determina el profesor?

c. ¿A cuántas clases faltaría usted si viviera a 3.2 millas del campus, según el modelo?

14. El director administrativo de Bupkus, Inc., obtuvo datos sobre 100 empleados respecto a las pruebas de ingreso que se les practicó en el momento de la contratación y las calificaciones subsiguientes que recibieron los empleados por parte del supervisor un año después. Los puntajes del examen oscilaron entre 0 y 10 y la calificación era sobre un sistema de 5 puntos. El director intenta utilizar el modelo de regresión para predecir la clasificación (R) que recibirán con base en el puntaje del examen (S). Los resultados son:

$$\Sigma S = 522 \quad \Sigma R = 326 \quad \Sigma SR = 17{,}325$$
$$\Sigma S^2 = 28{,}854 \quad \text{y} \quad \Sigma R^2 = 10{,}781$$

Desarrolle e interprete el modelo de regresión. ¿Qué puede predecir el director respecto a la clasificación de un empleado que obtuvo 7 en el examen?

Nota: Mantenga sus cálculos de los ejercicios 9 a 14 para utilizarlos durante el resto de este capítulo. Utilizando los mismos datos usted evitará tener que calcular *SCx, SCy, y SCxy* cada vez. Usted ganará experiencia adicional con otros problemas al final del capítulo.

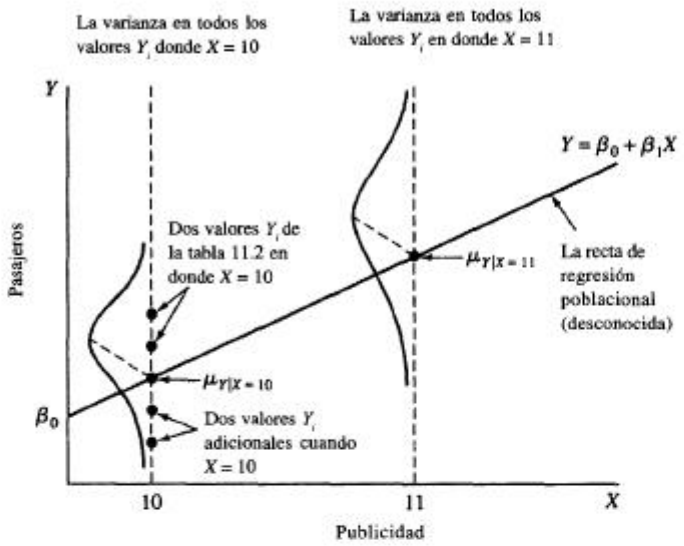
11.5 Supuestos del modelo de regresión lineal

Para comprender mejor el modelo lineal se deben examinar los cuatro supuestos sobre los cuales se construye.

Supuesto 1: El término de error ε es una variable aleatoria distribuida normalmente.

Como se dijo anteriormente, si se fija que X es igual a un valor dado muchas veces, los valores resultantes de Y variarán. Note las observaciones 1, 5, 7, 10 y 14 de la tabla 11.2. En cada caso $X = 10$. Sin embargo, Y_i es diferente cada vez. Algunas veces Y_i está por encima de la recta de regresión haciendo que el término de error $(Y_i - \hat{Y}_i)$ sea positivo mientras que en otros momentos Y_i es menor que $\hat{Y}_i$, creando un error negativo. Se asume que estos términos de error se distribuyen normal y aleatoriamente alrededor de la recta de regresión poblacional. Esto se muestra en la figura 11.6 para un valor X de 10.

Figura 11.6
La distribución normal de los valores y alrededor de la recta de regresión poblacional desconocida



Debido a que Y_i es diferente cada vez, lo mejor que la recta de regresión puede hacer es estimar el valor **promedio** de Y. Por tanto, la recta de regresión poblacional pasa por la media de aquellos valores Y en donde $X = 10$. Este punto se indica como $\mu_{Y|X=10}$ en la figura 11.6. Allí se observa una distribución normal de los términos de error por encima y por debajo de la recta de regresión. Lo mismo ocurriría si se tuviera $X = 11$ muchas veces. Resultarían muchos valores Y *diferentes*. Estos valores Y están distribuidos normalmente por encima y por debajo de la recta de regresión y pasan a través de la media de dichos valores donde $X = 11$, $\mu_{Y|X=11}$.

Supuesto 2: Varianzas iguales de los valores *Y*.

El modelo MCO asume que la varianza en los valores Y es la misma para todos los valores de X. Esto también se muestra en la figura 11.6. La variación en los valores Y por encima y por debajo de la recta de regresión en donde $X = 10$ es igual a la variación en los valores Y en donde $X = 11$. Esto es cierto para todos los valores de X. Este supuesto se denomina *homoscedasticidad*.

Homoscedasticidad	Las varianzas en los valores Y son las mismas en todos los valores de X

Desafortunadamente, este supuesto se contraviene con frecuencia cuando se trabaja con datos de corte seccional. Por ejemplo, se asume que se desea desarrollar un modelo de regresión en el cual los ingresos de los consumidores se utilicen para predecir o explicar sus gastos de consumo, Cons = f(Ing). Si se recolectaran datos sobre los consumidores en diferentes intervalos de ingreso durante un año dado se estarían utilizando datos de corte seccional ya que se incluyeron las observaciones a través de diferentes secciones de estrato de ingresos; los pobres, el promedio y los ricos. Como lo muestra la figura 11.7, se puede encontrar un rango muy estrecho en los valores para el consumo en los niveles bajos de ingreso, mientras que para los consumidores más ricos, la variación en sus gastos de consumo es mucho mayor. Los valores Y_i se dispersan más ampliamente a medida que el ingreso incrementa. Esto es lo que se denomina *heteroscedasticidad*.

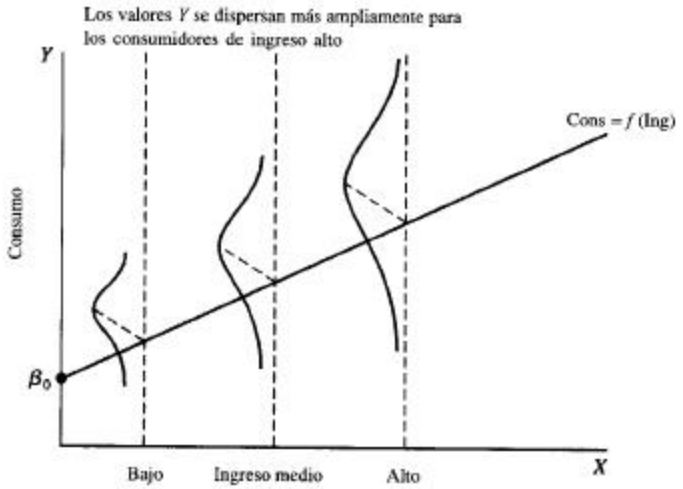


Figura 11.7
Heteroscedasticidad en la varianza de los valores Y

Supuesto 3: Los términos de error son independientes uno del otro.

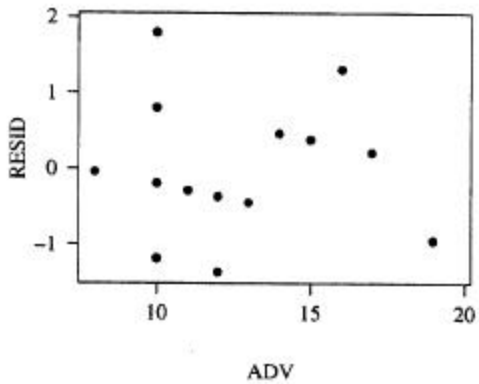
MCO se basa en el supuesto de que los términos de error son independientes uno del otro. El término de error encontrado para un valor de Y_i no se relaciona con el término de error para cualquier otro valor de Y_i. Esta hipótesis puede probarse analizando un diagrama de los errores de los datos muestrales. Si no puede observarse ningún patrón se puede asumir que los términos de error no se relacionan.

Pantalla en Minitab 11.3

Residual Data for Hop Scotch (Datos residuales de Hop Scotch)

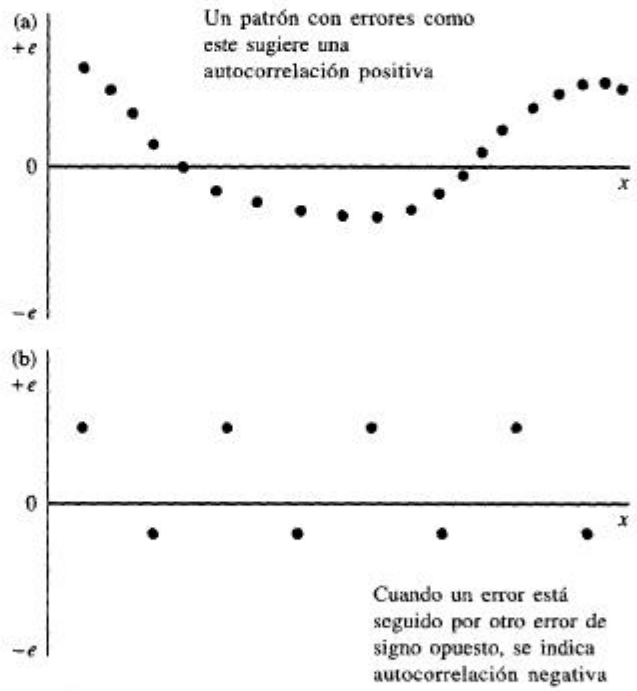
ADV	PASS	Y-HAT	RESID
10	15	15.1994	-0.19942
12	17	17.3621	-0.36215
8	13	13.0368	-0.03679
17	23	22.7686	0.23137
10	16	15.1994	0.80058
15	21	20.6060	0.39400
10	14	15.1994	-1.19942
14	20	19.5247	0.47532
19	24	24.9313	-0.93127
10	17	15.1994	1.80058
11	16	16.2807	-0.28074
13	18	18.4434	-0.44337
16	23	21.6873	1.31268
10	15	15.1994	-0.19942
12	16	17.3621	-1.36205

ADV = Advertising o publicidad
Pass = Passanger o pasajeros
Y-HAT = $\hat{Y}$
RESID = ERROR



El diagrama de los errores y sus valores para los datos de Hop Scotch se muestran en la pantalla 11.3 de Minitab. Es difícil detectar algún patrón discernible. Esto sugiere que los errores realmente son independientes. Se compara esto con un diagrama residual que puede aparecer como figura 11.8 a). Es evidente que los errores no son aleatorios y que están relacionados claramente. El patrón comienza con varios errores positivos, seguidos por varios errores negativos, y luego nuevamente varios errores positivos. Por el contrario, si se tuviera que lanzar una moneda varias veces, ¿se obtendrían varias caras seguidas de varios sellos y luego nuevamente varias caras? Es altamente improbable. Mientras que los lanzamientos de la moneda son eventos independientes, estos residuales no lo son. Están relacionados. Se puede decir que el valor de un error es una función del error anterior. Es más probable que un error positivo sea seguido por otro error positivo, mientras que un error negativo está relacionado con un segundo error negativo. Tal condición, la cual contraviene el supuesto de independencia de errores, se denomina **autocorrelación positiva** porque los signos iguales se agrupan. La **autocorrelación negativa** se representa en la figura 11.8b). Cada error es seguido de un error de signo opuesto. Este patrón de signos alternantes sugiere que los términos de error no son independientes.

Figura 11.8
Posibles diagramas de errores



Autocorrelación Ocurre cuando los términos de error no son independientes.

En realidad, los diagramas residuales nunca son tan obvios o tan fáciles de leer como lo pueden sugerir los diagramas anteriores. Afortunadamente existe una forma más confiable para detectar la autocorrelación con base en la **prueba de Durbin-Watson**. Es más probable que la autocorrelación ocurra en el uso de datos de *series de tiempo* en los cuales, a diferencia de los datos de corte seccional discutidos anteriormente, las observaciones de alguna variable se recolectan durante varios períodos (semanas, meses, años, etc.). Por ejemplo, se pueden

compilar datos para la tasa de desempleo mensual durante varios meses. Estos datos difieren de los datos de corte seccional, los cuales se recolectan para algún punto específico en el tiempo. El estadístico de Durbin-Watson se calcula así

Estadístico de Durbin-Watson	$d = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2}$	[11.14]

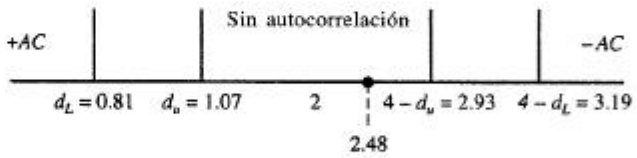
en donde e_t es el error en el período de tiempo t y e_{t-1} es el error en el período anterior. La fórmula (11.14) requiere que el término de error $(Y_i - \hat{Y}_i)$ se calcule para cada período y es muy difícil calcularlo manualmente. Un computador en Minitab reveló que el estadístico de Durbin-Watson para los datos Hop Scotch es de 2.48. Este valor se utiliza para probar la hipótesis de que no existe correlación entre términos de error sucesivos, así:

$$H_0: \rho_{\varepsilon_t, \varepsilon_{t-1}} = 0 \text{ (No existe autocorrelación)}$$

$$H_A: \rho_{\varepsilon_t, \varepsilon_{t-1}} \neq 0 \text{ (Existe autocorrelación)}$$

en donde ρ es el coeficiente de correlación para errores sucesivos. El valor Durbin-Watson se compara con los valores críticos tomados de la tabla *K* del apéndice III, para un nivel de significancia del 1 o del 5%. Se supone que se selecciona un valor α del 1%. Dado que $n = 15$, y k el número de variables independientes es 1, el valor Durbin-Watson inferior es $d_L = 0.81$, y el valor superior Durbin-Watson es $d_U = 1.07$. Estos valores se aplican entonces a la escala en la figura 11.9. Si el valor Durbin-Watson es menor que $d_L = 0.81$, se sugiere una autocorrelación positiva y se rechaza la hipótesis nula. Si es mayor que $(4 - d_L) = 3.19$, se sugiere la autocorrelación negativa y se rechaza la hipótesis nula. Si está entre $d_u = 1.07$ y $(4 - d_U) = 2.93$, no se rechaza la hipótesis nula. Si el valor Durbin-Watson cae en cualquiera de las dos regiones restantes, la prueba no es conclusiva. En este caso particular, el valor Durbin-Watson de 2.48 cae en la región de la escala que indica que la correlación no existe y no se rechaza la hipótesis nula. Generalmente hablando, si el valor Durbin-Watson es cercano a 2, no se rechaza la hipótesis nula.

Figura 11.9
Una prueba Durbin-Watson



Supuesto 4: El supuesto de linealidad.

Como se expresó en el supuesto 1, si *X* se deja igual que un valor muchas veces, ocurrirá una distribución normal de los valores *Y*. Esta distribución tiene una media, $\mu_{Y|X}$. Esto es cierto para todo valor de *X*. *MCO* asume que estas medias quedan en una recta, como se sugirió anteriormente en la figura 11.6.

Ejercicios de la sección

15. ¿Qué se entiende por *homoscedasticidad* y *heteroscedasticidad*? Haga los gráficos apropiados para ilustrar estos dos términos.

16. ¿Qué se entiende por autocorrelación? Diferencie entre una *autocorrelación positiva* y una *autocorrelación negativa*. Realice las gráficas que representan estos dos tipos.

17. Explique claramente cómo se utiliza la prueba de Durbin-Watson para probar la autocorrelación. Incluya una discusión de la naturaleza de la fórmula utilizada para calcular el estadístico Durbin-Watson.

18. ¿Cuál es la naturaleza del supuesto de linealidad sobre el modelo MCO?

11.6 El error estándar de estimación: Una medida de bondad de ajuste

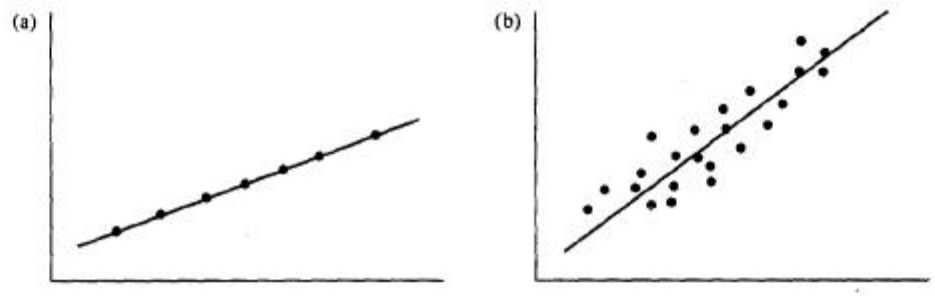
La recta de regresión, como ya se ha dicho, con frecuencia se le denomina la recta del ajuste óptimo. Se ajusta o representa la relación entre X y Y mejor que cualquier otra recta. Sin embargo, debido a que simplemente proporciona el mejor ajuste, no existe garantía de que sea buena. A los autores les encantaría poder medir qué tan bueno es el mejor ajuste.

En realidad, hay por lo menos dos medidas de bondad de ajuste: 1) el error estándar de estimación, y 2) el coeficiente de determinación. Por ahora veremos el análisis de este último concepto hasta que más adelante, en este capítulo, se estudie el análisis de correlación. Trataremos la descripción del error estándar de estimación en este momento.

El **error estándar de estimación**, Se, es una medida del grado de dispersión de los valores Y_i alrededor de la recta de regresión. Mide la variación de los puntos de datos por encima y por debajo de la recta de regresión. Refleja la tendencia a desviarse del valor real de Y cuando se utiliza el modelo de regresión para fines predictivos. En este sentido, es una medida del error "típico".

Si todos los puntos de datos se situaran perfectamente sobre una recta como en la figura 11.10 a), la recta de regresión pasaría por cada uno. En este caso afortunado no se presentarán errores en los pronósticos, y el error estándar de estimación sería cero. Sin embargo, los datos rara vez son así de cooperativos y usualmente habrá alguna dispersión en los datos como en la figura 11.10 b). El error estándar de estimación mide esta variación

Figura 11.10
Diagramas de dispersión posibles



promedio de los puntos de datos alrededor de la recta de regresión que se utiliza para estimar Y y por ende proporciona una medida del error que se presentará en dicha estimación. La fórmula (11.15) ilustra este principio. Vale la pena destacar que el numerador refleja la diferencia entre los valores reales de Y, Y_i, y el estimado $\hat{Y}_i$:

El error estándar de estimación

$$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2}} \qquad [11.15]$$

Desafortunadamente, la fórmula (11.15) no es conveniente a nivel computacional. Es necesario disponer de un método más fácil de calculo manual. Vale la pena recordar que σ^2 es la varianza de los errores de regresión. Uno de los supuestos básicos del modelo *MCO* es que esta varianza en los errores alrededor de la recta de regresión es la misma para todos los valores de *X*. Entre menor sea el valor de σ^2, menor será la dispersión de los puntos de datos alrededor de la recta.

Debido a que σ^2 es un parámetro, probablemente permanecerá desconocida, y es necesario estimar su valor con los datos muestrales. Una estimación insesgada de σ^2 es el cuadrado medio del error (*CME*). En el capítulo anterior sobre ANOVA , se aprendió que el *CME* es la suma del cuadrado del error (*SCE*) dividida por sus grados de libertad. A diferencia del análisis de regresión, *SCE* es

La suma de cuadrados del error	$SCE = SCy - \frac{(SCxy)^2}{SCx}$	[11.16]

En un modelo de regresión simple, se imponen dos restricciones en el conjunto de datos, debido a que se deben estimar dos parámetros, β_0 y β_1. Por tanto hay n - 2 grados de libertad y *CME* es:

Cuadrado medio del error	$CME = \frac{SCE}{n-2}$	[11.17]

El error estándar de estimación es entonces

El error estándar	$Se = \sqrt{CME}$	[11.18]

En el caso actual de Hop Scotch Airlines, se tiene que:

$$SCE = SC_y - \frac{(SC_{xy})^2}{SC_x}$$
$$= 171.7333 - \frac{(148.9333)^2}{137.73333}$$
$$= 10.6893$$
$$CME = \frac{10.6893}{15 - 2}$$
$$= 0.82226$$
$$S_e = \sqrt{0.82226}$$
$$= 0.90678 \text{ o } 0.907$$

La pantalla de Minitab 11.4 muestra que el error estándar es 0.9068.

Pantalla en Minitab 11.4

Regression Analysis (Análisis de regresión)

```
The regression equation is (La ecuación de regresión es)
PASS = 4.39 + 1.08 ADV

Predictor          Coef        Stdev      t-ratio        p
Constant         4.3863       0.9913         4.42    0.001
ADV              1.08132      0.07726       13.99    0.000

s = 0.9068          R-sq = 93.8%          R-sq(adj) = 93.3%

Analysis of Variance (Análisis de varianza)

SOURCE          DF          SS          MS          F          P
Regression       1      161.04      161.04     195.86      0.000
Error           13       10.69        0.82
Total           14      171.73

Unusual Observations
obs.    ADV       PASS        Fit     Stdev.Fit      Residual    St.Resid
10     10.0     17.000     15.199         0.302         1.801       2.11R

R denotes an obs. with a large st. resid.

Durbin-Watson statistic = 2.48

            Fit      Stdev.Fit          95.0% C.I.          95.0% P.I.
         15.199          0.302     (14.547, 15.852)    (13.134, 17.265)
MTB >
```

Pass = Pasajeros
Adv = Publicidad
Stdev = Desviación estándar
t-ratio = razón t
R-sq = r^2
DF = Grados de libertad
SS = Suma de cuadrados
MS = Cuadrado medio
Fit = Ajuste
CI = Intervalo de confianza

El error estándar siempre se expresa en las mismas unidades que la variable dependiente Y, en este caso miles de pasajeros. Por tanto, el error estándar de 0.907, o 907 pasajeros mide la variabilidad de los valores Y alrededor de la recta de regresión ajustada.

El error estándar de estimación es muy similar a la desviación estándar de una sola variable que se analizó en el capítulo 3. Si se recolectaran datos sobre los ingresos de $n = 100$ personas, se podría calcular fácilmente la desviación estándar. Esto proporcionaría una medida de dispersión de los datos de ingresos alrededor de su media.

En análisis de regresión se tienen dos variables, X y Y. El error estándar de estimación es una medida de la dispersión de los valores Y alrededor de su media, dado un valor X específico.

Como el error estándar de estimación es similar a la desviación estándar para una sola variable, puede interpretarse de manera similar. Vale la pena recordar que la regla empírica establece que si los datos están distribuidos normalmente, un intervalo de una desviación estándar por encima de la media y una desviación estándar por debajo de la media comprenderá el 68.3% de todas las observaciones; un intervalo de dos desviaciones

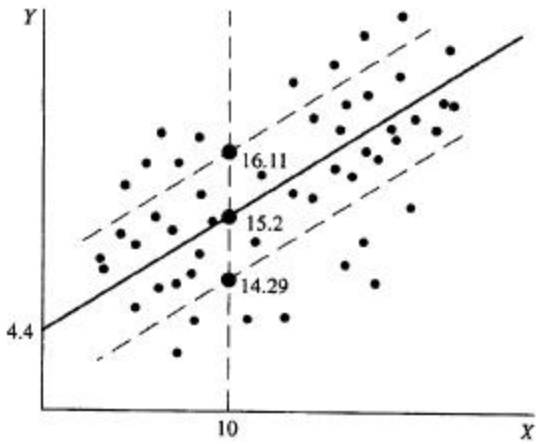
estándar a cada lado de la media contiene el 95.5% de las observaciones; y tres desviaciones estándar a cada lado de la media comprenden el 99.7% de las observaciones.

Lo mismo puede decirse del error estándar de estimación. En el ejemplo actual, en donde $X = 10$,

$$\hat{Y}_i = 4.4 + 1.08(10)$$
$$= 15.2$$

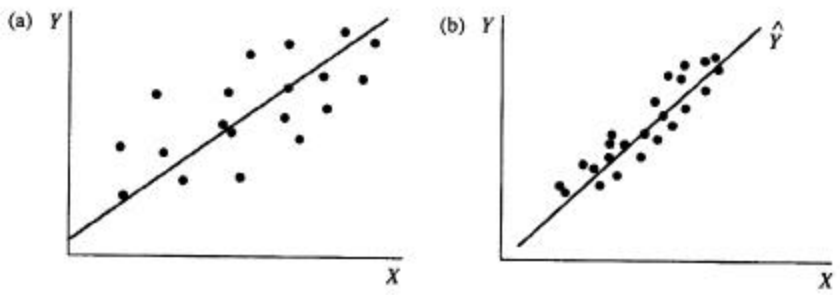
Vale la pena recordar que este valor de 15.2 es el estimado del valor promedio que se obtendría para Y si se determinara que X es igual a 10 muchas veces. Para ilustrar el significado del error estándar de estimación, se localizan los puntos que están a un *Se* (es decir, 0.907) por encima y por debajo del valor promedio de 15.2. Estos puntos son 14.29 (15.2 - 0.907) y 16.11 (15.2 + 0.907). Si se trazan las rectas pasando por cada punto paralelo a la recta de regresión como en la figura 11.11, aproximadamente el 68.3% de los puntos de datos caerán dentro de estas rectas. El 31.7% de las observaciones restantes estarán fuera de este intervalo. En este caso, el 68.3% de las veces cuando se invierte US$10,000 en publicidad, el número de pasajeros estará entre 14,290 y 16,110. El 31.7% del tiempo restante, el número de pasajeros excederá de 16,110 o será menor que 14,290.

Figura 11.11
Error estándar de estimación



Dada esta interpretación de *Se*, se deduce que entre más dispersos estén los datos originales, mayor será *Se*. Los datos para la figura 11.12 a) están mucho más dispersos que los de la figura 11.12 b). El *Se* para la figura 11.12 a) por tanto, sería mayor. Después de todo, si se van a involucrar el 68.3% de las observaciones a un *Se* de la recta de regresión, el intervalo debe ser más amplio si los datos están más dispersos.

Figura 11.12
Una comparación del error estándar de estimación



Ejercicios de la sección

19. Utilizando sus cálculos del ejercicio 9, calcule e interprete el error estándar de estimación para State University. Haga una gráfica en la interpretación. ¿Cómo puede utilizarse como medida de bondad de ajuste?

20. Con base en los datos del ejercicio 10, ¿cuál es el error estándar de estimación para el Departamento de Recursos Humanos de Florida State? ¿Cómo interpretaría los resultados? Utilice una gráfica.

21. Calcule e interprete el error estándar de estimación para el ejercicio 11 sobre el banco de Atlanta.

22. The Overland Group, del ejercicio 12, ahora desea conocer el error estándar de estimación.

23. ¿Cuál es el error estándar de estimación que va a experimentar el profesor del ejercicio 13?

11.7 Análisis de correlación

El modelo de regresión ha proporcionado un panorama claro de la relación entre los gastos publicitarios de Hop Scotch Airlines y el número de valientes viajeros que hacen fila en el puesto expendedor de pasajes. El valor positivo para b_1 indica una relación directa. A medida que la publicidad aumenta, también lo hace el número de pasajeros. Ahora es útil obtener una medida de la fuerza de esa relación. Esta es la función del **coeficiente de correlación**, desarrollado por Carl Pearson a finales de siglo, y algunas veces se le llama el coeficiente de correlación producto-momento de Pearson. Representado con una r, el coeficiente de correlación puede asumir cualquier valor entre –1 y + 1; es decir,

$$-1 \leq r \leq +1$$

Un valor de r = –1 indica una relación negativa perfecta entre X y Y, tal como se observa en la figura 11.13 a). Todas las observaciones quedan en una línea recta perfecta con una pendiente negativa. Por tanto, X y Y se moverán en direcciones opuestas. La figura 11.13 b) muestra una relación positiva perfecta entre X y Y con $r = +$

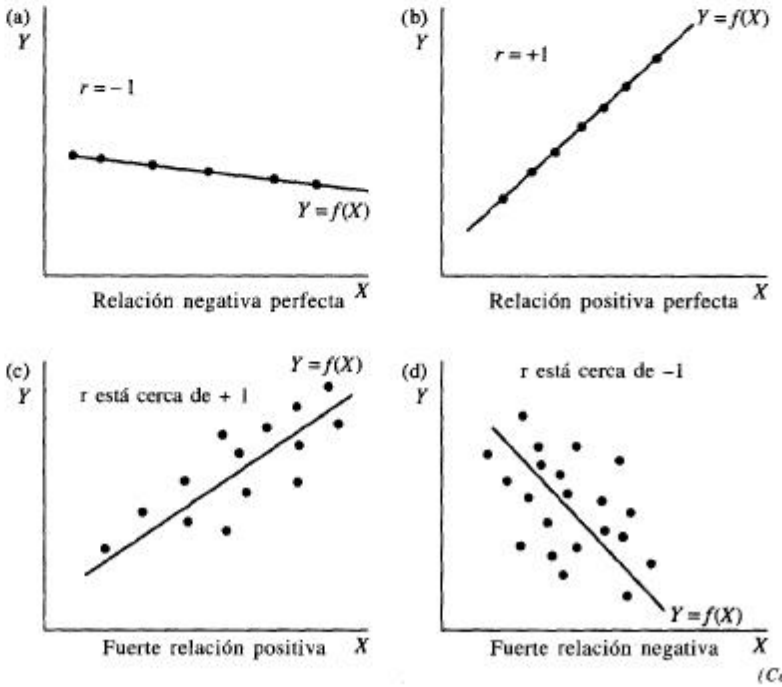
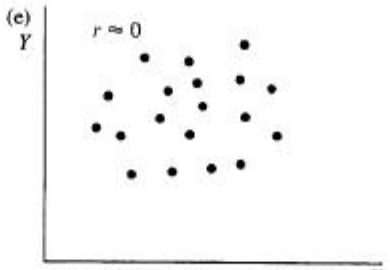


Figura 11.13
Posibles valores para el coeficiente de correlación r

(Continúa)

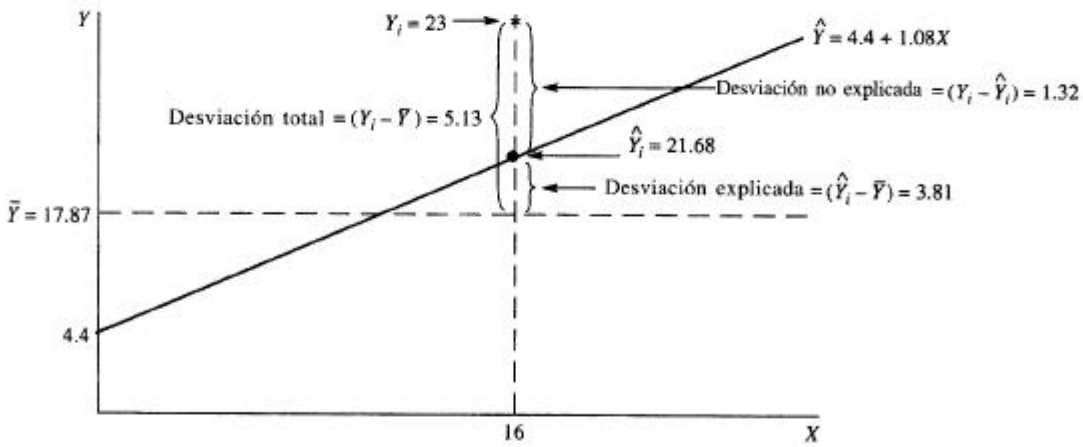


Poca o ninguna relación; no se puede dibujar una recta que ajuste

1. Como se anotó anteriormente, en toda relación entre dos variables existe la posibilidad de que exista alguna variación alrededor de la recta de regresión. Esto se observa en las figuras 11.13 c) y 11.13 d), las cuales muestran relaciones fuertes pero menos perfectas. En ambos casos el valor absoluto de r se aproxima a 1. Por el contrario, la figura 11.13 e) muestra muy poca o ninguna relación entre X y Y, y r se aproxima a cero. En general, entre mayor sea el valor absoluto de r, más fuerte será la relación entre X y Y.

Para comprender plenamente lo que mide el coeficiente de correlación, se deben desarrollar tres medidas de desviación. La **desviación total** de Y es la cantidad por la cual los valores individuales de Y, (Y_i) varían de su media $\overline{Y}$, $(Y_i - \overline{Y})$. Utilizando como ejemplo el mes 13 de los datos sobre Hop Scotch de la tabla 11.2, la figura 11.14 muestra la desviación total como (23 - 17.87) = 5.13. El valor de Y_i de 23 queda 5.13 por encima de la recta horizontal que representa a $\overline{Y}$ de 17.87. Si todas la $n = 15$ desviaciones totales se calculan y se elevan al cuadrado, y los resultados se suman, entonces se tendrá la *suma de cuadrados de las desviaciones totales, SCT.*

Figura 11.14
Desviaciones para Hop Scotch Airlines



Suma de cuadrados total	$SCT = \Sigma(Y_i - \overline{Y})^2$	[11.19]

Esta desviación total, como se observa en la figura 11.14, puede dividirse en desviación explicada y desviación no explicada. La **desviación explicada** es la diferencia entre lo que predice el modelo de regresión $\hat{Y}_i$, y el valor

promedio de Y, $(\hat{Y}_i - \bar{Y})$. Es esa porción de la desviación total la que es explicada por el modelo de regresión. En $X = 16, \hat{Y}_i = 4.4 + 1.08\,(16) = 21.68$. La figura 11.14 muestra que la desviación explicada es (21.68 - 17.87) = 3.81. Si esta porción de la desviación total, la cual es explicada por el modelo de regresión, se calcula para todas las $n = 15$ observaciones y luego se eleva al cuadrado, y los resultados se suman, entonces se tendrá la *suma de cuadrados de la regresión, SCR.*

Suma de cuadrados de la regresión	$SCR = \Sigma(\hat{Y}_i - \bar{Y})^2$	[11.20]

La **desviación no explicada** que se observa en la figura 11.14 es esa porción de la desviación total que no es explicada por el modelo de regresión. Es decir, es el error $(Y_i - \hat{Y}_i)$. La figura 11.14 muestra que es (23 - 21.68) = 1.32. Si se calculan estos términos de error para todos los $n = 15$ observaciones y luego se eleva al cuadrado y luego los resultados se suman, se tendrá la suma del cuadrado del error, *SCE.*

Suma del cuadrado del error	$SCE = \Sigma(Y_i - \hat{Y}_i)^2$	[11.21]

El coeficiente de correlación se calcula así:

Coeficiente de correlación	$r = \sqrt{\dfrac{\text{Variación explicada}}{\text{Variación total}}} = \sqrt{\dfrac{SCR}{SCT}}$	[11.22]

Se observa precisamente lo que r está midiendo. Está comparando el monto total de la desviación alrededor de $\bar{Y}$, *SCT*, con la porción de ésta que está explicada por el modelo de regresión *SCR*. Como raíz cuadrada de *SCR/SCT*, el coeficiente de correlación proporciona una medida relativa de la capacidad del modelo para explicar las desviaciones en los valores Y_i. Por ende mide la fuerza de la relación entre Y y la variable explicativa X.

La fórmula (11.22) es difícil de calcular manualmente. Una fórmula más conveniente es

Forma de calcular el coeficiente de correlación	$r = \dfrac{SCxy}{\sqrt{(SCx)(SCy)}}$	[11.23]

En el caso de Hop Scotch, se tiene que

$$r = \frac{148.93333}{\sqrt{(137.7333)(171.7333)}} = 0.9683$$

Esto indica una relación positiva fuerte entre los pasajeros y la cantidad de dinero invertido en fines publicitarios.

Vale la pena recordar que el error estándar de estimación *Se*, que se calculó anteriormente, es una medida de la bondad de ajuste. Proporciona una medida cuantificable de qué tan bien se ajusta el modelo a los datos que se han recolectado.

El **coeficiente de determinación** r^2 es otra medida quizá más importante de la bondad de ajuste. Se halla

Coeficiente de determinación: una medida de bondad de ajuste	$r^2 = \sqrt{\frac{\text{desviación explicada}}{\text{desviación total}}} = \sqrt{\frac{SCR}{SCT}}$	[11.24]

Una fórmula más conveniente de cálculo es

Fórmula computacional para el coeficiente de determinación	$r^2 = \frac{(SCxy)^2}{\sqrt{(SCx)(SCy)}}$	[11.25]

Proporciona una medida de bondad de ajuste porque revela qué porcentaje del cambio en Y se explica por un cambio en X.

El coeficiente de determinación para Hop Scotch es:

$$r^2 = \frac{(SCxy)^2}{(SCx)(SCy)}$$
$$= \frac{(148.9333)^2}{(137.7333)(171.7333)}$$
$$= 0.93776 \text{ o } 0.94$$

Como se puede esperar, r^2 puede determinarse más fácil, simplemente elevando al cuadrado el coeficiente de correlación r.

$$r^2 = (0.9683)^2 = 0.94$$

Esto establece que el 94% del cambio en el número de pasajeros se explica mediante un cambio en la publicidad. La pantalla de Minitab 11.4, presentada anteriormente, muestra que r^2 es 93.8%.

Este r^2 tiene significado sólo para las relaciones lineales. Dos variables pueden tener un r^2 de cero y sin embargo estar relacionadas en sentido curvilíneo. Además, no se interpreta este valor como si el 94% del cambio en los pasajeros fuera causado por un cambio en la publicidad. La correlación no significa causa. Este asunto se enfatiza en la siguiente sección.

Ejercicios de la sección

24. ¿Cómo puede utilizarse el coeficiente de determinación como medida de bondad de ajuste? Realice la gráfica para ilustrar.

25. ¿Cuál es la fuerza de la relación entre GPA y las ofertas de trabajo en el ejercicio 9?

26. Calcule e interprete el coeficiente de correlación y el coeficiente de determinación para el Departamento de Recursos Humanos de Florida State en el ejercicio 10.

27. ¿Cuántos cambios en las casas vendidas pueden explicarse por la tasa de interés del ejercicio 11?

28. ¿Cuál es la fuerza del modelo del profesor Mundane utilizado en el ejercicio 13 para explicar las ausencias de los estudiantes?

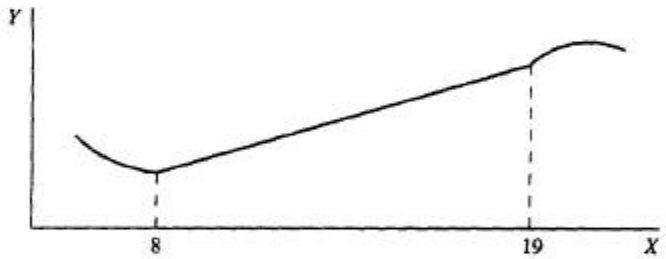
11.8 Limitaciones del análisis de regresión

Aunque los análisis de regresión y correlación con frecuencia han demostrado ser de utilidad en la toma de decisiones para una gran variedad de negocios y de asuntos económicos, existen ciertas limitaciones en su aplicación e interpretación. Estos no pueden determinar relaciones causa-efecto. La correlación no implica causalidad. Este punto fue elaborado dramáticamente por un estadístico británico quien "probó" que las cigüeñas traían a los bebés. Él recolectó datos sobre las tasas de natalidad y el número de cigüeñas en Londres y descubrió una correlación muy alta - algo así como $r = 0.92$. Por tanto, él concluyó que la historia sobre las cigüeñas y los bebés era cierta.

Sin embargo, como ya se habrá sospechado, esta no es la forma como funciona la correlación. Parece que a esta especie de cigüeñas les gusta anidar en la parte superior de la chimenea de los londinenses. Por tanto, cuando la población era densa y la tasa de natalidad era alta, habían muchas chimeneas para atraer a estas aves, de allí la alta correlación entre la tasa de natalidad y las cigüeñas. En realidad, tanto las cigüeñas como los nacimientos eran *causados* por un tercer factor, la densidad poblacional, que el investigador ignoró a su conveniencia. Vale la pena recordar que la correlación no significa causalidad.

Adicionalmente, se debe tener cuidado de no utilizar el modelo de regresión para predecir Y para valores de X que estén fuera del rango del conjunto original de datos. Los valores para X en el conjunto de datos de Hop Scotch oscilan desde tan bajo como 8 hasta tan alto como 19. Se ha aislado la relación entre X y Y sólo para ese rango de valores X. No se sabe cuál es la relación fuera de ese rango. Todo lo que se sabe es que puede aparecer como en la figura 11.15. Como se puede observar, para los valores fuera del rango de 8 a 19, la relación $X - Y$ es totalmente diferente de lo que se puede esperar dado el ejemplo.

Figura 11.15
Una posible relación $X - Y$



Otra falla del análisis de correlación y de regresión se hace evidente cuando dos variables no relacionadas parecen presentar alguna relación. Se asume que se desea analizar la correlación entre el número de elefantes nacidos en el zoológico de Kansas City y las toneladas de trucha que recogen los pescadores del Golfo de Tallahassee, Florida. Se halla $r = 0.91$. ¿Se concluirá que existe una relación? Tal conclusión es obviamente disparatada. A pesar del valor r, la simple lógica indica que no hay relación entre estas dos variables. No se ha abarcado la **correlación espúrea**, que es la correlación que ocurre simplemente por suerte. No hay sustituto para el sentido común en el análisis de correlación ni en el de regresión.

11.9 Pruebas para los parámetros poblacionales

Los resultados estadísticos sugieren una relación entre los pasajeros y la publicidad de Hop Scotch Airlines. Los valores que no son cero para el coeficiente de regresión (pendiente) de $b_1 = 1.08$ y el coeficiente de correlación de $r = 0.968$ indican que, a medida que los gastos publicitarios cambian, cambia el número de pasajeros.

Sin embargo, estos resultados se basan en una muestra de sólo $n = 15$ observaciones. Como siempre se pregunta si ¿existe alguna relación a nivel poblacional? Podría ser que debido al error de muestreo los parámetros

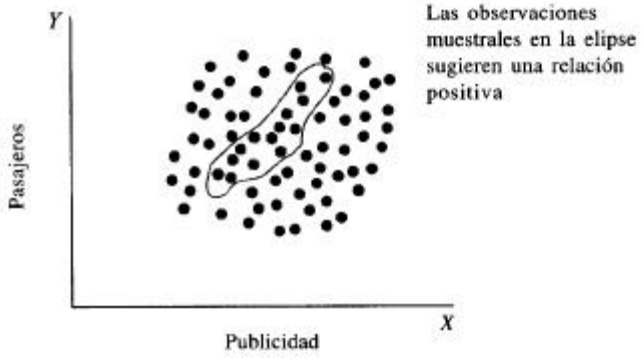
poblacionales son cero y se deben probar los parámetros poblacionales para asegurar que el hallado estadístico difiere *significativamente* de cero.

A. Pruebas para β_1

Si la pendiente de la recta de regresión poblacional real pero desconocida es cero, no existe relación entre los pasajeros y la publicidad contraria a los resultados muestrales. Si se hace un diagrama de dispersión para la población de todos los puntos de datos *X*, *Y*, puede aparecer como la figura 11.16. La ausencia de cualquier patrón indica que no existe relación. Al recolectar la muestra, se puede haber incluido sólo 15 de tales observaciones dentro de la elipse. Tomados por sí solos, estos datos sugieren de manera falsa una relación positiva. Se debe probar la hipótesis:

$$H_o: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Figura 11.16
Un diagrama de dispersión para la población de todos los puntos *X*-*Y*



Esta prueba emplea el estadístico *t*

La prueba *t* para el coeficiente de regresión poblacional	$t = \dfrac{b_1 - \beta_1}{s_{b_1}}$	[11.26]

y tiene $n - 2$ grados de libertad, en donde s_{b1} es el error estándar de la distribución muestral de b_1. Reconoce que muestras diferentes dan valores diferentes para b_1. Por tanto, si β_1 es realmente cero, estos valores para b_1 se distribuirían alrededor de cero como se muestra en la figura 11.17. Se puede calcular s_{b_1} mediante

Error estándar del coeficiente de regresión	$s_{b_1} = \dfrac{Se}{\sqrt{SCx}}$	[11.27]

Dados los valores para Hop Scotch,

$$s_{b_1} = \frac{0.907}{\sqrt{137.73333}} = 0.07726$$

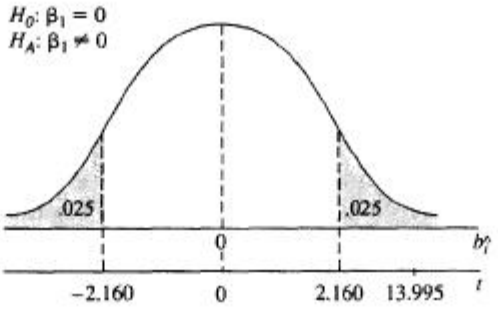
y

$$t = \frac{1.0813 - 0}{0.07726} = 13.995$$

Si se selecciona un valor α del 5%, $t_{0.05,13} = \pm 2.160$. La regla de decisión es:

Regla de decisión: "No rechazar si t está entre $\pm$ 2.160, de lo contrario rechazar".

Figura 11.17
Distribución de b_1 si $\beta_1 = 0$



Debido a que $t = 13.995$, la hipótesis nula de que $\beta_1 = 0$ se rechaza. Al nivel del 5% parece existir una relación entre pasajeros y publicidad. Esto puede confirmarse mediante la pantalla de Minitab 11.4 (que se mostró anteriormente). El valor p de 0.000 también aparece en la pantalla.

Si la hipótesis nula no hubiera sido rechazada, se concluiría que la publicidad y los pasajeros no están relacionados. Descartando el modelo, se utilizaría una variable explicativa diferente.

Debido a que se ha rechazado la hipótesis nula de que $\beta_1 = 0$, la pregunta natural es, "¿Cuál es su valor?" Esta pregunta puede responderse calculando un intervalo de confianza (I.C.) para β_1.

I.C. para	$\beta_1 = b_1 \pm t(s_{b_1})$	[11.28]

Si se utiliza un nivel de confianza del 95%,

$$\text{I.C. para } \beta_1 = 1.08 \pm (2.160)(0.07726)$$
$$0.913 \leq \beta_1 \leq 1.247$$

Esto significa que se puede estar 95% seguro de que el coeficiente de regresión para toda la población de todos los valores X, Y está entre 0.913 y 1.247.

B. Pruebas para el coeficiente de correlación poblacional ρ

Gran parte del trabajo realizado para probar las inferencias sobre el coeficiente de regresión puede aplicarse al coeficiente de correlación. El propósito y los fundamentos son muy similares. Como el análisis respecto a la correlación entre pasajeros y publicidad se basa en los datos muestrales, el error de muestreo podría llevarnos a conclusiones no apropiadas. Es decir, los datos muestrales produjeron un coeficiente de correlación de no cero de $r = 0.9683$ debido al error de muestreo como el de la figura 11.16. Puede ser que la correlación a nivel poblacional sea cero y que una muestra engañosa hizo que se asumiera equivocadamente una relación. Por consiguiente se debe probar la hipótesis

$$H_0: \rho = 0$$
$$H_A: \rho \neq 0$$

en donde ρ es el coeficiente de correlación a nivel poblacional. De nuevo, se utiliza la prueba *t*.

Una prueba *t* para el coeficiente de correlación poblacional	$t = \frac{r - \rho}{s_r}$	[11.29]

en donde s_r es el error estándar del coeficiente de correlación y puede hallarse así

El error estándar del coeficiente de correlación	$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$	[11.30]

Entonces,

$$s_r = \sqrt{\frac{1 - 0.93776}{15 - 2}} = 0.069$$

y

$$t = \frac{0.9683 - 0}{0.069} = 13.995$$

Si α es 5%,

La **regla de decisión** es: "No rechazar si *t* está entre ± 2.160. De otro modo rechazar".

Debido a que $t = 14.033 > 2.160$, se rechaza la hipótesis nula. A un nivel de significancia del 5%, se concluye que el coeficiente de correlación poblacional no es cero y que los pasajeros y la publicidad están relacionados. Al igual que con la prueba para β_1, si la hipótesis nula no se rechaza se concluye que la publicidad no tiene poder explicativo y el nuevo modelo tendrá que especificarse.

El hecho que el valor *t* de 14.033 sea casi el mismo tanto para β_1 como para ρ no es coincidencia. Siempre se obtendrán los resultados idénticos de estas dos pruebas de hipótesis en un modelo de regresión simple y, en realidad, posiblemente no se realizarán pruebas de hipótesis tanto para β_1 como para ρ. Sin embargo, deberían acostumbrarse a ambas pruebas puesto que esta igualdad no se mantiene en un modelo de regresión múltiple, como se verá en el siguiente capítulo.

Ejercicios de la sección

29. Utilizando la prueba de hipótesis apropiada, al nivel del 5%, ¿es GPA una variable explicativa significativa de las ofertas de trabajo en el ejercicio 9? Asegúrese de mostrar los cuatro pasos.

30. En el ejercicio 10, ¿la relación entre la tasa de interés y las ventas de vivienda es significativa? Pruebe la hipótesis a un nivel de significancia del 1%.

31. En el ejercicio 11, ¿la tasa de interés es significante al nivel del 10%? Pruebe la significancia del coeficiente de correlación al 10%. ¿Cómo difiere esta prueba de la prueba para β_1?

32. Pruebe la hipótesis del profesor Mundane para β_1 al nivel del 5% en el ejercicio 13. ¿Qué concluye? Compárela con su prueba para ρ.

33. Pruebe la significancia de los puntajes del examen en el ejercicio 14 al nivel del 5%. ¿Cuál es el valor ρ para esta prueba?

11.10 Intervalos de confianza en el análisis de regresión

El análisis de regresión puede pronosticar y predecir valores para la variable dependiente. Una vez que se ha determinado la ecuación de regresión, se puede desarrollar un estimado *puntual* para la variable dependiente sustituyendo un valor dado para X en la ecuación y despejando Y.

Además, el investigador puede estar interesado en los estimados *por intervalo*. Ya hemos visto que con frecuencia se prefieren más que los estimados puntuales. Existen por lo menos dos estimados por intervalo que se relacionan comúnmente con los procedimientos de regresión.

El primero es un estimado por intervalo para el valor promedio de Y dado cualquier valor X. Se puede estimar la media *poblacional* para *todos* los valores de Y (no sólo los $n = 15$ de la muestra) cuando X es igual a algún valor dado. Se puede desear el número *promedio* de pasajeros en *todos* los meses en los que se gastó US$10,000 en publicidad (por ejemplo, $X = 10$). Esto es lo que se denomina **media condicionada**.

Un segundo intervalo de confianza importante busca estimar un *valor único* de Y dado que X se fija en una cantidad específica. Este estimado se llama **intervalo de predicción** (I.P.). Por tanto, mientras que la media condicionada es una estimación del valor promedio de Y en todos los meses en los cuales X es igual a un monto especificado, los estimados por intervalo de predicción Y en cualquier mes, en el cual X se fija en un monto dado.

A. La media de *Y* condicionada a un valor de *X*

Se supone que se desea desarrollar un estimado por intervalo para la media condicionada de Y, $\mu_{y|x}$. Esta es la media poblacional para todos los valores de Y, con la condición de que X sea igual a un valor específico. Vale la pena recordar que si X es igual a algún monto dado (por ejemplo $X = 10$) muchas veces, se obtendrán muchos valores diferentes de Y. El intervalo que se está calculando aquí es un estimado de la media de todos los valores de Y. Es decir, es una estimación de intervalo para el valor promedio de Y, con la condición que X se deje igual 10 veces.

En realidad el intervalo de confianza para el valor promedio condicional de Y tiene dos posibles interpretaciones, de la misma manera que los intervalos de confianza que se realizaron en el capítulo 7. Se asume que se está calculando, por ejemplo, un intervalo de confianza del 95%.

Primera interpretación: Como se explicó anteriormente, si se deja X igual la misma cantidad de veces, se obtendrán muchos valores diferentes de Y. Entonces se puede estar 95% seguro de que la media de esos valores Y ($\mu_{y|x}$) caerá dentro del intervalo especificado.

Segunda interpretación: Si se tomaran muchas muestras de los valores de X y Y, y se construyera un intervalo de confianza con base en cada muestra, 95% de ellos contendría $Y\,\mu_{y|x}$, el valor promedio real pero desconocido de Y dado que $X = 10$.

Para calcular este intervalo para el valor promedio condicional de Y, se debe primero determinar S_y, el **error estándar de la media condicionada**. El error estándar de la media condicionada reconoce que se utiliza una muestra para calcular b_0 y b_1 en la ecuación de regresión. Por tanto, b_0 y b_1 están sujetos al error de muestreo. Si se fuera a tomar un conjunto diferente de $n = 15$ meses y determinar una ecuación de regresión, sería probable que se obtuvieran valores diferentes para b_0 y b_1. El propósito de S_y es tener en cuenta los diferentes valores de b_0 y b_1 que resultan del error de muestreo. Se determina mediante:

Error estándar de la media condicionada

$$S_Y = Se\sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SCx}} \qquad [11.31]$$

en donde Se = es el error estándar de estimación

X_i = es el valor dado para la variable independiente

El intervalo de confianza para la media condicionada es entonces:

Intervalo de confianza para la media condicionada

$$\text{I.C. para } \mu_{y|x} = \hat{Y}_i \pm tS_Y \qquad [11.32]$$

en la cual $\hat{Y}_i$ es el estimador puntual hallado de la ecuación de regresión. El valor t se basa en $n - 2$ grados de libertad porque se deben calcular los dos valores b_0 y b_1 de los datos muestrales.

Si Hop Scotch desea desarrollar el intervalo para la media condicionada en donde $X_i = 10$ y la media de todos los 15 valores de X es $\overline{X} = 12.47$, la fórmula (11.31) da:

$$S_y = 0.907\sqrt{\frac{1}{15} + \frac{(10 - 12.47)^2}{137.73333}} = 0.303$$

Si se utiliza un nivel de confianza del 95%, $t_{0.05,13} = \pm 2.160$. Ya que:

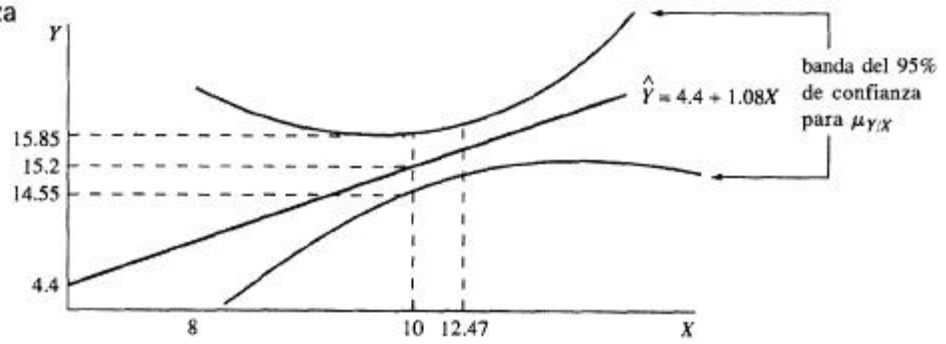
$$\hat{Y}_i = 4.4 + 1.08(10) = 15.2$$

La fórmula (11.32) halla

$$\text{I.C. para } \mu_{y|x} = 15.2 \pm 2.160(0.303)$$
$$14.55 \leq \mu_{y|x} \leq 15.85$$

Hop Scotch puede estar 95% seguro de que si se invierten US$10,000 en publicidad muchas veces (teóricamente, un número infinito de veces), existe un 95% de probabilidad de que la media de todos los valores resultantes para los pasajeros estará entre 14,550 y 15,850. Minitab también genera este intervalo, como se observa en la pantalla 11.4 (que se presentó anteriormente).

Figura 11.18
Límites de confianza para $\mu_{y|x}$



Se calcularía el intervalo de confianza para $\mu_{y|x}$ en valores diferentes de X. Estos intervalos conformarán toda una **banda de confianza** para $\mu_{y|x}$. Vale la pena notar que en la figura 11.18, la banda se vuelve más ancha en cada extremo porque el análisis de regresión se basa en las medias, y entre más se aleja de la media de $\overline{X} = 12.47$, los resultados se vuelven menos precisos. Por tanto, para mantener el nivel de confianza del 95%, la banda debe ser más amplia.

B. El intervalo de predicción para un valor único de *Y*

El intervalo de confianza elaborado anteriormente es para el valor de la media poblacional de todos los valores de Y cuando X es igual a una cantidad dada muchas veces. En otras ocasiones puede ser útil construir un intervalo de confianza para un valor único de Y que se obtiene cuando X se deja igual a un valor dado una sola vez. Hop Scotch puede estar interesado en predecir el número de clientes el próximo mes si invierte US$ 10,000 en publicidad. Esto difiere del problema anterior en el cual la preocupación se centraba en el valor promedio de Y si X se fija en 10 muchas veces.

Ahora el objeto es predecir un valor único de Y si X se fija en una cantidad dada una sola vez. En lugar de intentar predecir la media de muchos valores Y, obtenidos con la condición de que X sea igual a 10 muchas veces, se desea predecir un valor único para Y, obtenido si X se fija en 10 una sola vez. Deténgase a pensar en este problema durante un minuto. Los promedios tienden a estar centrados alrededor de la mitad del conjunto de datos. Por tanto, son más fáciles de predecir debido a que se conoce en dónde están. Sin embargo, los valores individuales están muy dispersos y por tanto son más difíciles de predecir. Por ende, un intervalo de confianza del 95% para un valor único de Y debe ser más ancho que el de una media condicionada.

Este intervalo de confianza para el intervalo de predicción de Y también tiene dos interpretaciones. Tales interpretaciones se proporcionan bajo el supuesto de que los intervalos que se calcularon son intervalos del 95%, aunque, por supuesto, pueden utilizarse otros niveles de confianza.

Primera interpretación: Si se determina que X es igual a alguna cantidad sólo una vez, se podría obtener un único valor resultante de Y. Se puede estar 95% seguro de que dicho valor único de Y cae dentro del intervalo especificado.

Segunda interpretación: Si se tomaran muchas muestras y cada una se utilizara para construir un intervalo de confianza de predicción, el 95% de ellos contendrían el valor verdadero para Y.

Para calcular este intervalo de predicción, primero se debe calcular el **error estándar del pronóstico**, S_w (no confundirse con el error estándar de la media condicionada S_y). Este error estándar del pronóstico explica el hecho de que los valores individuales estén más dispersos que las medias. El error estándar del pronóstico S_w refleja el error de muestreo inherente al error estándar de la media condicionada S_y más la dispersión adicional, porque se está tratando con un valor individual de Y. La fórmula (11.33) se utiliza en su cálculo:

Error estándar del pronóstico

$$S_{y_i} = Se\sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SCx}} \qquad [11.33]$$

El intervalo de predicción para un valor único de Y, Y_x entonces será:

Intervalo de confianza para el intervalo de predicción

$$\text{I.C. para } Y_x = \hat{Y}_i \pm tS_{Y_i} \qquad [11.34]$$

Ahora se construye un intervalo de confianza para un valor único de Y cuando $X = 10$ y se compara con el intervalo para la media condicionada elaborado anteriormente.

$$S_{yi} = Se\sqrt{1 + \frac{1}{15} + \frac{(10 - 12.47)^2}{137.73333}}$$
$$= 0.907\sqrt{1.1114}$$
$$= 0.956$$

Debido a que:

$$\hat{Y}_i = 4.4 + 1.08(10)$$
$$= 15.2$$

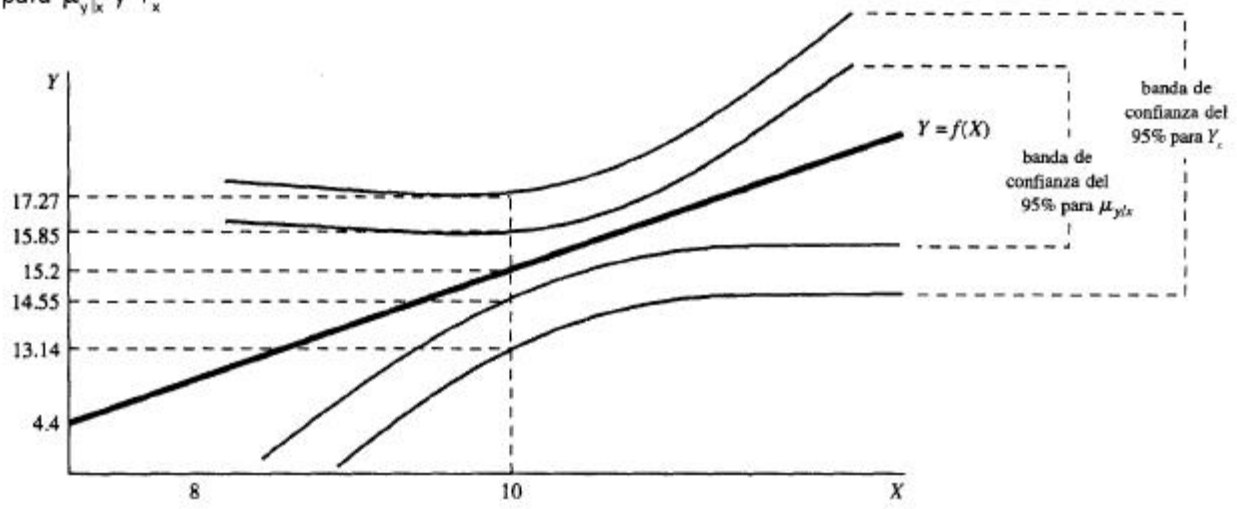
Se obtiene:

$$\text{I.C. para } Y_x = \hat{Y}_i \pm tS_{y_i}$$
$$= 15.2 \pm (2.160)(0.956)$$
$$= 15.2 \pm 2.065$$
$$13.14 < Y_x < 17.27$$

De nuevo la pantalla en Minitab 11.4 confirma este intervalo. Hop Scotch puede estar 95% seguro de que si en un mes cualquiera X_i = US$10,000, el valor único resultante de Y estará entre 13,140 y 17,270 pasajeros.

Como se prometió, este intervalo es más amplio que el primero porque se está trabajando con menos valores individuales predecibles. Se comparan en la figura 11.19.

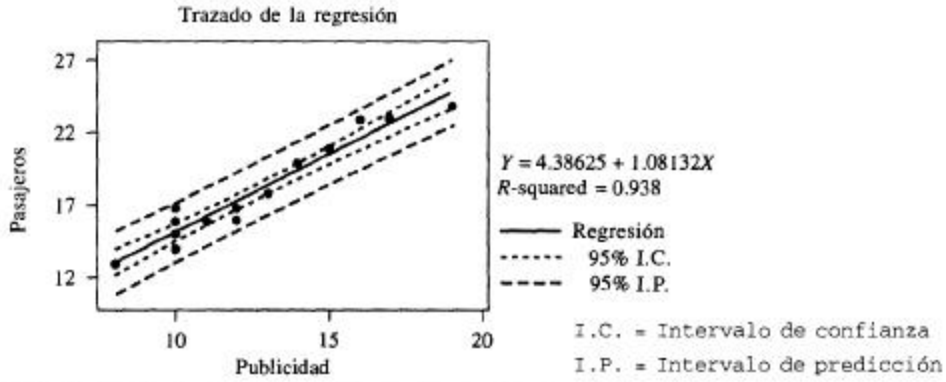
Figura 11.19
Estimados por intervalo para $\mu_{y|x}$ y Y_x



Estas bandas de confianza se observan en la pantalla de Minitab 11.5, aunque las curvaturas no son tan evidentes.

Pantalla en Minitab 11.5.

Confidence bands for $\mu_{y|x}$ and Y_x (Bandas de confianza para $\mu_{y|x}$ y Y_x)



C. Factores que influyen en el ancho del intervalo

Dado un nivel de confianza, es preferible minimizar la amplitud del intervalo. Entre más pequeño sea el intervalo, más precisa será la predicción de $\mu_{y|x}$ o de Y_x. Sin embargo, existen fuerzas que trabajan en contra del intento de producir un intervalo más estrecho.

La primera es el grado de dispersión de los datos originales. Entre más dispersos estén los datos originales, mayor será el *Se*, el error estándar de estimación. Dada la aritmética de las fórmulas (11.31) y (11.33), un *Se* mayor resulta en un intervalo más amplio.

El tamaño de la muestra es un segundo factor en la determinación de la amplitud del intervalo. Como se ha observado en los capítulos anteriores, un tamaño muestral grande termina en un error estándar más pequeño. De nuevo, dada la aritmética descrita anteriormente, un error estándar pequeño resulta en un intervalo pequeño.

Además, como ya se ha visto, un valor de X relativamente cercano a $\overline{X}$ producirá un intervalo pequeño porque la regresión se basa en los promedios. Por tanto, un tercer factor que influye en la amplitud del intervalo es la distancia a la cual está un valor particular de X de $\overline{X}$.

Ejercicios de la sección

34. ¿Qué miden el error estándar de la media condicionada y el error estándar del pronóstico?

35. ¿En qué difiere el intervalo de confianza para la media condicionada del intervalo de predicción?

36. ¿Qué miden las bandas de confianza para la media condicionada y el intervalo de predicción y por qué tienen la forma que tienen? ¿Qué afecta la amplitud de estos intervalos?

37. El centro de ubicación laboral en State University, del ejercicio 9, desea un estimado por intervalo del 95% para el número promedio de ofertas laborales que muchos de sus graduados recibirán para quienes obtuvieron un GPA de 2.69. Calcule e interprete el intervalo apropiado.

38. Fred tiene un GPA de 2.69 (ver ejercicios 9 y 37). Calcule el intervalo del 95% para el número de ofertas laborales que recibirá. ¿Por qué difiere de su respuesta del ejercicio 37?

39. Si el economista del Departamento de Recursos Humanos de Florida State, del ejercicio 10, identifica muchos consumidores con ingresos de US$14,200, ¿cuál es el intervalo del 99% para el consumo promedio de todos esos consumidores?

40. Si el economista del ejercicio 39 identifica un consumidor con un ingreso de US$14,500

a. ¿Cuál es la estimación puntual de su consumo?

b. ¿Cuál es el estimado por intervalo del 99% de su consumo?

11.11 Análisis de varianza en la regresión

El modelo de regresión presenta una descripción de la naturaleza de la relación entre las variables dependiente e independiente. Se utilizó una prueba t para probar la hipótesis que $\beta_1 = 0$. Una prueba similar puede realizarse con el uso del análisis de varianza (ANOVA) con base en la prueba F. El procedimiento ANOVA mide la cantidad de variación en el modelo de muestreo. Como se anotó anteriormente, existen tres fuentes de variación en un modelo de regresión: la variación explicada por la regresión (*SCR*), la variación que permanece sin explicar debido al error (*SCE*), y la variación total (*SCT*), la cual es la suma de las dos primeras. Esto puede resumirse en una tabla de ANOVA, cuya forma general se presenta en la tabla 11.3.

Tabla 11.3 Tabla general de ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón F
Regresión	*SCR*	k	$\text{CME} = \frac{\text{SCR}}{\text{K}}$	$\frac{\text{CMR}}{\text{CME}}$
Error	*SCE*	$n-k-1$	$\text{CME} = \frac{\text{SCE}}{n-k-1}$	
Total	*SCT*	$n-1$		

La razón *CMR/CME* proporciona una medida de exactitud del modelo porque es la razón de la desviación promedio al cuadrado que se explica con el modelo, y la desviación promedio al cuadrado se queda sin explicar. Entre mayor sea esta razón, el modelo tendrá mayor poder explicativo. Es decir, una prueba F alta señala que el modelo posee un poder explicativo significativo. Para determinar qué es alto, el valor F debe compararse con un valor crítico tomado de la tabla *G* del apéndice III.

La fórmula computacional para *SCE* se dio en la fórmula (11.16). *SCR* puede calcularse como:

Suma de cuadrados de la regresión

$$SCR = \frac{(SCxy)^2}{SCx} \qquad [11.35]$$

Utilizando los datos de Hop Scotch, la fórmula 11.16 da:

$$\begin{aligned} SCR &= SCy - \frac{(SCxy)^2}{SCx} \\ &= 171.73333 - \frac{(148.93333)^2}{137.73333} \\ &= 10.69 \end{aligned}$$

y la fórmula 11.35 produce:

$$SCR = \frac{(148.93333)^2}{137.73333}$$
$$= 161.0441$$

SCT se halla como la suma de *SCR* y *SCT*, como se muestra en la tabla 11.4. El valor *F* tiene 1 y 13 grados de libertad debido a que fue formado con el cuadrado medio de la regresión y el cuadrado medio del error, tal como se observa en la tabla 11.4. La pantalla en Minitab 11.4 también proporciona la tabla ANOVA.

Tabla 11.4
Tabla ANOVA para Hop Scotch Airlines

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón *F*
Regresión	161.04	1	161.04	196.39
Error	10.69	13	0.82	
Total	171.73	14		

Se puede fijar $\alpha = 0.05$ para contrastar la hipótesis de que $\beta_1 = 0$. Entonces $F_{0.05,1,13} = 4.67$ produce una regla de decisión que establece que se debería rechazar la hipótesis nula si el valor *F* excede de 4.67. Como $196.39 > 4.67$, se rechaza la hipótesis nula y se concluye con un 95% de confianza en que la publicidad tiene poder explicativo. Este es el mismo resultado obtenido en la prueba *t*.

En realidad, en regresión simple, la prueba *F* y la prueba *t* son análogas. Ambas darán los mismos resultados. El valor *F* es el cuadrado del valor *t*. En regresión múltiple, la prueba *F* produce una prueba más general para determinar si alguna de las variables independientes en el modelo tienen poder explicativo. Cada variable se prueba después por separado con la prueba *t* para determinar si es una de las variables significativas.

Problemas resueltos

1. **Función de consumo de Keynes** En el año de 1936, en su famoso libro, *The General Theory of Employment, Interest and Money* (Teoría general del empleo, interés y dinero), el notable economista británico John Maynard Keynes propuso una relación teórica entre el ingreso y los gastos en consumo personal. Keynes argumentó que a medida que el ingreso aumentaba, el consumo se incrementaría en una cantidad menor. Esta relación teórica se ha probado empíricamente muchas veces desde 1936.

   Milton Friedman, profesor de economía en la Universidad de Chicago y premio Nobel de economía, recolectó datos sobre el ingreso y el consumo en los Estados Unidos durante un período prolongado. Aquí se muestran las 10 observaciones sobre los niveles anuales de consumo e ingreso utilizados por Friedman en su estudio. Utilizando estos datos, se deriva una función de consumo bajo el supuesto de que existe una relación lineal entre el consumo y el ingreso. Las cifras se encuentran en miles de millones de dólares corrientes:

Año	Ingreso	Consumo
1950	284.8	191.0
1951	328.4	206.3
1952	345.5	216.7
1953	364.6	230.0
1954	364.8	236.5
1955	398.0	254.4
1956	419.2	266.7
1957	441.1	281.4
1958	447.3	290.1
1959	483.7	311.2

a. Debido a que el consumo depende del ingreso, el consumo es Y o variable dependiente. Friedman buscó una función de consumo de la forma

$$\hat{C} = b_0 + b_1 I$$

en donde C es el consumo e I es el ingreso.

$$\Sigma X = 3{,}877.4 \qquad \Sigma XY = 984{,}615.32 \qquad \Sigma Y^2 = 630{,}869.49$$

$$\Sigma Y = 2{,}484.3 \qquad \Sigma X^2 = 1{,}537{,}084.88$$

$$SCx = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$= 1{,}537{,}084.88 - \frac{(3{,}877.4)^2}{10}$$

$$= 33{,}661.804$$

$$SCy = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

$$= 630{,}869.49 - \frac{(2{,}484.3)^2}{10}$$

$$= 13{,}694.841$$

$$SCxy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$

$$= 984{,}615.32 - \frac{(3{,}877.4)(2{,}484.3)}{10}$$

$$= 21{,}352.838$$

$$b_1 = \frac{SCxy}{SCx}$$

$$= \frac{21{,}352.838}{33{,}661.804}$$

$$= 0.634$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$= 248.43 - (0.634)(387.74)$$

$$= 2.603$$

Por tanto:

$$\hat{C} = 2.603 + 0.63I$$

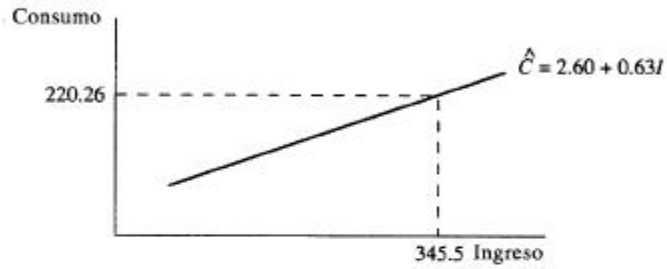
Estos no son los mismos valores que Friedman encontró debido a que se utilizó sólo una pequeña porción de su conjunto de datos. Sin embargo, el modelo confirma la teoría de Keynes. El coeficiente de 0.63 demuestra que por cada US$1 (o US$1,000,000,000) de incremento en el ingreso, el consumo se incrementará en 63 centavos de dólar (o US$630,000,000). Quienes han tomado un curso de introducción a la macroeconomía reconocerán que 0.63 es la propensión marginal al consumo. La constante, o término intercepto, de 2.603 es el nivel de consumo cuando el ingreso es cero. Los economistas con frecuencia argumentan que esta

interpretación del término intercepto no es válida debido a que un sistema económico siempre generará ingreso positivo. La función de consumo, por tanto, se grafica sin el intercepto, como en la figura. Si $I = 345.5$, como en 1952, el modelo predice

$$\hat{C} = 2.603 + 0.63(345.5) = 220.26$$

El consumo en realidad fue de 216.7 en 1952, resultando un error de US$3.560.000.000

$$\hat{C} = 2.603 + 0.63(345.5) = 220.26$$



b. El coeficiente de determinación es:

$$r^2 = \frac{(SCxy)^2}{(SCx)(SCy)}$$
$$= \frac{(21{,}352.838)^2}{(33{,}661.804)(13{,}694.841)}$$
$$= 0.989$$

Un cambio en el ingreso explica más del 98% de la variación en el consumo. La información respecto a los valores b_0, b_1 y r^2 son vitales para quienes aconsejan al Congreso y al presidente sobre asuntos de política económica nacional.

2. **Acciones de la Reserva Federal para frenar la inflación** Después de aproximadamente seis años de expansión continuada, la economía de los Estados Unidos comenzó a presentar signos de presiones inflacionarias en otoño de 1988. Un artículo de la edición de septiembre de *The Wall Street Journal* describió los esfuerzos de la junta de la Reserva Federal por calmar estos aires inflacionistas. Esto debía realizarse restringiendo el suministro de dinero a través del incremento en la tasa de descuentos que los bancos comerciales deben pagar por prestar de la Reserva Federal. En febrero de 1988, Manuel H. Johnson, vicepresidente de la Reserva Federal, dijo ante una audiencia de una conferencia de Cato Institute, que las acciones de la Reserva respecto a la tasa de descuentos podrían predecirse sobre la base de la tasa de los fondos federales, la cual es el costo que los bancos cobran entre ellos para los créditos de un día para otro. Sin embargo, durante lo que restó de 1988, los controladores de la Reserva argumentaron que la tasa de los fondos federales no estaba sirviendo como predictor adecuado de los cambios en la tasa de descuento, y que este comportamiento deficiente dificultaba que los inversionistas intentaran predecir qué nivel de la tasa de interés permitiría la Reserva Federal.

Aquí se presentan los valores para la tasa de los fondos federales y la tasa de descuento desde mediados de 1987 hasta mediados de 1988. ¿Estos datos sustentan los cargos de los controladores de la Reserva Federal?

Fecha	Tasa de Fondos Federales (%)	Tasa de descuento (%)	Fecha	Tasa de fondos Federales (%)	Tasa descuento (%)
Junio de 1987	8.0	7.5	Dic. de 1987	7.0	5.5
Jul. de 1987	7.5	7.5	Ene. de 1987	6.0	5.5
Ago. de 1987	7.0	7.0	Feb. de 1987	7.0	5.5
Sept. de 1987	6.5	6.5	Marzo de 1987	7.5	5.5
Oct. de 1987	6.0	6.0	Abril de 1987	7.0	6.0
Nov. de 1987	6.0	5.5	mayo de 1987	7.5	6.5
				83.0	74.5

Debido a que Johnson argumentó que la tasa de los fondos federales podría explicar la conducta de la tasa de descuentos, los fondos federales se ven como variable independiente.

a. La naturaleza de la relación entre la tasa de fondos federales y la tasa de descuento puede analizarse a través del análisis de regresión y de correlación.

$$\Sigma X = 83 \qquad \Sigma Y^2 = 469.25$$
$$\Sigma Y = 74.5$$
$$\Sigma XY = 518.5 \qquad \bar{Y} = 6.21$$
$$\Sigma X^2 = 579 \qquad n = 12$$
$$SCx = 4.9166667$$
$$SCy = 6.72917$$
$$SCxy = 3.20833$$
$$b_1 = 0.6525$$
$$b_0 = 1.6949$$

Por tanto,

$$\hat{Y} = 1.69 + 0.653\,X$$

El coeficiente de determinación es:

$$r^2 = \frac{(3.20833)^2}{(4.92)(6.73)}$$
$$= 0.3111$$
$$r = 0.56$$

Los controladores de la Reserva tienen razón en su crítica de la tasa de los fondos federales como predictor de los cambios en la tasa de descuento. Sólo el 31% de los cambios en la tasa de descuentos se explican mediante los cambios en la tasa de los fondos federales.

b. Una medida de bondad de ajuste que refleja la capacidad de la tasa de los fondos federales para predecir la tasa de descuento es el error estándar de estimación.

El error estándar de estimación es:

$$SCE = SCy - \frac{(SCxy)^2}{SCx}$$
$$= 6.7292 - \frac{(3.208)^2}{4.9166}$$
$$= 4.63033$$
$$CME = \frac{4.63033}{10}$$
$$= 0.463033$$
$$Se = \sqrt{0.463033}$$
$$= 0.6808$$

Típicamente, el estimado de la tasa de descuento está en error en 0.68 de un punto porcentual.

c. Una prueba de la significancia del coeficiente de correlación sería muy útil en este punto. Sea el nivel de confianza 95%. Con 10 grados de libertad el valor crítico de t es por tanto ± 2.228.

Las hipótesis son:

$$H_0: \rho = 0$$
$$H_A: \rho \neq 0$$

Regla de decisión: "No rechazar si H_0 si t está entre ± 2.228. De lo contrario rechazar".

$$t = \frac{r}{S_r}$$
$$= \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$
$$= \frac{0.56}{\sqrt{(1 - 0.31)/10}}$$
$$= \frac{0.56}{0.2627}$$
$$= 2.13$$

La hipótesis nula no puede rechazarse. A pesar del hallazgo muestral de una relación positiva entre las tasas de los fondos federales y la tasa de descuento, no se puede rechazar la hipótesis de que no hay correlación. El coeficiente de correlación muestral no es significativo al nivel del 5%.

d. Una prueba de la significancia del coeficiente de regresión de $b_1 = 0.6525424$ también es sabia. La prueba se realizará al nivel de 99%. Con 10 grados de libertad el valor crítico de t es ± 3.169.

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Regla de decisión: "No rechazar si t está entre ± 3.169. De lo contrario rechazar. La prueba requiere".

$$t = \frac{b_1}{S_{b_1}}$$

en donde:

$$S_{b_1} = \frac{Se}{\sqrt{SCx}}$$
$$= 0.681/\sqrt{4.92} = 0.307$$
$$= \frac{0.652542}{0.307}$$
$$= 2.126$$

La hipótesis de que $\beta_1 = 0$ no puede rechazarse. El valor para b_1 no es significativamente diferente de cero al nivel del 1%. Hay muy poca confianza o nada de confianza en la tasa de los fondos federales como predictor de la tasa de descuento. Sería imprudente de parte de los inversionistas confiar en los fondos federales como indicador del comportamiento de la tasa de descuento y de otras tasas de interés.

3. **Análisis adicional de la tasa de descuento** Con base en los resultados del problema anterior, el banquero profesional y los inversionistas pueden encontrar poco alivio en la habilidad de los fondos federales para predecir la tasa de descuento. Utilizando el modelo de regresión para elaborar un estimado puntual de la tasa de descuento no parece prudente. Para analizar aún más la relación entre estas dos variables, si existe alguna, se puede calcular los estimados por intervalo de la tasa de descuento.

   a. Las personas empleadas en la banca y las finanzas estarían interesadas en un estimado por intervalo para el valor promedio de la tasa de descuento, si los fondos federales se mantuvieran constantes durante varios meses. Claro que esto es un estimado por intervalo de la media condicionada de la tasa de descuento:

$$\text{I.C. para } \mu_{y|x} = \hat{Y} \pm tS_Y$$

y requiere el cálculo del error estándar de la media condicionada, S_y y $\hat{Y}$ como estimador puntual de la tasa de descuento. Debido a que la tasa de los fondos federales parecía moverse alrededor del 7% con frecuencia, es en esta tasa en la que se calculará el intervalo de confianza.

Para calcular S_y y $\hat{Y}$, se tiene que:

$$S_Y = Se\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{SCx}}$$
$$= 0.681\sqrt{\frac{1}{12} + \frac{(7 - 6.9167)^2}{4.92}}$$
$$= 0.1982$$

También,

$$\hat{Y} = b_0 + b_1X$$
$$= 1.6949 + 0.6525424(7)$$
$$= 6.2627$$

Si se calcula el intervalo al nivel del 95% de confianza, el valor crítico de t es $t_{0.05,n-2} = \pm 2.228$. Entonces se tiene que:

$$\text{I.C. para } \mu_{y|x} = \hat{Y} \pm tS_Y$$
$$= 6.2627 \pm (2.228)(0.1982)$$
$$5.82 < \mu_{y|x} < 6.70$$

Los banqueros pueden estar 95% seguros que si la tasa de los fondos federales es del 7% durante varios meses, la tasa de descuento promedio que deben pagar para prestar dinero de la Reserva Federal caerá entre el 5.82% y el 6.70%. Sus planes y políticas pueden formularse de acuerdo a esta expectativa.

b. Si un banquero desea hacer planes para el mes próximo, estaría interesado en saber cuál sería la tasa de descuento para ese mes, dado que la tasa de los fondos federales era del 7%. El banquero, por tanto calcularía un intervalo de predicción para el siguiente mes así:

$$\text{I.C. para } Y_x = \hat{Y} \pm tS_{y_x}$$

Esto requiere del cálculo del error estándar del pronóstico, S_{y_x}. Asumiendo un nivel de significancia del 95% y una tasa de los fondos federales del 7%, el banquero procedería así:

$$S_{y_x} = Se\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SC_x}}$$
$$= 0.70927$$

Debido a que $\hat{Y} = 6.2627$, se tiene que

$$\text{I.C. para } Y_x = 6.2627 \pm (2.228)(0.70927)$$
$$4.68 < Y_x < 7.85$$

El banquero podría formular planes para las operaciones del próximo mes comprendiendo que él podría estar 95% seguro de que si la tasa de los fondos federales fue del 7%, la tasa de descuento quedaría entre el 4.68 y el 7.85%. Este es un rango más amplio que el encontrado para la media condicionada de la tasa de descuento.

Ciertamente parecería que es cuestionable la afirmación de Johnson respecto al uso de la tasa de fondos federales para estimar o predecir la tasa de descuento. El r^2 es más bien bajo y las pruebas para la significancia de ρ y β_1 sugieren que las hipótesis de $\rho = 0$ y $\beta_1 = 0$ no pueden rechazarse en ninguno de los niveles aceptables de significancia.

Para ser justo, debe argumentarse que la tasa de los fondos federales debería retrasarse un mes. Es decir, la tasa de descuento en cualquier mes (período *t)* es una función de la tasa de fondos federales para el mes anterior (período *t - 1*). Esto permitiría a la Reserva Federal un tiempo para ajustar la tasa de descuento a la tasa del último mes de los fondos federales, ya que la Reserva Federal no puede responder inmediatamente a los cambios en la tasa de los fondos federales. Esto se expresa así:

$$TD_t = f(FF_{t-1})$$

en donde *TD* es la tasa de descuento y *FF* es la tasa de los fondos federales. Este modelo retrasado da:

$$\hat{Y} = 0.6 + 0.8X$$

Con $r^2 = 60\%$ y $Se = 0.47$. Esto representa una mejora sobre el modelo natural, el cual no incluye la variable retrasada.

4. **El efecto de productividad sobre el PIB real** Una emisión reciente de la revista *Fortune* informó la relación entre la productividad de los trabajadores y las tasas de cambio en el nivel de producción de la nación, medida en términos reales. El mensaje era que el incremento de productividad durante los años 80 podría servir como factor explicativo para el crecimiento del PIB. Con el crecimiento en la productividad y los cambios en el PIB medidos en porcentajes, y el PIB como variable dependiente, los datos anuales para ese período puede resumirse así:

$$\Sigma X = 32.5 \qquad \Sigma Y^2 = 483.72$$
$$\Sigma Y = 62.2 \qquad n = 9$$
$$\Sigma XY = 255.4 \qquad \Sigma X^2 = 135.25$$

El modelo es:

$$\hat{Y} = 0.69596273 + 1.721118X$$

el que indica que si la productividad incrementó un punto porcentual, el PIB real incrementará en 1.72%. r^2 es 0.98407, y $Se = 0.35$.

Para efectos de formular la política nacional tributaria, la cual argumentan algunos economistas al margen de la oferta, tiene un impacto directo en la productividad de los trabajadores. Los miembros de planeación de Washington probaron la significancia tanto del coeficiente de correlación muestral como del coeficiente de regresión muestral. Cada uno probó ser significativo al nivel del 10%.

Ellos mismos solicitaron luego un intervalo de confianza por cada coeficiente poblacional a un nivel del 10%:

$$\text{I.C. para } \beta_1 = b_1 \pm tS_{b_1}$$
$$S_{b_1} = \frac{Se}{\sqrt{SC_x}} = 0.08275$$
$$\text{I.C. para } \beta_1 = 1.72 \pm (1.895)(0.08275)$$
$$1.56 < \beta_1 < 1.88$$

Los miembros de planeación pueden basar la formulación de la política tributaria nacional en la condición de que pueden estar 90% seguros de que el coeficiente de regresión poblacional está entre 1.56 y 1.88.

Lista de fórmulas

[11.3]	$Y = b_0 + b_1X$	Fórmula para una recta que tiene intercepto, b_0, y pendiente, b_1.
[11.9]	$SCx = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$	Suma de cuadrados para X
[11.10]	$SCy = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$	Suma de cuadrados para Y
[11.11]	$SCxy = \Sigma XY - \frac{(\Sigma Y)(\Sigma Y)}{n}$	Suma de productos cruzados
[11.12]	$b_1 = \frac{SCxy}{SCx}$	La pendiente de la recta de regresión mide el cambio unitario en Y dado un cambio unitario en X.
[11.13]	$b_0 = \bar{Y} - b_1\bar{X}$	El intercepto es el valor de Y cuando X es igual a cero.
[11.14]	$d = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2}$	El estadístico de Durbin-Watson

[11.15]	$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}}$	El error estándar de estimación es la medida de la dispersión de los valores de Y alrededor de su media.
[11.16]	$SCE = SCy - \frac{(SCxy)^2}{SCx}$	Suma de cuadrados del error
[11.17]	$CME = \frac{SCE}{n-2}$	Cuadrado medio del error
[11.18]	$Se = \sqrt{CME}$	El error estándar de estimación es la medida de la dispersión de los valores de Y alrededor de su media.
[11.19]	$SCT = \Sigma(Y_i - \bar{Y})^2$	Suma de cuadrados total
[11.20]	$SCR = \Sigma(\hat{Y}_i - \bar{Y})^2$	Suma de cuadrados de la regresión
[11.21]	$SCE = \Sigma(Y_i - \hat{Y}_i)^2$	Suma de cuadrados del error
[11.22]	$r = \sqrt{\frac{SCR}{SCT}}$	Coeficiente de correlación
[11.23]	$r = \frac{SCxy}{\sqrt{(SCx)(SCy)}}$	Forma de calcular el coeficiente de correlación
[11.24]	$r^2 = \frac{SCR}{SCT}$	Coeficiente de determinación
[11.25]	$r^2 = \frac{(SCxy)^2}{(SCx)(SCy)}$	Forma de calcular el coeficiente de determinación
[11.26]	$t = \frac{b_1 - \beta_1}{S_{b_1}}$	Prueba t para el coeficiente de regresión poblacional
[11.27]	$S_{b_1} = \frac{Se}{\sqrt{SCx}}$	Error estándar del coeficiente de regresión
[11.28]	I.C. para $\beta_1 = b_1 \pm t(s_{b_1})$	Intervalo de confianza para el coeficiente de regresión poblacional
[11.29]	$t = \frac{r - \rho}{s_r}$	Prueba t para el coeficiente de correlación poblacional
[11.30]	$S_r = \sqrt{\frac{1 - r^2}{n-2}}$	Error estándar del coeficiente de correlación
[11.31]	$S_Y = Se\sqrt{\frac{1}{n} + \frac{(X_i + \bar{X})^2}{SCx}}$	Error estándar de la media condicionada
[11.32]	I.C. para $\mu_{y\|x} = \hat{Y}_i \pm ts_Y$	Intervalo de confianza para la media condicionada
[11.33]	$s_{Y_i} = Se\sqrt{1 = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SCx}}$	Error estándar de pronóstico
[11.34]	I.C. para $Y_X = \hat{Y}_i \pm ts_{Y_i}$	Intervalo de confianza para el intervalo de predicción
[11.35]	$SCR = \frac{(SCxy)^2}{SCx}$	Suma de cuadrados de regresión

Ejercicios del capítulo

41. Los residentes de un pueblo pequeño están preocupados sobre el incremento en los costos de la vivienda en la zona. El alcalde considera que los precios de la vivienda fluctúan con los valores de la tierra. Los datos sobre 10 casas vendidas recientemente y el costo del terreno sobre el cual se construyeron se observan en la siguiente tabla en miles de dólares. Se trata el costo de las casas como la variable dependiente. Haga e interprete el modelo de regresión. Sobre esta base ¿parece que el alcalde está en lo cierto?

Valores de la tierra	Costo de la casa	Valores de la tierra	Costo de la casa
7.0	67.0	3.8	36.0
6.9	63.0	8.9	76.0
5.5	60.0	9.6	87.0
3.7	54.0	9.9	89.0
5.9	58.0	10.0	92.0

42. Calcule e interprete el coeficiente de determinación para el ejercicio 41.

43. Pruebe la hipótesis de que los valores de la tierra son significantes al nivel del 10% en el ejercicio 41.

44. Calcule e interprete el intervalo de confianza del 90% para el coeficiente de regresión en el ejercicio 41.

45. El gobierno estudiantil en la universidad local intenta determinar si el precio de admisión al salón de juegos del centro estudiantil tiene un impacto en el número de estudiantes que utilizan las instalaciones. El costo de admisión y el número de estudiantes que ingresan al salón se registran durante 12 viernes seguidos y se muestran en la siguiente tabla. Haga e interprete el modelo de regresión.

Precio	Número de boletas	Precio	Número de boletas
US$1.25	95	US$1.00	98
1.50	83	1.50	85
1.75	75	2.00	75
2.00	72	2.50	65
2.10	69	1.10	98
1.00	101	1.50	86

46. Calcule e interprete el intervalo de confianza del 99% para el coeficiente de regresión en el ejercicio 45.

47. Para reducir los crímenes, el presidente ha presupuestado más dinero para poner más policía en las calles de nuestra ciudad. ¿Qué información ofrece el modelo de regresión con base en estos datos sobre el número de policías en patrullas y el número diario de crímenes reportados? Utilice las fórmulas para ilustrar que el modelo *MCO* realmente se basa en las desviaciones con respecto a la media calculando,

$$SCx = \Sigma(X - \bar{X})^2 \qquad SCy = \Sigma(Y - \bar{Y})^2$$

$$SCxy = \Sigma(X - \bar{X})(Y - \bar{Y})$$

Policía	Número de crímenes reportados
13	8
15	9
23	12
25	18
15	8
10	6
9	5
20	10

48. Tía Bea desea obtener este verano más rendimiento de sus plantas de tomate Big Boy incrementando el número de veces que utiliza fertilizante. Con base en los datos mostrados aquí, ¿el coeficiente para el modelo de regresión sugiere que esto es posible? Utilice las fórmulas para ilustrar que el modelo *MCO* se basa en las desviaciones con respecto a la media calculando

$$SCx = \Sigma(X - \bar{X})^2 \qquad SCy = \Sigma(Y - \bar{Y})^2$$
$$SCxy = \Sigma(X - \bar{X})(Y - \bar{Y})$$

Uso de fertilizante	Rendimiento (libras)
4.00	12.00
9.00	20.00
5.00	15.00
8.00	17.00
2.00	7.00

49. Doce distritos escolares en el área de Chicago están interesados en saber si el incremento en las tasas de impuesto predial podrían relacionarse con el número de alumnos en clase de las escuelas locales. ¿Parece ser este el caso con base en los datos que se muestran a continuación?

Tasas de valorización	Estudiantes por clase	Tasas de valorización	Estudiantes por clase
1.20	32	1.30	25
1.20	36	1.30	21
1.10	25	1.20	35
1.30	20	1.40	16
1.10	39	1.40	39
1.20	42	1.30	37

a. Si se piensa que más alumnos requieren impuestos más elevados, ¿cuál es la variable dependiente? Calcule e interprete el modelo de regresión. ¿Las clases más grandes parecen estar relacionadas con los impuestos más altos?

b. Calcule e interprete el coeficiente de determinación y el coeficiente de correlación. ¿Parece ser útil este modelo?

c. Calcule e interprete el error estándar de estimación.

50. Pruebe la significancia tanto para el coeficiente de correlación como para el coeficiente de regresión a un nivel del 10% para el ejercicio 49. ¿Qué le dicen los resultados?

51. Calcule e interprete el intervalo de confianza del 95% para β_1 en el ejercicio 49.

52. Con base en las cifras presentadas por el servicio de renta interna (SRI), un grupo nacional de ciudadanos ha expresado su preocupación porque el presupuesto para éste no sea utilizado efectivamente. El SRI argumentó que el incremento en el número de contribuyentes que presentan su declaración de renta explica los problemas de presupuesto. A continuación se proporcionan los datos relevantes:

Año	Declaración de renta (en millones)	Presupuesto del SRI (en miles de millones de dolares)
1	116	US$6.7
2	116	6.2
3	118	5.4
4	118	5.9
5	120	3.7
6	117	5.9
7	118	4.7
8	121	4.2

a. Construya el modelo de regresión, ¿parece plausible el argumento del SRI?

b. Calcule e interprete el coeficiente de determinación.

c. Calcule e interprete el error estándar de estimación.

53. Cuál es el intervalo de confianza del 95% para el intervalo de predicción del ejercicio 52 si hay 119 declaraciones registradas?

54. Una teoría financiera popular sostiene que existe una relación directa entre el riesgo de una inversión y el rendimiento que promete. El riesgo de una acción se mide por medio de su valor β. A continuación se presentan los rendimientos y los valores β para 12 acciones ficticias sugeridas por la empresa de inversiones de Guess & Pickum. ¿Estos datos parecen confirmar esta teoría financiera de una relación directa?

Acción	Rendimiento (%)	Valor β
1	5.4	1.5
2	8.9	1.9
3	2.3	1.0
4	1.5	0.5
5	3.7	1.5
6	8.2	1.8

Acción	Rendimiento (%)	Valor β
7	5.3	1.3
8	0.5	-0.5
9	1.3	0.5
10	5.9	1.8
11	6.8	1.9
12	7.2	1.9

Típicamente los inversionistas consideran el rendimiento como una función del riesgo. Utilice una interpretación tanto del coeficiente de regresión como del coeficiente de correlación en su respuesta.

55. El servicio de emergencia para ciertas áreas rurales de Ohio con frecuencia es un problema, especialmente durante los meses de invierno. El jefe del Departamento de Bomberos de Danville, Township está preocupado por el tiempo de respuesta a las llamadas de emergencia. Ordena una investigación para determinar si la distancia del lugar de la llamada, medida en millas, puede explicar el tiempo de respuesta, medido en minutos. Con base en 37 emergencias, se recolectaron los siguientes datos:

$$\Sigma X = 234 \quad \Sigma X^2 = 1{,}796$$
$$\Sigma Y = 831 \quad \Sigma Y^2 = 20{,}037$$
$$\Sigma XY = 5{,}890$$

a. ¿Cuál es el tiempo de respuesta a una llamada que proviene de ocho millas de la estación de bomberos?

b. ¿Qué tan dependiente es dicha estimación, con base en el grado de dispersión de los puntos de datos alrededor de la recta de regresión?

56. Refiriéndonos al ejercicio 55, a un nivel de confianza del 90%, qué puede decir sobre la significancia del:

a. ¿Coeficiente de regresión?

b. ¿Coeficiente de correlación?

57. Respecto al ejercicio 55, con un nivel de confianza del 90%, ¿qué intervalo de tiempo predeciría usted para una llamada proveniente Zeke Zipple, quien vive a 10 millas de la estación de bomberos?

58. Respecto al ejercicio 55, con un nivel de confianza del 90%, ¿cuál es el intervalo de tiempo promedio que usted predeciría para muchas llamadas provenientes de una distancia de 10 millas de la estación?

59. Utilizando los datos del ejercicio 55, el jefe de bomberos está interesado en una estimación del 95% para el coeficiente de regresión poblacional. Interprete sus resultados para el jefe.

Ejercicios en computador

Usted acaba de ser contratado por su nuevo suegro, presidente de Jesse James National Bank. Su primera tarea es estimar un modelo de regresión que predecirá los depósitos. Su suegro ha sugerido varias variables explicativas, incluyendo las tasas de interés, un índice para el clima económico general en la zona, y el número de negocios recientemente constituidos. Usted también considera que los niveles poblacionales podrían utilizarse como un predictor.

Usted ha recolectado datos para todas estas variables y debe decidir ahora cuál puede utilizarse mejor como variable explicativa, pues usted desea especificar un modelo de regresión simple (en realidad todas las variables pueden ser significativas en un modelo de regresión múltiple al igual que las analizadas en el siguiente capítulo). Ingrese al archivo "BANK" en su disco de datos. Éste contiene los datos de los depósitos (DP) en millones de dólares, la tasa de interés (INT) que el banco paga por los depósitos, un índice de la actividad económica (IAC), la población (POP) en cientos para diferentes áreas en las cuales se encuentran ubicadas sucursales del banco, y el número de nuevos negocios (NEG) en tales áreas.

Compare el poder explicativo de cada variable en modelos de regresión simple independientes. Proporcione un análisis comparativo de cada modelo con respecto a todas las características de regresión y correlación que usted considere que pueden proporcionar alguna información útil. ¿Cuál modelo recomendaría usted? Prepare su informe estadístico final como se describe en el apéndice I.

PUESTA EN ESCENA

La guerra de las "Colas" entre Coca Cola y Pepsi, que se mencionó en la sección denominada *Escenario* a comienzos de este capítulo, discutía los esfuerzos realizados por las dos compañías por ganar participación en el mercado a expensas una de la otra. Cada una ha intentado varias estrategias para lograr estas metas. Ninguna compañía parece interesarse en participar en una guerra de precios prolongada para incrementar las ventas. Pepsi ha confiado plenamente en nombres de celebridades, utilizando personalidades muy conocidas para promover su producto, mientras que Coca-Cola parece preferir los esquemas promocionales relacionados con películas populares y héroes de tiras cómicas.

Como analista de Coca-Cola, su trabajo es utilizar los datos proporcionados aquí para saber si los cambios en los precios son efectivos para promover las ventas. Estos datos se tomaron de los mercados de prueba seleccionados en toda la nación para paquetes de 12 botellas de cada bebida. Las ventas se aproximaron a la unidad más próxima en miles.

Utilice todas las herramientas de regresión y correlación que aprendió en este capítulo para analizar completamente el impacto de los cambios en precios en las ventas para ambos productos. Desarrolle e interprete el modelo de regresión para ambas compañías haciendo regresión de las ventas sobre el precio. Construya e interprete todos los intervalos de confianza y prueba todas las hipótesis pertinentes. ¿Parece que una compañía ha sido más exitosa en el uso de los cambios en precios para promover las ventas? Como empleado de Coca-Cola, ¿qué recomendaría usted?

Ventas de Pepsi	Precio de Pepsi	Ventas de Coca-Cola	Precio de Coca-Cola
25.00	US$2.58	35.00	US$2.10
21.00	3.10	25.00	3.52
18.00	3.25	21.00	2.10
35.00	2.10	19.00	2.55
29.00	2.90	23.00	3.50
24.00	2.85	31.00	2.00
18.00	4.21	24.00	3.50
16.00	5.26	31.00	2.99
18.00	5.85	20.00	2.99
32.00	2.50	19.00	2.25

Del escenario a la vida real

¿De qué lado está en la guerra de Coca-Cola y Pepsi? ¿Cuáles son sus razones? ¿Sabe usted cuál producto está ganando la guerra y por cuánto? Visite los sitios Web de estas compañías para ayudar a responder estas preguntas. Primero visite el sitio de Coca-Cola (*www.cocacola.com/sports*). Explore y haga un listado de las zonas características en Coke Home Page. Algunas áreas son About the Coca-Cola Company (Sobre la empresa Coca-Cola), una tienda de regalos, y la zona de juegos. En la sección About the Coca-Cola Company haga clic en cada uno de los botones de la máquina expendedora para aprender diferentes hechos sobre la empresa de interés para los inversionistas.

¿En alguna información de estas se menciona a Pepsi? Sobre todo, ¿el sitio está dirigido más hacia el consumidor o hacia el inversionista? ¿Cuáles estrategias competitivas de mercadeo, diferentes al precio, se sugieren en este sitio, por ejemplo cobrables y una tienda de regalos?

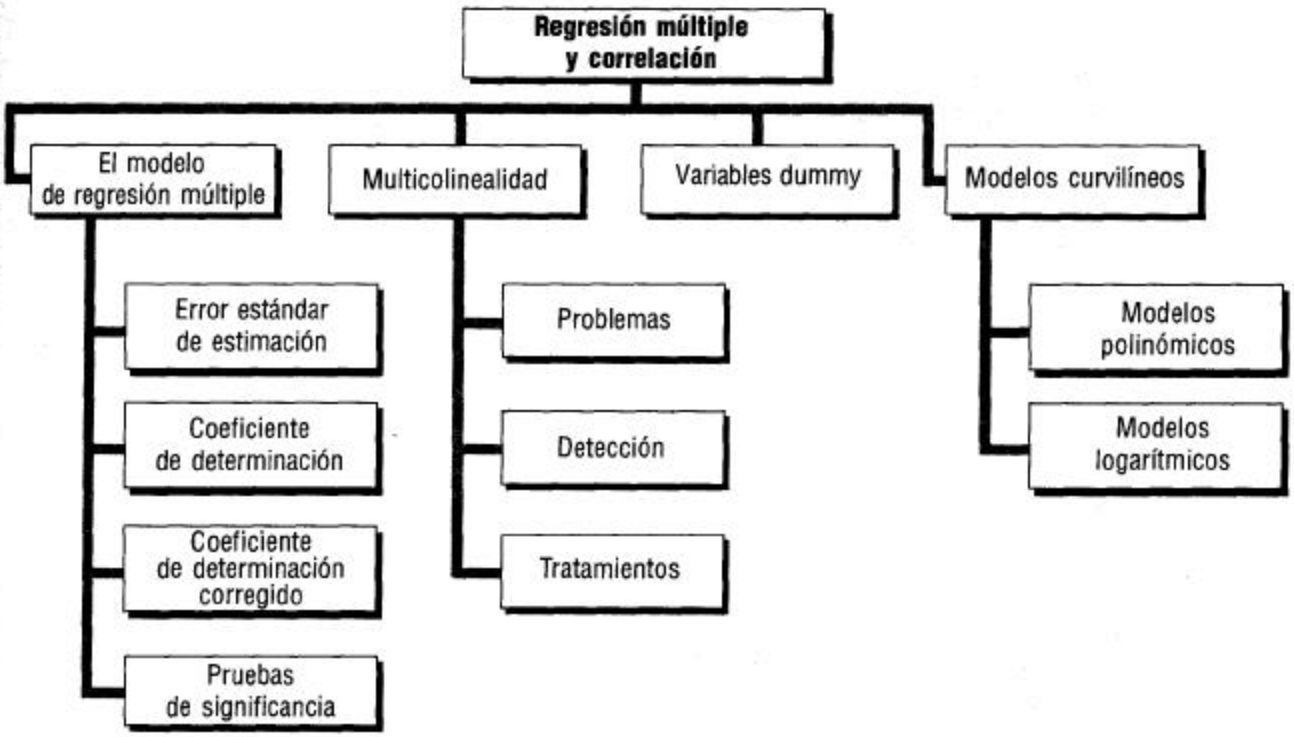
Ahora visite el sitio de Pepsi (*www.pepsico.com*). Haga una lista de las zonas de características de este Home Page. Sobre todo, ¿el sitio está dirigido a los consumidores o a los inversionistas? ¿Qué estrategias competitivas de mercadeo diferentes al precio se sugieren en este sitio? Haga clic en el área de Annual Report (reporte anual). Luego seleccione "PepsiCo Facts" seguido por "beverages". Lea toda la información fáctica. ¿Hay gráficas que muestren la participación en el mercado mundial y de Estados Unidos? Si es así, ¿Cuál es la participación de mercado para Coca-Cola y para Pepsi? ¿Qué motivos se sugieren para los vacíos en la participación de mercado a través de sus visitas a los sitios Web de estas dos compañías?

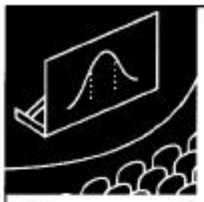
12

Regresión múltiple y correlación

Plan del capítulo

En un modelo de regresión, al utilizar más de una variable explicativa es posible incrementar el poder explicativo y la utilidad del modelo en la toma de muchas decisiones de negocios. Este capítulo discute la construcción de dichos modelos de regresión múltiple y muestra cómo pueden utilizarse para facilitar la toma de decisiones en los negocios.





ESCENARIO

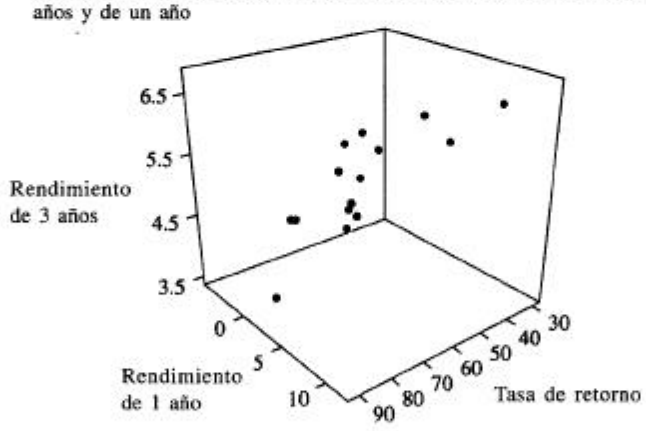
Como preparación para su grado, a finales de este año, usted ha asumido el cargo de practicante con Griffen Associates, una empresa de inversiones en Chicago. Como medida de sus habilidades financieras, la compañía le ha encomendado la tarea de analizar el desempeño del mercado de los fondos mutuos que operan como competencia de Griffen. Se han recolectado datos para el rendimiento de tres años (R3A) y para el rendimiento de un año (R1A) para 15 fondos competitivos.

El Sr. Griffen le solicita un informe sobre el desempeño de estos competidores, con base en diversos factores, incluyendo sus tasas de rendimiento, sus activos totales, y si cada fondo tiene una partida de carga por ventas.

Existe un interés en particular en Griffen Associates de saber si ha habido algún cambio durante los últimos tres años en el desempeño de estos fondos. Griffen está considerando cambios significativos en muchos de sus procedimientos operativos y ciertos gerentes que han estado con la firma durante varios años están preocupados sobre los resultados de tales cambios. Al analizar el comportamiento de las empresas competitivas a través del tiempo, estos gerentes consideran que ganan algún discernimiento para la futura dirección de Griffen Associates.

Este proyecto requerirá que usted configure y analice un modelo de regresión múltiple que pueda proporcionar la información necesaria para discernir y establecer los procedimientos operativos importantes que Griffen Associates está considerando.

Uso de las tasas de rendimiento para analizar los rendimientos de tres años y de un año



12.1 Introducción

En el capítulo 11 se analizó cómo una sola variable explicativa podría utilizarse para predecir el valor de la variable dependiente. Se considera cuánto más poderoso podría volverse el modelo si se utilizaran más variables explicativas. Esto es precisamente lo que el modelo de regresión múltiple hace, permitiendo incorporar dos o más variables independientes. El modelo de regresión múltiple con k variables independientes se expresa como:

El modelo de regresión múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad [12.1]$$

en donde β_i son los coeficientes de regresión y ε es el término de error aleatorio. Se estima el modelo utilizando los datos muestrales así:

El modelo de regresión múltiple estimado

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad [12.2]$$

en donde $\hat{Y}$ es el valor estimado para la variable dependiente y b_i son los estimados para los coeficientes poblacionales β_i. Los b_i se denominan **coeficientes parciales (o netos) de regresión** y tienen la misma interpretación que en la regresión simple. Por tanto, b_1, es la cantidad por la cual Y_i cambiará si X_i cambia en una unidad, *asumiendo que todas las otras variables independientes se mantienen constantes*. Esta suposición no fue necesaria bajo la regresión simple porque no había otras variables independientes para mantener constantes.

La regresión múltiple implica los mismos supuestos citados en el capítulo anterior para la regresión simple, más otros dos. El primer supuesto requiere que el número de observaciones n, exceda el número de variables independientes k, en por lo menos 2. En la regresión múltiple hay $k + 1$ parámetros por estimar: los coeficientes para las variables independientes k más el término del intercepto. Por tanto, los grados de libertad relacionados con el modelo son g.l. = $n - (k + 1)$. Si se va a retener incluso un grado de libertad, n debe exceder a k en por lo menos 2, de manera que $n - (k + 1)$ es por lo menos 1.

El segundo supuesto involucra la relación entre las variables independientes. Requiere que ninguna de las variables independientes esté linealmente relacionada. Por ejemplo, si $X_1 = X_2 + X_3$, o quizá $X_1 = 0.5X_2$, entonces una relación lineal existiría entre dos o más variables independientes y surgiría un problema grave. Este problema es la **multicolinealidad**.

Multicolinealidad La multicolinealidad existe si dos o más variables independientes están relacionadas linealmente.

La multicolinealidad puede hacer que los signos algebraicos de los coeficientes sean opuestos a lo que la lógica puede dictar, mientras que incrementan vastamente el error estándar de los coeficientes. Un análisis más completo sobre multicolinealidad se hará más tarde en este capítulo.

12.2 El modelo de regresión múltiple para Hop Scotch Airlines

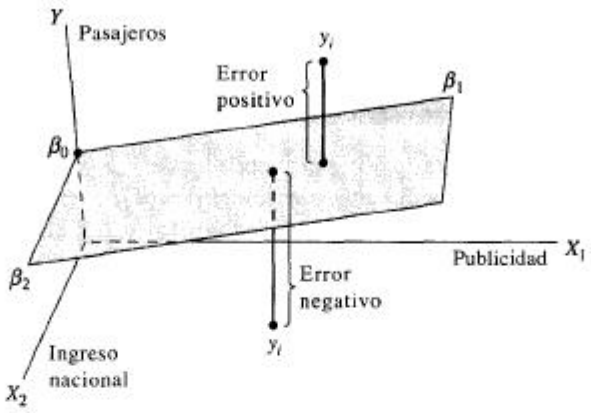
En el capítulo 11, Hop Scotch Airlines utilizó publicidad, en un modelo de regresión simple, para explicar y predecir el número de pasajeros. El modelo tenía un error estándar de 0.907 y r^2 del 94%. Se supone que Hop Scotch desea incorporar una segunda variable explicativa dentro de su modelo para explicar el número de pasajeros. Con base en el principio de que el ingreso es la determinante primaria de la demanda, Hop Scotch escoge el ingreso nacional como segunda variable. El modelo se vuelve entonces

El modelo regresión múltiple para Hop Scotch

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 \qquad [12.3]$$

Con dos variables explicativas, un diagrama de dispersión puede elaborarse en un espacio tridimensional formando un plano de regresión, como se observa en la figura 12.1. La variable dependiente se coloca en el eje vertical único. Un modelo con tres o más variables independientes requieren un *hiperplano* y es difícil de representar gráficamente.

Figura 12.1
Plano de regresión para Hop Scotch Airline



en donde $\hat{Y}$ es el valor proyectado para los pasajeros expresado en miles
X_1 es la publicidad expresada en cientos de dólares
X_2 es el ingreso nacional expresado en dólares.

Plano de regresión Los coeficientes de las dos variables independientes se representan mediante las pendientes del plano de regresión.

Los valores para b_0, b_1 y b_2 en la fórmula (12.3) se hallan como los de la regresión múltiple. Se desean estimaciones de los coeficientes que minimizan la suma de los errores al cuadrado: $\sum (Y_i - \hat{Y}_i)^2$ Esto proporcionará el modelo de mínimos cuadrados que se ajuste mejor a los datos que se muestran en la tabla 12.1. La última columna representa la nueva variable del *ingreso nacional* en billones de dólares.

Tabla 12.1
Datos de la regresión múltiple para Hop Scotch Airlines

Observación (meses)	Pasajeros (Y) (en miles)	Publicidad (X_1) (en miles de dólares)	Ingreso nacional X_2 (en billones de dólares)
1	15	10	2.40
2	17	12	2.72
3	13	8	2.08
4	23	17	3.68
5	16	10	2.56
6	21	15	3.36
7	14	10	2.24
8	20	14	3.20
9	24	19	3.84
10	17	10	2.72
11	16	11	2.07
12	18	13	2.33
13	23	16	2.98
14	15	10	1.94
15	16	12	2.17

Calcular un modelo de regresión múltiple manualmente es muy tedioso y demanda mucho tiempo. El procedimiento requiere $k + 1$ ecuaciones simultáneas con $k + 1$ desconocidas, en donde k es el número de variables al lado derecho. Por tanto, se dispensará todo esfuerzo para solucionar manualmente el modelo de regresión y se confiará solamente en el computador para la mayoría de los cálculos. La atención se centrará en los fundamentos necesarios para comprender e interpretar el modelo de regresión múltiple.

La pantalla 12.1 es una impresión en Minitab para los datos que aparecen en la tabla 12.1. Se puede observar claramente que si se aproximan los coeficientes para facilitar en algo la discusión, el modelo es

$$Pass = 3.53 + 0.84Adv + 1.44NI$$

en donde *Pass*, *Adv* y *NI* son pasajeros en miles, gastos publicitarios en miles de dólares e ingreso nacional en billones de dólares, respectivamente. Por tanto, el modelo predice que si se incrementa la publicidad una unidad (US$1,000), los pasajeros aumentarán 0.84 unidades (840 pasajeros) si el ingreso nacional no cambia. Además, si el ingreso nacional incrementa en una unidad (US$1 billón), los pasajeros incrementarán en 1,440, si la publicidad se mantiene constante.

Pantalla 12.1

Regression Analysis (Análisis de regresión)

The regression equation is (La ecuación de regresión es)

PASS = 3.53 + 0.840 ADV + 1.44 NI

Predictor	Coef	Stdev	t-ratio	p
Constant	**3.5284**	0.9994	3.53	0.004
ADV	**0.8397**	0.1419	5.92	0.000
NI	**1.4410**	0.7360	1.96	0.074

s = 0.8217 R-sq = 95.3% R-sq (adj) = 94.5%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	**163.632**	81.816	121.18	0.000
Error	12	8.102	0.675		
Total	14	**171.733**			

SOURCE	DF	SEQ SS
ADV	1	161.044
NI	1	2.588

PASS = Pasajeros
ADV = Publicidad
Stdev = Desviación estándar
t-ratio = razón t
R-sq = r^2
DF = Grados de libertad
SS = Suma de cuadrados
MS = Cuadrado medio
NI = Ingreso nacional

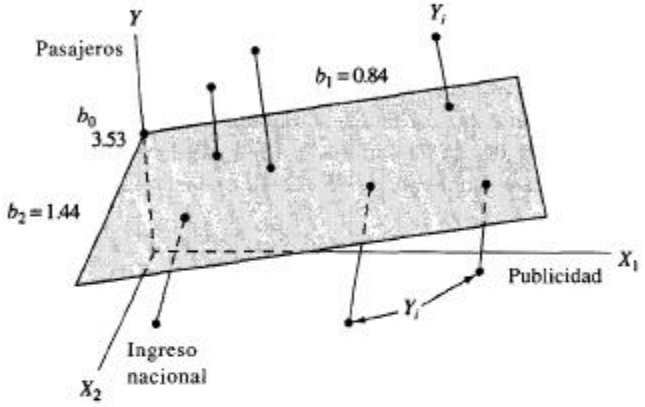
12.3 Evaluación del modelo

Después de haber estimado el modelo, es necesario evaluarlo para determinar si proporciona un ajuste y explicación satisfactorios para los datos que se han recolectado. Existen varias pruebas que pueden practicarse para tomar esta decisión. Estas pruebas son muy similares a las realizadas en el modelo de regresión simple. Sin embargo, como se anotó anteriormente, se enfatizará en los fundamentos que se encuentran tras estas pruebas, dejando los cálculos reales al computador. Los cálculos matemáticos y fórmulas necesarias se presentarán sólo para demostrar *conceptualmente* para qué está diseñada la prueba.

A. El error estándar de estimación

Al igual que la regresión simple, el error estándar de estimación puede utilizarse como medida de bondad de ajuste. Tiene la misma interpretación que con la regresión simple. Mide los grados de dispersión de los valores Y_i alrededor del plano de regresión, tal como se observa en la figura 12.2. Claro que entre menos dispersión se presente, más pequeño será el *Se* y más preciso será el modelo en su predicción y pronóstico.

Figura 12.2
Plano de regresión para Hop Scotch



El error estándar se calcula de la misma forma que la regresión simple.

El error estándar de estimación

$$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - k - 1}} \qquad [12.4]$$

en donde $n - k - 1$ es el número de grados de libertad y k es el número de variables a la derecha. El numerador bajo el radical de la fórmula (12.4) es la suma de los errores elevada al cuadrado y se minimizará de acuerdo con el concepto de mínimos cuadrados ordinarios. La pantalla 12.2 muestra una impresión en Minitab de los valores reales para los pasajeros (Y_i), el valor proyectado para los pasajeros ($\hat{Y}_i$), y el error ($Y_i - \hat{Y}_i$), y el error al cuadrado $(Y_i - \hat{Y}_i)^2$. La suma de esta última columna es la suma de los errores al cuadrado y se tiene que es 8.1016. El error estándar entonces será:

$$Se = \sqrt{\frac{8.1016}{15 - 2 - 1}} = 0.8217$$

Pantalla 12.2

Hop Scotch Data (Datos de Hop Scotch)

Row	PASS Y_i	Y-HAT $\hat{Y}_i$	RESIDUAL $(Y_i - \hat{Y}_i)$	RESDSQ $(Yi - \hat{Y}i)^2$
1	15	15.3834	-0.38338	0.14698
2	17	17.5238	-0.52382	0.27438
3	13	13.2429	-0.24294	0.05902
4	23	23.1055	-0.10547	0.01112
5	16	15.6139	0.38607	0.14906

RESDSQ = Cuadrado residual
Row = Fila
PASS = Pasajero

(Continúa)

```
   6      21      20.9650       0.03497     0.00122
   7      14      15.1528      -1.15282     1.32900
   8      20      19.8948       0.10519     0.01106
   9      24      25.0154      -1.01536     1.03095
  10      17      15.8445       1.15551     1.33521
  11      16      15.7475       0.25248     0.06375
  12      18      17.8015       0.19850     0.03940
  13      23      21.2571       1.74287     3.03761
  14      15      14.7205       0.27947     0.07810
  15      16      16.7313      -0.73128     0.53477
MTB >                                      8.10106
```

La impresión en Minitab en la pantalla 12.1 muestra el error estándar como 0.8217. Esto representa una mejoría sobre el error estándar del modelo de regresión simple en el capítulo anterior, el cual se informó que era 0.907.

B. Coeficiente de determinación múltiple

Al igual que con la regresión simple, el coeficiente de determinación múltiple se utiliza como una medida de bondad de ajuste. En aras de la conveniencia, el término *múltiple* se supone con frecuencia dado el contexto de la discusión, y la expresión se abrevia al coeficiente de determinación, lo mismo que para el modelo de regresión simple. Otra similitud entre el modelo simple y un modelo que contiene dos o más variables explicativas es la interpretación del coeficiente de determinación. En ambos casos, la porción del cambio en Y se explica mediante todas las variables independientes en el modelo.

Coeficiente de determinación Es el que mide la fuerza de la relación entre Y y las variables independientes.

Para medir dicha porción del cambio total en Y, explicada por el modelo de regresión, se utiliza la relación de la variación explicada con la variación total, de la misma manera como se hizo en el caso de la regresión simple. Como se dijo en el capítulo 11, por *variación* se entiende la variación en los valores Y observados (Y_i) de su media ($\overline{Y}$). La variación en Y, que se explica mediante el modelo, se refleja por la suma de los cuadrados de la regresión (SCR). La variación total en Y es a su vez medida por la suma total de cuadrados (SCT). Por tanto,

Coeficiente de determinación múltiple

$$R^2 = \frac{SCR}{SCT} \qquad [12.5]$$

Debido a que $SCT = SCR + SCE$, también se tiene que:

$$R^2 = 1 - \frac{SCE}{SCT} \qquad [12.6]$$

Vale la pena destacar que el coeficiente es r^2 en el modelo simple y R^2 en la discusión actual.

De la impresión en Minitab de la pantalla 12.1 que se mostró anteriormente, observamos que:

$$R^2 = \frac{SCR}{SCT}$$
$$= \frac{163.632}{171.733}$$
$$= 0.953$$

El R^2 también puede leerse directamente de la pantalla 12.1 como R - sq = 95.3%. Así, el 95.3% del cambio en el número de pasajeros que transporta Hop Scotch se explica mediante los cambios en la publicidad y en el ingreso nacional. Esto se compara de manera favorable con $r^2 = 0.93$ para el modelo de regresión simple del capítulo 11 que contiene sólo publicidad. Al incorporar *NI* (national income o ingreso nacional) como segunda variable independiente, se ha incrementado el poder explicativo del modelo de 93 a 95.3%. Al igual que con el modelo de regresión simple, siempre se encuentra que $0 \leq R^2 \leq 1$. Claro que entre mayor sea R^2, mayor poder explicativo tendrá el modelo.

C. El coeficiente de determinación corregido

Debido a su importancia, R^2 se reporta en la mayoría de los paquetes de computador. Es una forma rápida y fácil de evaluar el modelo de regresión para determinar qué tan bien se ajusta el modelo a los datos. Además de los coeficientes de regresión en sí mismos, R^2 es quizás el estadístico más comúnmente observado y el que se observa más de cerca en el análisis de regresión.

Sin embargo, para estadísticos descuidados –o poco escrupulosos– es posible inflar artificialmente R^2. Se puede incrementar R^2 simplemente adicionando otra variable independiente al modelo. Incluso, si alguna variable absurda sin un poder realmente explicativo se incorpora al modelo, R^2 aumentará. Hop Scotch podría "inflar" su R^2 adicionando al modelo como variable explicativa, las toneladas de trucha de mar que los practicantes de pesca deportiva cogen en la costa de Florida. Ahora, obviamente la pesca no tiene nada que ver con la lista de pasajeros de Hop Scotch. Sin embargo, probablemente existe aunque sea una pequeña correlación casual, bien sea positiva o negativa, entre la pesca y el viaje aéreo. Incluso un diminuto grado de correlación inflará R^2. Al sumar varias de estas variables explicativas "absurdas", R^2 podría incrementarse de manera ilegítima hasta que se aproximara al 100%. Un modelo de esta naturaleza puede parecer que se ajusta muy bien a los datos, pero produciría resultados infortunados en todo intento por predecir o pronosticar el valor de la variable independiente.

Por consiguiente, es una práctica común en regresión múltiple y correlación reportar el **coeficiente de determinación corregido**. Representado con el símbolo $\bar{R}^2$, y se lee "R barra al cuadrado", este estadístico se ajusta a la medida del poder explicativo para el número de grados de libertad. Debido a que el grado de libertad para *SCE* es $n-k-1$, agregar otra variable explicativa termina en la pérdida de otro grado de libertad. $\bar{R}^2$ *decrecerá* si se adiciona una variable que no ofrece suficiente poder explicativo como para justificar su pérdida en los grados de libertad. Si se reduce demasiado, se debe considerar su retiro.

El coeficiente de determinación corregido se obtiene dividiendo *SCE* y *SCT* por sus respectivos grados de libertad.

Coeficiente de determinación múltiple corregido

$$\bar{R}^2 = 1 - \frac{SCE/(n-k-1)}{SCT/(n-1)} \qquad [12.7]$$

Una fórmula más conveniente a nivel computacional para $\bar{R}^2$ es:

Coeficiente de determinación múltiple corregido

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1} \qquad [12.8]$$

Debido a que el numerador de la fórmula (12.7) es el *CME*, puede decirse que $\bar{R}^2$ es una combinación de dos medidas del desempeño del modelo de regresión: el cuadrado medio de error y el coeficiente de determinación.

Los datos para el modelo de Hop Scotch son:

$$\bar{R}^2 = 1 - (1 - 0.953)\frac{15 - 1}{15 - 2 - 1}$$
$$= 0.945$$

Después de corregir los grados de libertad, se obtiene $\bar{R}^2$ de 94.5%.

Como puede esperarse, la mayoría de los programas de computación también informan el coeficiente de determinación corregida. La pantalla 12.1 revela $R - s_e$ (adj) = 94.5%.

D. Evaluación del modelo como un todo

Dado el modelo de regresión, una de las primeras preguntas que se plantean es: "¿Tiene algún valor explicativo?" Esto puede responderse mejor realizando el análisis de varianza (ANOVA). El procedimiento del ANOVA prueba si *alguna* de las variables independientes tiene una relación con la variable dependiente. Si una variable independiente no está relacionada con la variable Y, su coeficiente debería ser cero. Es decir, si X_i no está relacionada con Y, entonces $\beta_i = 0$. El procedimiento ANOVA prueba la hipótesis nula de que todos los valores β son cero contra la alternativa de que *por lo menos un* β *no es cero*. Es decir

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$
$$H_A: \text{Al menos un } \beta \text{ no es cero}$$

Si no se rechaza la hipótesis nula, entonces no hay relación lineal entre Y y cualquiera de las variables independientes. Por otra parte, si la hipótesis nula se rechaza, entonces por lo menos una variable independiente está relacionada linealmente con Y.

El proceso ANOVA, necesario para probar la hipótesis nula, se presentó en el capítulo 10. Se establece una tabla ANOVA y se utiliza la prueba F. La tabla 12.2 proporciona el formato general para una tabla ANOVA para la regresión múltiple.

Tabla 12.2
Una tabla ANOVA generalizada

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Valor F
Entre muestras (tratamiento)	SCR	K	$\frac{SCR}{K}$	$F = \frac{CMR}{CME}$
Dentro de las muestras (error)	SCE	$n-k-1$	$\frac{SCE}{n-k-1}$	
Variación total	SCT	$n-1$		

Vale la pena destacar la similitud de la tabla 12.2 a las tablas ANOVA que ya se han visto. El grado de libertad para la suma de los cuadrados de la regresión es igual a k, el número de variables independientes del modelo, mientras que el grado de libertad para la suma de cuadrados del error es $n - k - 1$. Cada una de las sumas de cuadrados se encuentra exactamente igual que en la regresión simple.

$$SCT = \Sigma(Y_i - \bar{Y})^2 \qquad [12.9]$$

$$SCR = \Sigma(\hat{Y}_i - \bar{Y})^2 \qquad [12.10]$$

$$SCT = \Sigma(Y_i - \hat{Y}_i)^2 \qquad [12.11]$$

La tabla 12.3 proporciona los resultados en una tabla ANOVA para Hop Scotch Airlines. Esta información también puede observarse en la pantalla 12.1 (presentada anteriormente) y en la pantalla 12.3 (repetida nuevamente por conveniencia).

Tabla 12.3
Tabla ANOVA para Hop Scotch

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Valor *F*
Entre muestras (tratamiento)	163.632	2	81.816	121.18
Dentro de las muestras (error)	8.102	12	0.675	
Variación total	171.733	14		

Pantalla 12.3

Analysis of variance (Análisis de varianza)

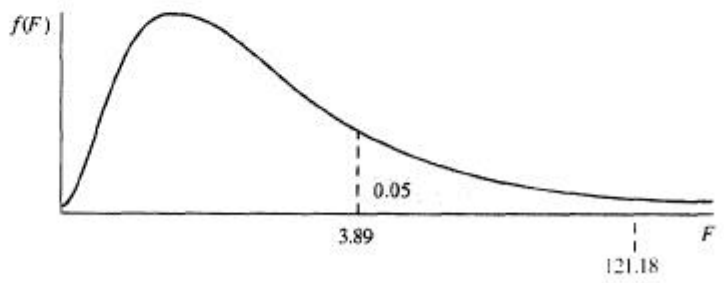
SOURCE	DF	SS	MS	F	p
Regression	2	163.632	81.816	121.18	0.000
Error	12	8.102	0.675		
Total	14	171.733			

SOURCE	DF	SEQ SS
ADV	1	161.044
NI	1	2.588

DF = Grados de libertad
SS = Suma de cuadrados
MS = Cuadrado medio
ADV = Publicidad
NI = Ingreso nacional

Debido a que la razón *F* es *CMR/CME*, los grados de libertad necesarios para realizar una prueba *F* vista en la tabla 12.3 son 2 y 12. Para probar la hipótesis, al nivel del 5%, se tiene de la tabla G (apéndice III) que $F_{0.05,2,12}$ es 3.89. La regla de decisión es: no rechazar si $F \leq 3.89$; rechazar si $F > 3.89$. Esto se muestra en la figura 12.3. debido a que $F = 121.18 > 3.89$, la hipótesis nula se rechaza. Se puede concluir al nivel del 5% que existe una relación lineal entre *Y* y por lo menos una de las variables independientes.

Figura 12.3
Prueba *F* para el modelo de regresión de Hop Scotch



E. Pruebas individuales para los coeficientes de regresión parcial

El siguiente paso lógico es probar cada coeficiente individualmente para determinar cuál es (cuáles son) significante(s). Primero se prueba la publicidad. El proceso es idéntico a la regresión simple:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

en donde β_1 es el coeficiente de regresión poblacional para la publicidad. No rechazar la hipótesis nula significa que la publicidad no contribuye con poder explicativo alguno al modelo, dado que el ingreso nacional ya está incluido.

Se utiliza la prueba t estándar con $n - k - 1$ grados de libertad.

Prueba de hipótesis para la significancia del coeficiente de regresión parcial	$t = \dfrac{b_1 - \beta_1}{s_{b_1}}$	[12.12]

en donde s_{b_1} es el error estándar del coeficiente de regresión. De la misma manera que con la mayoría de estadísticos relacionados con la regresión múltiple, s_{b_1} es difícil de calcular manualmente. Por fortuna, la mayoría de los paquetes para computador reportan esta información. Como se observa en la impresión de Minitab en la pantalla 12.4, el valor t para la publicidad es

$$t = \frac{0.8398 - 0}{0.1419} = 5.92$$

Pantalla 12.4

Regression Analysis (Análisis de regresión)

The regression equation is (La ecuación de regresión es)
PASS = 3.53 + 0.840 ADV + 1.44 NI

Predictor	Coef	Stdev	t-ratio	p
Constant	3.5284	0.9994	3.53	0.004
ADV	0.8397	0.1419	5.92	0.000
NI	1.4410	0.7360	1.96	0.074

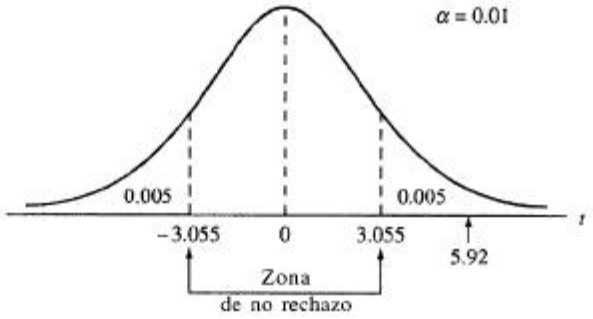
Stdev = Desviación estándar
t-ratio = razón t
PASS = Pasajeros

Si se selecciona un valor α del 1%, $t_{.01,12} = 3.055$. Como se observa en la figura 12.4,

Regla de decisión: "No rechazar si t está entre ± 3.055. De lo contrario rechazar".

Debido a que $t = 5.92 > 3.055$, se rechaza la hipótesis nula. Al nivel de significancia del 1%, la publicidad contribuye significativamente al poder explicativo del modelo, aun después de haber incluido el ingreso nacional. Esto se confirma mediante el valor p en la pantalla 12.4 de 0.000. El valor p, es el valor α más bajo en el que se puede fijar y sin embargo rechazar la hipótesis nula. Debido a que el valor α del 1% es mayor que 0.000 se rechaza la hipótesis nula.

Figura 12.4
Prueba de significancia para publicidad



Vale la pena recordar del capítulo 11 que cuando la publicidad era la única variable explicativa se reportó un valor t de 13.995. ¿Por qué es diferente ahora? El valor t de 5.92 en este modelo mide la contribución adicional de publicidad dado que el ingreso nacional ya está incluido. Al rechazar la hipótesis nula, se ha determinado al nivel de significancia del 1% que la publicidad contribuye significativamente al poder explicativo del modelo, incluso después que se ha adicionado el ingreso nacional.

La misma prueba de significancia se realiza sobre la segunda variable explicativa, el ingreso nacional.

$$H_0: \beta_2 = 0$$
$$H_A: \beta_2 \neq 0$$

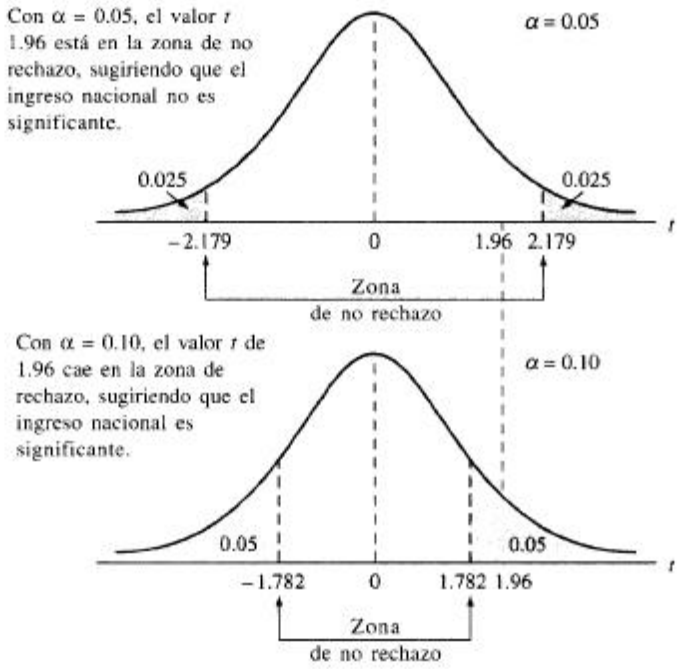
La pantalla 12.4 revela que:

$$t = \frac{1.441 - 0}{0.7360} = 1.96$$

Si $\alpha = 5\%$, $t_{0.05,12} = 2.179$. Entonces, como lo muestra la figura 12.5,

Regla de decisión: "No rechazar si t está entre ± 2.179. De lo contrario rechazar".

Figura 12.5
Pruebas de significancia para el ingreso nacional



Claramente, la hipótesis nula $\beta_2 = 0$ *no* se rechaza. No se puede concluir, al nivel de significancia del 5%, que el ingreso nacional contribuye al poder explicativo del modelo si la publicidad ya está incluida como variable explicativa.

De acuerdo con el valor p, se puede bajar el nivel de significancia a sólo 7.4% y sin embargo rechazar la hipótesis nula. Si el valor α se fija más bajo que 7.4%, tal como 5%, no se rechaza la hipótesis nula. Pero si α se fija al 10%, $t_{0.10,12} = \pm 1.782$,

Regla de decisión: "No rechazar si t está entre $\pm$ 1.782. De lo contrario rechazar".

Debido a que el valor t es 1.96, la hipótesis nula se rechaza al nivel de significancia del 10%. Esto también se refleja en la figura 12.5.

Ejercicios de la sección

1. Dado el modelo de regresión $\hat{Y} = 40 + 3X_1 - 4X_2$,

a. Interprete los coeficientes

b. Estime Y si $X_1 = 5$ y $X_2 = 10$.

2. Dada la ecuación de regresión, con valores t en paréntesis,

$$\hat{Y} = 100 + 17X_1 + 80X_2$$
$$(0.73) \quad (6.21)$$

¿Cómo podría usted mejorar este modelo?

3. Para la ecuación

$$\hat{Y} = 100 - 20X_1 - 40X_2$$

a. ¿Cuál es el impacto estimado de X_1 sobre Y?

b. ¿Qué condiciones respecto a X_2 deben observarse al responder la parte *a*?

4. Una función de demanda se expresa como

$$\hat{Q} = 10 + 12P + 8I$$

en donde Q es la cantidad demandada, P es el precio, e I es el ingreso del consumidor. ¿Cómo respondería a esta ecuación?

5. Un modelo de regresión con $n = 20$ observaciones y tres variables independientes reporta un valor F de 3.89. ¿Alguna de las variables independientes es significante al nivel del 1%?

6. Se presenta una regresión de consumo sobre el ingreso y la riqueza, con valores t en paréntesis. ¿Las variables independientes son significantes al nivel del 5%? Eran 100 observaciones.

$$\hat{C} = 52 + 17.3I + 4.6W$$
$$(12.2) \quad (0.87)$$

7. Un modelo de regresión del consumo (C) sobre el ingreso (I) y la riqueza (W) produjo los siguientes resultados:

$$R^2 = 0.86$$
$$\bar{R}^2 = 0.79$$
$$F = 17.42$$
$$\hat{C} = 402 + 0.83I + 0.71W$$
$$(0.71) \quad (6.21) \quad (5.47)$$

en donde los valores t se muestran en paréntesis. Eran 25 observaciones en el conjunto de datos.

a. ¿Cuál es el significado del término intercepto?

b. ¿Los coeficientes son significantes al nivel del 5%?

c. ¿El modelo es significante al nivel del 10%?

8. Batex Associates vende aceite de calefacción a los residentes de las áreas rurales de Virginia. El director del departamento de mercadeo en Batex desarrolló un modelo de regresión del consumo de aceite de calefacción (en galones) de sus clientes sobre la temperatura local (TEMP), la población (POP) por condados, y el precio del aceite. Los resultados están dados en la siguiente impresión en Minitab.

a. ¿Cuál es el modelo?

b. ¿Los datos sugieren que las tres variables son significantes al nivel del 5%?

c. ¿Cuál es el nivel de significancia más bajo para cada variable?

d. ¿Cómo sugeriría usted mejorar el modelo?

e. ¿Qué tan fuerte es la relación?

```
The regression equation is
OIL = 20.7 - 0.853 TEMP + 0.113 POP + 0.00193 PRICE

Predictor      Coef        Stedv   t-ratio       p
Constant     20.706        3.455      5.99   0.000
TEMP        -0.8530       0.6220     -1.37   0.085
POP         0.11287      0.02603      4.34   0.000
PRICE      0.001929     0.003570      0.54   0.595

s = 1.478  R-sq = 95.8%     R-sq (adj) = 95.2%
```

```
OIL     = Aceite
POP     = Población
t-ratio = razón t
Stdev   = Desviación
          estándar
R-sq    = r²
```

9. Un campo de la economía denominado capital humano con frecuencia sostiene que el ingreso de una persona (I) podría determinarse sobre la base de su 1) nivel de educación (E), 2) capacitación (T), y 3) nivel general de salud (H). Utilizando 25 empleados en una pequeña empresa textil en Carolina del Norte, un investigador hizo regresión de ingreso sobre las otras tres variables y obtuvo los siguientes resultados:

$$\hat{I} = \underset{}{27.2} + \underset{(3.70)}{3.7E} + \underset{(6.21)}{1.7T} + \underset{(4.32)}{3.05H} + \underset{(6.79)}{}$$

$$R^2 = 0.67 \quad F = 5.97$$

I se mide en unidades de 1,000 dólares, E y T se miden en años, y H se mide en términos de un índice escalado de la salud propia:

a. Si la educación propia incrementa en dos años, ¿qué pasa con su ingreso?

b. ¿El modelo es significante al nivel del 5%? Plantee la hipótesis, la regla de decisión, y la conclusión.

c. Determine cuál(es) variable(s) es(son) significante(s) al nivel del 10%. Plantee la hipótesis, la regla de decisión y la conclusión.

d. ¿Cuál es el valor del coeficiente de determinación corregido?

10. ¿Qué significa si la hipótesis nula en una prueba para una β_i no se rechaza?

11. En referencia al problema anterior si H_0: β_i no se rechaza, según el modelo, ¿qué pasará a Y si X cambia en una unidad? ¿En dos unidades?

12. Considerando el siguiente modelo con $n = 30$.

$$\hat{Y} = 50 + 10X_1 + 80X_2$$
$$R^2 = 0.78 \qquad S_{b_1} = 2.73 \qquad S_{b_2} = 4.71$$

¿Qué variable(s) es (son) significantes al nivel del 5%? Plantee la hipótesis, la regla de decisión y saque una conclusión.

13. Los economistas han sostenido durante mucho tiempo que la demanda de dinero por una comunidad se ve afectada por 1) el nivel de ingreso y 2) la tasa de interés. A medida que el ingreso aumenta, las personas desean guardar más dinero para facilitar el incremento en sus transacciones diarias. A medida que la tasa de interés aumenta, las personas deciden mantener menos dinero debido a la oportunidad de invertirlo a una tasa de interés más elevada.

Un economista del gobierno federal hace regresión de la demanda de dinero (M) el ingreso (I) y las tasas de interés (r), en donde M se expresa en cientos de dólares e I en miles de dólares. El modelo es:

$$\hat{M} = 0.44 + 5.49I + 6.4r$$

Una tabla parcial ANOVA es

Fuente	Suma de cuadrados	Grados de libertad
Entre muestras	93.59	2
Dentro de muestras	1.42	9

a. De acuerdo con la teoría de demanda de dinero, ¿los signos de los coeficientes son como se esperaban? Explique.

b. Pruebe todo el modelo a $\alpha = 0.01$.

14. Dadas las condiciones del ejercicio 13, si el error estándar del coeficiente para I es 1.37 y el de r es 43.6, determine cuál variable es (o cuáles variables son) significante al nivel del 1%. Plantee la hipótesis, la regla de decisión y la conclusión.

15. Un analista económico de IBM desea pronosticar las ventas regionales (S) en cientos de dólares con base en el número de personal de ventas (P), el número de nuevos negocios que comienzan en la región (B), y alguna medida de los precios. Como representante de la última variable, el analista utiliza los cambios en el Índice de precios al consumidor (IPC). Luego recolecta datos para las 10 regiones de ventas y deriva el siguiente modelo y tabla ANOVA parcial:

$$\hat{S} = -1.01 + 0.422P + 0.091B - 1.8\text{IPC}$$

Fuente	Suma de cuadrados	Grados de libertad
Entre muestras	391.57	3
Dentro de muestras	31.33	6

a. Pruebe la significancia de todo el modelo al 1%. Plantee la hipótesis, la regla de decisión y la conclusión.

b. Si los errores estándar de los coeficientes de P, B, y IPC son 0.298, 0.138 y 2.15, respectivamente, pruebe cada coeficiente al nivel del 10%. Plantee la hipótesis, la regla de decisión y la conclusión en cada caso.

c. ¿Cómo puede usted conciliar los hallazgos de las partes *a* y *b*?

12.4 Presencia de multicolinealidad

Anteriormente se expresó el peligro de la multicolinealidad. Este problema surge cuando una de las variables independientes está relacionada linealmente con una o más de las otras variables independientes. Dicha situación contraviene una de las condiciones de la regresión múltiple. Específicamente, la multicolinealidad ocurre si existe una alta correlación entre dos variables independientes, X_i, y X_j. En el capítulo 11 se discutió el coeficiente de correlación r para la variable dependiente y una sola variable independiente. Si este mismo concepto se aplica a dos variables independientes, X_i y X_j, en regresión múltiple, se puede calcular el coeficiente de correlación r_{ij}. Si r_{ij} es alto, existe multicolinealidad.

¿Qué es alto? Desafortunadamente, no existe respuesta a esta pregunta crítica. No existe un punto de corte mágico en el cual se pueda determinar que la correlación es demasiado alta y por ende existe multicolinealidad. La multicolinealidad es un problema de grado. En cualquier momento dos o más variables independientes están relacionadas linealmente, entonces existe algún grado de multicolinealidad. Si la presencia se vuelve demasiado pronunciada, el modelo se ve afectado negativamente. Lo que se considera demasiado alto todavía es un llamado al juicio por parte del investigador. Es necesario algún discernimiento para hacer que dicho llamado se proporcione en esta sección.

Se asume que se están utilizando las técnicas de regresión para estimar una curva de demanda (o función de demanda) para su producto. Al reconocer que el número de consumidores está relacionado con la demanda, se seleccionan como variables explicativas.

X_1 = Todos los hombres en el área del mercado

X_2 = Todas las mujeres en el área del mercado

X_3 = Población total en el área del mercado

Obviamente, X_3 es una combinación lineal de X_1 y X_2 ($X_3 = X_1 + X_2$). La correlación r_{13} entre X_1 y X_3 y la correlación r_{23} entre X_2 y X_3 es muy alta. Esto garantiza la presencia de la multicolinealidad y crea muchos problemas en el uso de las técnicas de regresión. A continuación se presenta una discusión sobre algunos de los problemas más comunes.

A. Los problemas de la multicolinealidad

Uno de los problemas más preocupantes de la multicolinealidad surge de la incapacidad de separar los efectos individuales de cada variable independiente sobre Y. Ante la presencia de la multicolinealidad, es imposible desenmarañar los efectos de cada X_i. Se supone en el modelo

$$\hat{Y} = 40 + 10X_1 + 80X_2$$

X_1 y X_2 mostraron un alto grado de correlación. En este caso, el coeficiente de 10 para X_1 puede no representar el efecto verdadero de X_1 sobre Y. Los coeficientes de regresión se vuelven no confiables y no pueden tomarse como estimaciones del cambio en Y dado un cambio de una unidad en la variable independiente.

Además, los errores estándar de los coeficientes, s_{b_i}, se desbordan. Si se toman dos o más muestras del mismo tamaño, se encontraría una gran variación en los coeficientes. En el modelo especificado anteriormente, en lugar de 10 como coeficiente de X_1, una segunda muestra puede dar un coeficiente de 15 o 20. Si b_i varía mucho de una muestra a la siguiente, se debe cuestionar la exactitud.

La multicolinealidad incluso puede hacer que el signo del coeficiente sea opuesto al que la lógica dicta. Por ejemplo, si se incluye el precio como una variable en la estimación de la curva de demanda, se puede encontrar que tomó un signo positivo. Esto implica que como el precio de un bien sube, los consumidores compran más de éste. Esta es una evidente contravención de la lógica que está tras la teoría de la demanda.

B. Detección de la multicolinealidad

Quizás la forma más directa para probar la multicolinealidad es producir una **matriz de correlación** para todas las variables del modelo, como lo muestra la impresión en Minitab en la pantalla 12.5. El valor de 0.870 para la correlación entre las dos variables independientes indica que NI y ADV están relacionadas muy de cerca. Aunque no existe un valor predeterminado para r_{ij} que señale la aparición de la multicolinealidad, es muy probable que un valor de 0.870 lo suficientemente alto indique un problema importante.

Pantalla 12.5

```
Correlations (Correlaciones) (Pearson)

          ADV      PASS                    PASS = Pasajeros
PASS    0.968                             ADV  = Publicidad
NI      0.870     0.903                   NI   = Ingreso nacional
```

Puede eliminarse algo de conjetura utilizando una prueba t para determinar si el nivel de correlación entre dos variables independientes difiere significativamente de cero. Dada una relación de no cero entre publicidad e ingreso nacional de $r = 0.870$ en la muestra, se desea probar la hipótesis de que la correlación es cero al nivel poblacional. Se probará la hipótesis de que:

$$H_0: \rho_{12} = 0$$
$$H_A: \rho_{12} \neq 0$$

en donde ρ_{12} es el coeficiente de correlación poblacional para X_1 (ADV) y X_2 (NI). Se puede hacer esto utilizando las técnicas del capítulo 11. Allí se demostró que:

$$t = \frac{r_{12}}{S_r}$$

en donde r_{12} es la correlación muestral entre publicidad (X_1) e ingreso nacional (X_2) y

$$S_r = \sqrt{\frac{1 - r_{12}^2}{n - 2}}$$

Como ilustración, la hipótesis de que $\rho_{12} = 0$, en donde ρ_{12} es el coeficiente de correlación poblacional para las dos variables independientes, se realiza así:

$$S_r = \sqrt{\frac{1 - (0.87)^2}{15 - 2}}$$
$$= 0.1367$$

Por tanto,

$$t = \frac{0.870}{0.1367}$$
$$= 6.36$$

Si se determina α al 5%, el valor crítico de $t_{0.05,13} = 2.16$. Hay n - 2 (no n –k - 1) grados de libertad.

Regla de decisión: "No rechazar si $-2.16 \leq t \leq 2.16$. Rechazar si $t < -2.26$ o > 2.16".

Debido a que $t = 6.36 > 2.16$, se puede rechazar la hipótesis nula de que no hay correlación entre X_1 y X_2 ($\rho_{12} = 0$). Existe alguna multicolinealidad. Esto no significa que el modelo es defectuoso irrevocablemente. De hecho, muy pocos modelos están totalmente libres de multicolinealidad. Cómo manejar este problema será un tema que se discutirá en breve.

Otra forma de detectar la multicolinealidad es comparar los coeficientes de determinación entre la variable dependiente y cada una de las variables independientes. De la pantalla 12.5 se encuentra correlación entre pasajeros y publicidad que es $r^2 = (0.968)^2 = 0.937$, mientras que entre pasajeros e ingreso nacional es $r^2 = (0.903)^2 = 0.815$. Sin embargo las dos variables juntas revelaron un R^2 de sólo 0.953. Si se toman independientemente, las dos variables independientes explican el 93.7% y el 81.5%. Aparentemente existe alguna superposición en su poder explicativo. Incluir la segunda variable de *NI* hizo muy poco por incrementar la capacidad del modelo para explicar el número de pasajeros. Gran parte de la información sobre pasajeros, ya proporcionada por la publicidad, simplemente se duplica por el *NI*. Este es un indicio de que la multicolinealidad puede estar presente.

Una tercera forma de detectar la multicolinealidad es analizar el **factor de inflación de varianza** (FIV). El FIV relacionado con toda variable X se halla haciendo regresión de ésta sobre todas las otras variables X. El R^2 resultante se utiliza luego para calcular el FIV de esa variable. EL FIV para todo X_i representa la influencia de dicha variable en la multicolinealidad.

Factor de inflación de la varianza El FIV para toda variable independiente es una medida del grado de multicolinealidad en que contribuye dicha variable.

Debido a que hay sólo dos variables independientes en el modelo de Hop Scotch, se hace regresión de X_1 sobre las otras variables independientes (X_2), o se hace regresión de X_2 sobre las otras variables independientes (X_1) y da el mismo coeficiente de correlación ($r_{12} = 0.87$), como se muestra en la pantalla 12.5. El FIV para cualquier variable independiente dada X_i es:

Factor de inflación de la varianza para X_i

$$\text{FIV}(X_i) = \frac{1}{1 - R_i^2} \qquad [12.13]$$

en donde R_i^2 es el coeficiente de determinación obtenido al hacer la regresión de X_i sobre todas las otras variables independientes. Como se anotó anteriormente, la multicolinealidad produce un incremento en la variación, o error estándar, del coeficiente de regresión. El FIV mide el incremento en la varianza del coeficiente de regresión por encima del que ocurriría si no estuviera presente la multicolinealidad.

El *FIV* para publicidad es:

$$FIV(X_1) = \frac{1}{1 - (0.87)^2}$$
$$= 4.1$$

El mismo FIV para X_2 se hallaría ya que hay sólo dos variables independientes. Esto podría interpretarse como la varianza en b_1 y b_2 que es más de cuatro veces lo que debería ser sin la multicolinealidad en el modelo. Sin embargo, en general, la multicolinealidad no se considera un problema significativo a menos que el FIV de una sola X_i mida por lo menos 10, o la suma de los FIV's para todas las X_i sea de por lo menos 10. Claro que los paquetes de computación proporcionarán el FIV, tal como se muestra en la impresión de Minitab de la pantalla 12.6.

Pantalla 12.6

Regression Analysis (Análisis de regresión)

The regression equation is

PASS = 3.53 + 0.840 ADV + 1.44 NI

Predictor	Coef	Stdev	t-ratio	p	**VIF**
Constant	3.5284	0.9994	3.53	0.004	
ADV	0.8397	0.1419	5.92	0.000	**4.1**
NI	1.4410	0.7360	1.96	0.074	**4.1**

With only two explanatory variables, both will have the same VIF.
(Sólo con dos variables explicativas, ambas con el mismo VIF.)

PASS = Pasajeros
NI = Ingreso nacional
VIF = Factor de inflación de la varianza
t-ratio = razón estándar

Otros indicios de multicolinealidad incluyen grandes cambios en los coeficientes o en su signo cuando existe un cambio pequeño en el número de observaciones. Además, si la razón F es significante y los valores t no lo son, puede estar presente la multicolinealidad. Igualmente si la suma o eliminación de una variable produce grandes cambios en los coeficientes o sus signos, puede existir la multicolinealidad.

En resumen, ante la presencia de la multicolinealidad tenemos:

1. Incapacidad para separar el efecto neto de las variables independientes individuales sobre Y.
2. Un error estándar exagerado para los coeficientes b.
3. Signos algebraicos de los coeficientes que contravienen la lógica.
4. Una alta correlación entre variables independientes, y un FIV elevado.
5. Grandes cambios en los coeficientes o sus signos, si el número de observaciones se cambia en una sola observación.
6. Una razón F significativa combinada con razones t no significativas.
7. Grandes cambios en los coeficientes o en sus signos cuando se adiciona o se quita una variable.

C. Corrigiendo la multicolinealidad

¿Qué puede hacerse para eliminar o mitigar la influencia de la multicolinealidad? Quizá la solución más lógica es la eliminación de la variable causante. Si X_i y X_j están relacionadas muy de cerca, una de ellas puede sencillamente eliminarse del modelo. Después de todo, debido a la superposición, la inclusión de la segunda variable agrega muy poco a la explicación de Y.

La pregunta sería ¿cuál de ellas debería eliminarse? Haciendo referencia al modelo de Hop Scotch, puede ser aconsejable eliminar NI debido a que su correlación con Y es menor que la de publicidad. Las pruebas t practicadas también sugieren que NI no era significante al nivel del 5%.

Sin embargo, al eliminar simplemente una de las variables esto puede conllevar al **sesgo de especificación,** en el cual el formato del modelo está en desacuerdo con su base teórica. Debe evitarse la multicolinealidad, por ejemplo, si el ingreso se eliminara de una expresión funcional para la demanda de los consumidores. Sin embargo, la teoría económica, así como el simple sentido común, dicen que el ingreso debería incluirse en todo intento por explicar el consumo.

Sesgo de especificación Una especificación errónea de un modelo a causa de la inclusión o exclusión de ciertas variables que terminan en una contravención de los principios teóricos, esto es lo que se denomina sesgo de especificación.

Si se prohíbe eliminar una variable debido a algún sesgo resultante, se puede con frecuencia reducir la multicolinealidad cambiando la forma de la variable. Quizá dividiendo los valores originales de la variable causante por la población, para así obtener una cifra per cápita lo cual sería benéfico. Adicionalmente, dividir ciertas medidas monetarias por el índice de precios (como el Índice de Precios al Consumidor) y por ende obtener una medida en términos "reales", también es un método efectivo de eliminar la multicolinealidad. Ambos procedimientos podrían aplicarse al *NI*.

También es posible combinar dos o más variables. Esto podría hacerse con el modelo para la demanda del consumidor, el cual empleó X_1 = hombres, X_2 = mujeres, y X_3 = población total. Las variables X_1 y X_2 podrían sumarse para formar X_3. El modelo entonces constaría de una sola variable explicativa.

En cualquier evento, debería reconocerse que existe algún grado de multicolinealidad en la mayoría de los modelos de regresión que contienen dos o más variables independientes. Entre más grande sea el número de variables independientes mayor será la probabilidad de multicolinealidad. Sin embargo, esto no necesariamente resta méritos a la utilidad del modelo ya que el problema de multicolinealidad puede no ser grave. La multicolinealidad causará grandes errores en los coeficientes individuales, aunque el efecto combinado de estos coeficientes no sea drásticamente alterado. Un modelo de predicción diseñado para predecir el valor de Y, con base en todos los X_i tomados en combinación, darán una precisión considerable. Sólo los modelos explicativos creados para explicar la contribución al valor de Y por cada X_i, tienden a colapsarse ante la multicolinealidad.

Ejercicios de la sección

16. Defina la *multicolinealidad*. Explique claramente todos los problemas que puede producir un modelo de regresión.

17. ¿Por qué la multicolinealidad incrementa la probabilidad de un error Tipo II al probar una hipótesis sobre un coeficiente de regresión?

18. Describa las pruebas que pueden utilizarse para detectar la multicolinealidad.

19. ¿Cómo se calcula un factor de inflación de varianza? ¿Qué mide exactamente?

20. Un economista de la Federal Research Board propuso estimar el promedio industrial del Dow Jones utilizando como variables explicativas X_1, tasa de interés sobre los bonos corporativos AAA; X_2 tasas de interés sobre los valores de la Tesorería de Estados Unidos. Se pide su consejo. ¿Cómo respondería usted y qué problema estadístico es probable que encuentre?

12.5 Comparación de los coeficientes de regresión

Después de desarrollar el modelo completo, existe con frecuencia la tendencia a comparar los coeficientes de regresión para determinar cuál variable ejerce más influencia en Y. Esta tentación peligrosa debe evitarse. Para el modelo:

$$\hat{Y} = 40 + 10X_1 + 200X_2$$

en donde Y es toneladas de producción, X_1 unidades del insumo de trabajo, y X_2 es unidades de ingreso de capital, se puede concluir que el capital es más importante que el trabajo al determinar la producción, ya que tiene el coeficiente más grande. Después de todo, un incremento de una unidad de capital, manteniendo constante el trabajo, resulta en un incremento de 200 unidades en la producción. Sin embargo, tal comparación no es posible. Todas las variables se miden en unidades totalmente diferentes: una en unidades de peso, otra en número de personas y una tercera en máquinas.

Al medir todas las variables de la misma manera esto nos permite juzgar el impacto relativo de las variables independientes con base en el tamaño de sus coeficientes. Se supone que se plantea un modelo en términos de unidades monetarias, como:

$$\hat{Y} = 50 + 10{,}000X_1 + 20X_2$$

en donde Y está en dólares, X_1 en unidades de US$1,000, y X_2 está en centavos. A pesar del gran coeficiente para X_1, no es posible concluir que es de mayor impacto. Un incremento de US$1,000 (1 unidad) en X_1 incrementa Y en 10,000 unidades. Un incremento de US$1,000 (100,000 unidades) en X_2 incrementará Y en 2,000,000 unidades (100,000 x 20).

Incluso si se expresa Y, X_1, y X_2 en unidades de US$1, no se puede comparar el impacto relativo de X_1 y X_2 en los cambios en Y. Factores diferentes al coeficiente de una variable determinan su impacto total en Y. Por ejemplo, la varianza en una variable es muy importante al determinar su influencia en Y. La varianza mide con qué frecuencia y cuánto cambia una variable. Por tanto, una variable puede tener un coeficiente grande y cada vez que cambia afecta a Y notablemente. Pero si su varianza es muy pequeña y cambia sólo una vez en un milenio, su impacto global en Y será insignificante.

Para compensar estas deficiencias, algunas veces se mide la respuesta de Y a los cambios en los coeficientes de regresión estandarizados. **Los coeficientes de regresión estándar**, también denominados **coeficientes beta** (no confundir con el valor β, el cual es el coeficiente desconocido a nivel poblacional), reflejan el cambio en la respuesta promedio de Y, calculada en el número de desviaciones estándar de Y, de los cambios en X_i, medida en el número de desviaciones estándar de X_i. El efecto que se pretende con el cálculo de los valores β es hacer que los coeficientes «no tengan dimensiones".

El valor β para una variable explicativa X_i se calcula así:

Coeficiente beta o estandarizado para X_i	$\text{Beta} = \dfrac{b_i}{s_Y / s_{X_i}}$	[12.14]

en donde s_y y s_{xi} son desviaciones estándar de la variable dependiente Y y la variable independiente X_i respectivamente. Dado que estos valores son 3.502 para la variable dependiente pasajeros, y 0.605 para el ingreso nacional en este ejemplo, el valor β para el ingreso nacional se convierte:

$$\text{Beta} = \frac{1.441}{3.502/0.605} = 0.2436$$

Así, un cambio en una desviación estándar en el ingreso nacional da un cambio en la desviación estándar de pasajeros en 0.2436. De igual forma, β para publicidad es 0.7519. Esto puede sugerir que la publicidad tiene un impacto más fuerte en los pasajeros. Sin embargo, ante la presencia de la multicolinealidad, incluso estos coeficientes estandarizados sufren de muchas de las deficiencias que los coeficientes normales. Por ende, se considera una práctica deficiente medir la importancia de una variable con base en su coeficiente β.

12.6 Regresión paso a paso

Muchos paquetes modernos de computación proporcionan un procedimiento que ofrecen al estadístico la opción de permitir que el computador seleccione las variables independientes deseadas, de una lista de posibilidades establecida previamente. El estadístico proporciona los datos para varias variables explicativas potenciales y, luego, con ciertos comandos, da la orden al computador para que determine cuáles de las variables son las más adecuadas para formular el modelo completo.

De esta forma, el modelo de regresión se desarrolla por etapas; esto es lo que se conoce como **regresión paso a paso**. Puede tomar la forma de: 1) eliminación hacia atrás o 2) selección hacia delante. Se analizará cada una de ellas.

A. Eliminación hacia atrás

Para ejecutar la eliminación hacia atrás, se da la orden al computador para que calcule todo el modelo utilizando todas las variables independientes. Los valores t luego se computan para todos los coeficientes. Si alguno es insignificante, el computador elimina el que tenga el valor t más próximo a cero y calcula el modelo nuevamente. Esto continúa hasta que todos los b_i restantes sean significativamente diferentes de cero.

B. Selección hacia delante

Como su nombre lo indica, la selección hacia delante es el opuesto a la eliminación hacia atrás. Primero, la variable que esté más altamente correlacionada con Y se selecciona para ser incluida en el modelo. El segundo paso es la selección de una segunda variable con base en su capacidad para explicar Y, dado que la primera variable ya está en el modelo. La selección de la segunda variable se basa en su *coeficiente parcial de determinación*, el cual es una contribución de la variable al poder explicativo del modelo, dada la presencia de la primera variable.

Por ejemplo, se asume que la primera variable seleccionada es X_5. Se calcula todo modelo posible de dos variables en el cual una de tales variables es X_5. El modelo que produce el R^2 más alto es el que se selecciona. Este proceso continúa hasta que todas las variables X están en el modelo o hasta que la inclusión de otra variable no termine en un incremento significativo en R^2.

Aunque la regresión por pasos parece ser un método efectivo y conveniente para la especificación del modelo, deben tomarse ciertas precauciones. El proceso explorará los datos, a la espera de un modelo estadísticamente más preciso con el R^2 más alto. Sin embargo, un computador no puede pensar o razonar, y el modelo resultante puede "funcionar" estadísticamente pero puede ser contrario a todo principio lógico o teórico, y, por tanto, sufrir de un sesgo de especificación. La regresión paso a paso debería utilizarse con extrema precaución, y todo modelo formulado de esta forma debería escudriñarse muy de cerca.

12.7 Variables dummy

En los esfuerzos de búsqueda se pueden hallar muchas variables que son útiles para explicar el valor de la variable dependiente. Por ejemplo, años de educación, entrenamiento y experiencia son instrumentos para determinar el nivel de ingresos de una persona. Estas variables pueden medirse numéricamente, y prestarse al análisis estadístico.

Sin embargo, tal no es el caso con muchas otras variables que también son útiles al explicar los niveles de ingreso. Los estudios han demostrado que el género y la geografía también tienen un poder explicativo considerable. Una mujer que haya completado el mismo número de años de educación y capacitación que un hombre no obtendrá el mismo ingreso. Un trabajador del noreste puede no ganar lo mismo que un trabajador del sur haciendo un trabajo similar. Tanto el género como la geografía pueden ser variables explicativas altamente

útiles en el esfuerzo por predecir el ingreso propio. Debido a que ninguna variable puede expresarse de inmediato numéricamente, no pueden incluirse directamente en un modelo de regresión. Por tanto, se debe modificar la forma de estas variables no numéricas, de tal manera que se puedan incluir en el modelo y por ende ganar el poder explicativo adicional que ofrecen.

Las variables que no están expresadas de forma directa y cuantitativa se llaman **variables cualitativas** o **variables dummy.** Veamos otro ejemplo, las ventas de una empresa pueden depender de la estación. Los trajes de baño probablemente se venden mejor en primavera que en otoño o en invierno. Se venden más palas para la nieve en diciembre que en julio. Este factor estacional puede captarse tomando en cuenta el tiempo del año (otoño, invierno, primavera o verano), una variable que no puede medirse numéricamente. Si una persona es casada, soltera o divorciada puede afectar sus gastos para efectos recreacionales, mientras que el lugar de residencia (urbano, suburbano, o rural) probablemente tendrá impacto en la valorización de impuestos de una persona. En todos estos casos, las variables que se desean medir no se expresan numéricamente. Se deben entonces utilizar las variables dummy para obtener una descripción más completa del impacto de estas medidas no numéricas.

Variable dummy La variable que da cuenta de la naturaleza cualitativa de una variable e incorpora su poder explicativo dentro del modelo, se denomina variable dummy.

Como gerente regional de una cadena de tiendas por departamentos se desea estudiar la relación entre los gastos de los clientes y seleccionar las variables que pueden explicar tales gastos. Además de la selección lógica del ingreso como variable explicativa, se considera que el sexo de un cliente también puede hacer parte en la explicación de los gastos. Por tanto, se recolectan 15 observaciones para estas tres variables: los gastos en dólares, el ingreso en dólares y el sexo.

Pero ¿cómo se codifican los datos para sexo en el modelo? No se puede simplemente especificar M y F para masculino y femenino, debido a que estas letras no pueden manipularse matemáticamente. La solución se encuentra asignando valores de 0 o 1 a cada observación con base en el sexo. Por ejemplo, se puede decidir registrar un 0 si la observación es masculino y 1 si la observación es femenino. Lo contrario es igualmente posible. Se podría igualmente codificar 0 si es femenino y 1 si es masculino (en breve se analizarán los efectos de este esquema de codificación alterno).

Se supone que se decide registrar un 0 si la observación es masculino y un 1 si es femenino. El conjunto completo de datos para $n = 15$ observaciones aparece en la tabla 12.4 con Y en dólares y X_1 en unidades de US$1,000. Vale la pena notar que X_2 contiene sólo valores de 0 para masculino y 1 para femenino.

Tabla 12.4
Datos para el estudio de los gastos de los clientes

Observación	Gastos (Y)	Ingresos (X_1)	Sexo (X_2)
1	51	40	1
2	30	25	0
3	32	27	0
4	45	32	1
5	51	45	1
6	31	29	0
7	50	42	1
8	47	38	1
9	45	30	0
10	39	29	1
11	50	41	1
12	35	23	1
13	40	36	0
14	45	42	0
15	50	48	0

Utilizando los procedimientos de *MCO* que se discutieron en el capítulo 11, la ecuación de regresión es:

$$\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + b_2X_2 \\ &= 12.21 + 0.791X_1 + 5.11X_2 \\ &\qquad\quad (0.000) \quad (0.010)\end{aligned}$$

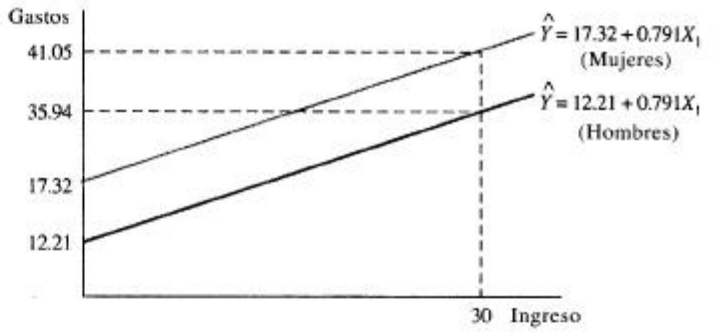
Los valores p aparecen en paréntesis

El uso de la variable dummy para sexo, en realidad produce dos rectas de regresión: una para hombres y otra para mujeres. Estas rectas tienen la misma pendiente pero diferentes interceptos. En otras palabras, la ecuación da dos rectas de regresión paralelas que comienzan en puntos diferentes sobre el eje vertical. Debido a que se codificó 0 para masculino, la ecuación se vuelve:

$$\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + b_2X_2 \\ &= 12.21 + 0.791X_1 + 5.11(0) \\ &= 12.21 + 0.791X_1\end{aligned}$$

para hombres. Esta recta tiene un intercepto de 12.21 y una pendiente de 0.791, y se muestra en la figura 12.6.

Figura 12.6
Rectas de regresión para los gastos



Para las mujeres, el valor codificado de 1 produce:

$$\begin{aligned}\hat{Y} &= 12.21 + 0.791X_1 + 5.11(1) \\ &= 17.32 + 0.791X_1\end{aligned}$$

Esta segunda recta tiene la misma pendiente que las rectas para hombres, pero tiene un intercepto de 17.32. Debido a que $X_2 = 1$ para mujeres, se determinó que el intercepto era $b_0 + b_2 = 12.21 + 5.11 = 17.32$.

Esto significa que para cualquier nivel de ingreso dado, los clientes mujeres gastan en promedio US$5.11 más que los hombres. Sea el ingreso igual a 30 (US$30,000). Entonces para las mujeres:

$$\begin{aligned}\hat{Y} &= 12.21 + 0.791(30) + 5.11(1) \\ &= 41.05\end{aligned}$$

y para los hombres:

$$\begin{aligned}\hat{Y} &= 12.21 + 0.791(30) + 5.11(0) \\ &= 35.94\end{aligned}$$

La diferencia de US$5.11 ocurre porque el valor codificado de 0 para los hombres cancela el coeficiente b_2 de 5.11, mientras que el valor codificado de 1 para mujeres termina en la adición de 5.11 a la ecuación.

Si se hubiera codificado la variable dummy asignando un 1 a una observación masculina y un 0 a una observación femenina, los resultados finales serían iguales. Un computador demuestra que la ecuación inicial es:

$$\hat{Y} = 17.32 + 0.791X_1 - 5.11X_2$$

Para las mujeres se tiene que:

$$\begin{aligned}\hat{Y} &= 17.32 + 0.791X_1 + 5.11(0) \\ &= 17.32 + 0.791X_1\end{aligned}$$

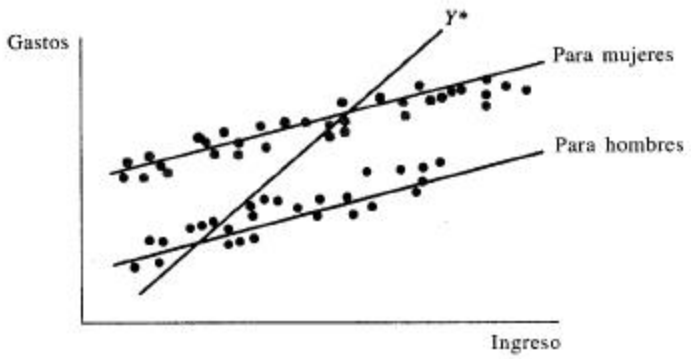
y para los hombres:

$$\begin{aligned}\hat{Y} &= 17.32 + 0.791X_1 - 5.11(1) \\ &= 12.21 + 0.791X_1\end{aligned}$$

Codificar la variable dummy de cualquier forma da los mismos resultados.

Si se colocaran los datos en un diagrama de dispersión, pueden aparecer como en la figura 12.7. En caso extremo, podrían aparecer dos diagramas casi totalmente independientes, uno para las observaciones masculinas y otro para las femeninas. Si se ignorara la variable dummy y se ajustara sólo una recta, su pendiente sería mucho más empinada que las otras dos, como la recta identificada con Y^*. El efecto atribuido sólo al ingreso por la recta de regresión debería adscribirse parcialmente al sexo.

Figura 12.7
Diagrama de dispersión para los gastos



Si una variable dummy tiene más de dos posibles respuestas, no se puede codificar como 0, 1, 2, 3 y así sucesivamente. Una variable con r posibles respuestas se expandirá para comprender un total de r - 1 variables. Por ejemplo, se puede incluir la tercera variable en el modelo para estudiar el efecto del estado civil sobre los gastos. Las posibles respuestas pueden incluir, casado(a), soltero(a), divorciado(a) y viudo(a). Además de X_1 para ingreso y X_2 para sexo, estas cuatro posibles respuestas requieren tres variables adicionales, X_3, X_4 y X_5, para codificar los datos sobre estado civil. Esto se logra ingresando sólo un 0 o un 1 por cada variable de la siguiente manera:

X_3 = 1 si es casado(a)
= 0 si no es casado(a)
X_4 = 1 si es soltero(a)
= 0 si no es soltero(a)
X_5 = 1 si es divorciado(a)
= 0 si no es divorciado(a)

No es necesaria ninguna entrada para viudo, porque si $X_3 = X_4 = X_5 = 0$, el proceso de eliminación revela la observación de ser viudo.

Se asume que 0 está codificado para hombres y 1 para mujeres en X_2. Las tres observaciones (OBS) que aparecen aquí para 1) un hombre casado con gastos de 30 e ingreso de 40, 2) una mujer divorciada con gastos de 35 e ingreso de 38, y 3) un hombre viudo con gastos de 20 e ingresos de 45.

OBS	Y	X_1	X_2	X_3	X_4	X_5
1	30	40	0	1	0	0
2	35	38	1	0	0	1
3	20	45	0	0	0	0

Por ejemplo, en la primera observación, X_2 sería 0, ya que la observación es hombre, y X_3 es 1, mientras que tanto X_4 como X_5 son 0 porque la observación es casado.

Ejercicios de la sección

21. Una empresa carbonífera desea configurar un modelo de regresión para predecir la producción (Y) que comprende como variables explicativas horas de trabajo (X_1) y si ocurrió un paro durante el período en estudio (X_2). Diseñe el modelo y explique.

22. Dado el modelo del problema anterior, ¿b_2 debería ser positivo o negativo? Explique.

23. Diga qué valores asignaría usted a las variables dummy para medir la raza de una persona si las categorías incluidas son 1) blanco, 2) negro, 3) asiático y 4) otra.

24. Los estudiantes de la Escuela Cosmopolita de Cosméticos (*Cosmopolitan School of Cosmetics*) aprenden a codificar datos sobre el color del cabello como 1 si es rubio, 2 si es pelirrojo, y 3 si es de otro color. Haga comentarios. ¿Qué aconsejaría usted?

25. El gerente de una empresa contable local creó un modelo de regresión para el tiempo que se toma una auditoría. El modelo era

$$\hat{Y} = 17 - 1.41X_1 + 1.73X_2$$

en donde: $\hat{Y}$ es el tiempo en horas

X_1 son los años de experiencia del auditor

X_2 si el auditor es un CPT (Contador público titulado) : 0 si lo es, 1 si no lo es.

a. Interprete el coeficiente para X_2.

b. ¿Esperaría usted que b_2 sea positiva? Explique

c. ¿Si el auditor tiene siete años de experiencia y es un CPT ¿cuánto tiempo le tomaría realizar la auditoría?

d. Si otro auditor tiene también siete años de experiencia pero no es CPT, ¿cuánto le tomaría terminar la auditoría según el modelo?

26. Si la variable dummy del ejercicio 25 fuera 1 si es CPT, 0 si no es CPT, ¿cuál esperaría que fuera el signo de b_2? Explique.

27. Un representante de mercadeo establece una ecuación de regresión para las unidades vendidas con base en la población del distrito de ventas y si el distrito tiene oficina principal a la cual deba reportar. El modelo es:

$$\hat{Y} = 78.12 + 1.01X_1 - 17.2X_2$$

en donde $\hat{Y}$ es unidad vendida

X_1 es la población en miles

X_2 es 0 si el distrito tiene oficina, 1 si no la tiene

a. Interprete $b_2 = -17.2$

b. ¿Cómo compararía las pendientes y los coeficientes de dos rectas de regresión suministradas por este modelo? Calcule y compare las dos ecuaciones de regresión.

c. Elabore una gráfica.

28. Considerando el problema anterior, si la población es 17,000 en un distrito que tiene una oficina y 17,000 en un distrito sin oficina, ¿cuál sería el número de unidades vendidas en cada uno? Realice una gráfica para ilustrar.

29. Los estudios han demostrado que en estados con regulaciones más liberales respecto al recibo de compensación por desempleo, las tasas de desempleo son mayores. Si un modelo de regresión para las tasas de desempleo incorpora una variable dummy, codificada con 1, si las regulaciones son liberales y con 0 si es de otro modo, ¿el coeficiente sería mayor que o menor que cero de acuerdo con estos estudios? Explique.

12.8 El caso curvilineal

A través de la discusión se ha asumido que la relación entre X y Y puede expresarse como una línea recta. Es decir, la relación es lineal. Sin embargo, este no es siempre el caso. Se puede encontrar que un modelo curvilineal (no lineal) puede proporcionar un mejor ajuste. Se supone que en el esfuerzo por predecir las declaraciones de impuestos con base en la población, Sam Jorden, alcalde de Plattsburg, recolecta los datos que se observaron en la tabla 12.5. Los datos, tanto para impuestos como para poblaciones, están en millones. Un diagrama de estos datos se encuentra en la figura 12.8 y sugiere que el modelo curvilineal es necesario. No parece que una línea recta produzca un ajuste adecuado.

Un polinomio de grado k

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon \qquad [12.15]$$

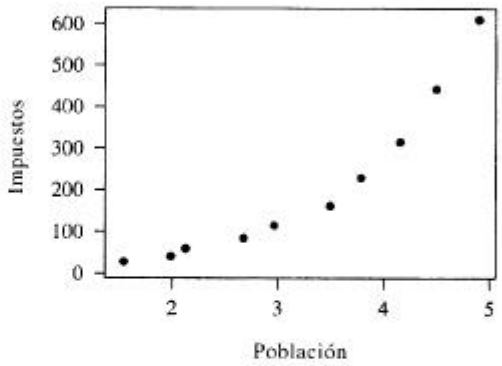
Tabla 12.5
Datos del alcalde Jorden sobre impuestos y población (en millones)

Impuestos	Población
85	2.68
118	2.98
164	3.50
228	3.79
31	1.57
43	2.01
61	2.15
611	4.90
316	4.16
444	4.50

Como se explicó en el capítulo anterior, en un modelo de regresión simple, el cambio en Y es constante. A medida que cambia X, Y cambia en un monto dado. En un modelo curvilineal, a medida que X cambia, Y cambia en una cantidad diferente. La figura 12.8 muestra que a medida que X se incrementa, Y aumenta la tasa de *incremento*.

Tales modelos curvilineales con frecuencia tienen buen ajuste utilizando una función polinómica de forma general.

Figura 12.8
Relación curvilineal para declaraciones de impuestos y población



Se dice que la fórmula (12.15) es un polinomio de grado k debido a que es la potencia más alta de cualquier variable explicativa. El modelo del alcalde Jorden puede tener mejor ajuste utilizando un polinomio de grado 2, o un polinomio de segundo orden, como:

Forma cuadrática del polinomio

$$\hat{Y} = b_0 + b_1X + b_2X^2 \qquad [12.16]$$

el cual es la forma cuadrática del modelo en el cual la segunda variable explicativa es simplemente el cuadrado del primero. En el caso del alcalde tenemos:

$$\hat{T} = b_0 + b_1POP + b_2(POP)^2$$

en donde T es impuestos y POP es población.

Se comparan los resultados de este modelo con los obtenidos si se estima un modelo lineal simple. La impresión en Minitab para el modelo simple en el cual se hace regresión de los impuestos (taxes) sobre la población (POP) y se muestra en la pantalla 12.7. Se nota que el valor R^2 es un respetable 86.1% con $\bar{R}^2$ del 84.3% y un error estándar de 76.38. Todo el modelo es

$$\hat{T} = -302.39 + 158.96POP$$

Pantalla 12.7 Un modelo lineal

Regression Analysis (Análisis de regresión)

The regression equation is
TAXES = -302 + 159 POP

Predictor	Coef	Stdev	t-ratio	p
Constant	-302.39	76.75	-3.94	0.004
POP	158.96	22.60	7.04	0.000

s = 76.38 R-sq = 86.1% R-sq (adj) = 84.3%

Taxes = Impuestos
POP = Población
Stdev = Desviación estándar
t-ratio = razón t

Si se utiliza el modelo cuadrático, como se observa en la impresión en Minitab en la pantalla 12.8, tanto $\bar{R}^2$ como el error estándar mejoran al 98.7% y 22.20 respectivamente. El modelo se vuelve:

$$\hat{T} = 325.36 - 277.98POP + 67.692(POP)^2$$

Obviamente, el modelo cuadrático proporciona el mejor ajuste.

Pantalla 12.8 Un ajuste cuadrático

Regression Analysis (Análisis de regresión)

```
The regression equation is
TAXES = 325 - 278 POP + 67.7 POPSQ

Predictor        Coef      Stdev    t-ratio       p
Constant       325.36      70.63       4.61   0.000
POP           -277.98      47.10      -5.90   0.000
POPSQ          67.692      7.226       9.37   0.000

s = 22.20        R-sq = 99.0%        R-sq (adj) = 98.7%
```

TAXES = Impuestos
POP = Población
t-ratio = razón t
Stdev = Desviación estándar

Un método alternativo para los modelos curvilineales puede lograrse mediante la transformación de los datos de alguna manera. Un método común de transformación implica el uso de logaritmos. Esta transformación logarítmica puede hacer que los datos sean *lineales en el logaritmo*. La tabla 12.6 muestra los datos originales del alcalde Jorden en las primeras dos columnas y sus logaritmos naturales en las últimas dos columnas. Entonces el alcalde simplemente hace regresión del logaritmo de los impuestos sobre el logaritmo de la población, como se observa en la impresión en Minitab en la pantalla 12.9. Se nota el mejoramiento en el error estándar de sólo 0.1680 y $\bar{R}^2$ de 82.2%, respectivamente. El modelo es

$$LOGTX = 2.0302 + 2.6147(LOGPOP)$$

Tabla 12.6
Transformación logarítmica

Impuestos	*POP*	*LOGTX*	*LOGPOP*
85	2.68	4.44265	0.98582
118	2.98	4.77068	1.09102
164	3.50	5.09987	1.25276
228	3.79	5.42935	1.33237
31	1.57	3.43399	0.45108
43	2.01	3.76120	0.69813
61	2.15	4.11087	0.76547
611	4.90	6.41510	1.58924
316	4.16	5.75574	1.42552
444	4.50	6.09582	1.50408

LOGTX = Logaritmo de impuestos
LOGPOP = Logaritmo de la población
POP = Población

Pantalla 12.9 Transformación logarítmica

Regression Analysis (Análisis de regresión)

```
The regression equation is
LOGTX = 2.03 + 2.61 LOGPOP

Predictor      Coef     Stdev  t-ratio       p
Constant     2.0302    0.1724    11.78   0.000
LOGPOP       2.6147    0.1478    17.69   0.000

s = 0.1680       R-sq = 97.5%       R-sq (adj) = 82.2%
```

Entonces, si la población es 3.2, log $POP = 1.163$ y

$$\begin{aligned} logTX &= 2.0302 + 2.6147(1.163) \\ &= 5.071 \end{aligned}$$

Tomando el antilogaritmo de 5.071 da 159.33; o debido a que los datos estaban originalmente en millones, los recibos de impuestos del alcalde se estimarían en US$159,330,000.

Puede ser necesario experimentar con diferentes formas funcionales para determinar cuál proporciona el mejor ajuste. En la búsqueda del modelo óptimo, los resultados de diferentes modelos logarítmicos pueden compararse con los obtenidos utilizando funciones polinómicas. El uso de computadores hace más práctica esta comparación.

Sin embargo, los resultados de tales comparaciones pueden ser inconsistentes. Un modelo puede reportar un coeficiente de determinación más alto que otro (eso es bueno) mientras que lleva un error estándar de estimación mayor (eso es malo). La pregunta entonces sería ¿cuál modelo utilizar?

La respuesta depende, al menos en parte, del propósito para el cual está destinado el modelo. Si se desea utilizar el modelo para explicar los valores presentes de Y y comprender por qué se comportan como lo hacen, se utiliza el modelo con el coeficiente de determinación más alto. Es decir, si el propósito es explicar, entonces el modelo con el valor explicativo más alto es el que debería utilizarse.

Si, por otra parte, el propósito del modelo es predecir los valores futuros de Y, se utiliza el modelo con el error estándar de estimación más bajo. Si se desea predecir, se gozará de más éxito con el modelo que genera el menor error de predicción.

Sin embargo, tal experimentación debería mantenerse al mínimo. Se considera cuestionable, incluso no ético, experimentar de forma salvaje con un modelo y luego con el otro. Se debería saber desde el comienzo, dada la naturaleza del estudio investigativo, qué procedimiento se debe seguir. Con frecuencia se hace la analogía de que buscar ciegamente el mejor modelo es similar a disparar una flecha al objetivo y luego sacar el blanco del punto en donde cayó la flecha.

Ejercicios de la sección

30. Grafique los siguientes datos. Compare un modelo lineal con la forma cuadrática y proporcione una evaluación comparativa de cada uno. Utilizando el mejor modelo, pronostique Y si $X = 22$.

Y	X	Y	X
2170	31	731	18
2312	32	730	18
2877	36	815	19
7641	48	1408	25
2929	36	2768	35
		1297	24

31. Utilizando los datos del ejercicio 30, realice una transformación logarítmica y pruebe los resultados de la regresión. ¿Cuáles son sus observaciones?

32. El alcalde Jorden desea estimar las declaraciones de impuestos con base en la formación de nuevos negocios. Él recolecta los datos de 22 ciudades que se muestran a continuación y que él considera parecidas a la propia.

a. Grafique el diagrama de dispersión.

b. Compare un modelo lineal con una forma cuadrática. ¿Cuál parece dar el mejor ajuste?

c. Proyecte la declaración de impuestos si hay 68 negocios nuevos.

Ciudad	Nuevos negocios	Declaración de impuestos	Ciudad	Nuevos negocios	Declaración de impuestos
1	47	US$10,154,589	12	68	US$26,272,898
2	51	18,215,568	13	37	6,074,615
3	57	27,171,076	14	58	18,215,568
4	68	26,272,898	15	51	13,546,448
5	57	6,074,615	16	48	17,500,544
6	45	5,693,092	17	65	14,801,029
7	85	43,918,912	18	68	18,215,568
8	87	46,334,860	19	85	43,918,912
9	48	11,781,520	20	68	42,738,224
10	57	17,500,544	21	86	45,117,748
11	68	26,272,898	22	58	25,391,750

33. Utilizando los datos del ejercicio 32, calcule el modelo de regresión logarítmica. ¿Cómo se compara con el desarrollado en el problema anterior? Calcule los impuestos si hay 68 nuevos negocios creados.

Lista de fórmulas

[12.1]	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$	Modelo de regresión múltiple
[12.4]	$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - k - 1}}$	Error estándar de estimación
[12.5]	$R^2 = \frac{SCR}{SCT}$	Coeficiente de determinación múltiple
[12.8]	$\overline{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - k - 1}$	Coeficiente de determinación múltiple corregido
[12.12]	$t = \frac{b_i - \beta_i}{s_{b_i}}$	Prueba t para la significancia de β_i
[12.13]	$FIV(X_i) = \frac{1}{1 - R_i^2}$	Factor de inflación de la varianza para X_i.
[12.14]	$\text{Beta} = \frac{b_i}{s_Y / s_{X_i}}$	El coeficiente beta o estandarizado para X_i
[12.15]	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon$	Un polinomio de grado k
[12.16]	$\hat{Y} = b_0 + b_1 X + b_2 X^2$	Forma polinómica cuadrática

Ejercicios del capítulo

Nota: La mayoría de los problemas requieren de computador.

34. Un director administrativo está intentando desarrollar un sistema diseñado para identificar qué atributos personales son esenciales para avanzar gerencialmente. A quince empleados que han sido ascendidos recientemente se les practica una serie de pruebas para determinar (X_1) sus habilidades comunicativas; (X_2) la habilidad para relacionarse con otros y (X_3) la habilidad para tomar decisiones. A la clasificación del trabajo de cada empleado (Y) se le hace regresión sobre estas tres variables. Los datos originales son los siguientes:

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
80	50	72	18	69	39	73	19
75	51	74	19	68	40	71	20
84	42	79	22	87	55	80	30
62	42	71	17	92	48	83	33
92	59	85	25	82	45	80	20
75	45	73	17	74	45	75	18
63	48	75	16	80	61	75	20
				62	59	70	15

a. Haga el modelo de regresión. Evalúelo determinando si muestra una relación significante entre la variable dependiente y las tres variables independientes.

b. ¿Qué puede decirse sobre la significancia de cada X_i?

35. ¿A qué causa puede usted atribuir la insignificancia de X_1 y X_3 en el ejercicio 34? Obtenga la matriz de correlación para estas variables y pruebe cada par para hallar multicolinealidad. Sea $\alpha = 5\%$.

36. Compare sus resultados del ejercicio 35 con los obtenidos con base en el FIV.

37. ¿El director administrativo del ejercicio 36 debería utilizar este modelo para identificar las características que hicieron a un empleado elegible para ascenso?

38. Como un proyecto de clase, un equipo de estudiantes de mercadeo diseña un modelo que explica la renta para la vivienda estudiantil que hay cerca de la universidad. La renta está expresada en dólares, PC son los pies cuadrados que tiene el apartamento o casa, y DIST es la distancia en millas de la casa al campus.

Renta	PC	DIST	Renta	PC	DIST
220	900	3.2	400	1,290	1.5
250	1,100	2.2	450	1,370	0.5
310	1,250	1.0	500	1,400	0.5
420	1,300	0.5	550	1,550	0.3
350	1,275	1.5	450	1,200	0.5
510	1,500	0.5	320	1,275	1.5

a. Realice un modelo. ¿Es significante al nivel del 1%?

b. Evalúe la significancia de ambos coeficientes.

c. ¿Los signos son apropiados? Explique.

39. Evalúe el modelo del problema anterior. ¿Parece útil para predecir la renta? Explique.

40. ¿Existe evidencia de multicolinealidad en el modelo del problema anterior? ¿Invalida el modelo para predecir la renta? ¿Por qué sí o por qué no?

41. Del modelo desarrollado anteriormente sobre la renta estudiantil, puede usted concluir que la distancia del campus es una determinante más fuerte de la renta que el número de pies cuadrados? ¿Por qué sí o por qué no?

42. Si dos apartamentos tienen el mismo espacio, pero uno es dos millas más cercano al campus, cómo difiere esta renta de la de la vivienda más distante?

43. Para expandir su modelo sobre las rentas de los estudiantes, los especialistas en mercadeo del problema anterior diseñaron un índice de lujo en el cual los estudiantes clasifican los aspectos atractivos de un apartamento con base en comodidades disponibles, como piscinas, cancha de tenis, servicio doméstico y otros lujos a los cuales los estudiantes están tradicionalmente acostumbrados. Para las 12 observaciones anteriores, este índice midió 22,23, 35, 40, 32, 55, 36, 41, 51, 50, 48 y 29. Incorpore la variable a su modelo para explicar las rentas. Analice y explique por qué usted obtuvo esos resultados. ¿Su modelo es mejor con esta variable adicional? ¿Qué problema es probable que usted encuentre y qué cambio haría para corregirlo?

44. Haga el cambio que sugirió en el problema anterior y discuta los resultados.

45. En el pasado, muchos economistas han estudiado los patrones de gasto de los consumidores en la economía. Un estudio famoso realizado por Milton Friedman concluye que el consumo es una función de *ingreso permanente*, que se define como el nivel de ingreso promedio que el consumidor espera recibir en el futuro. La teoría de *hábito-persistencia* de T.M. Brown argumenta que el consumo es moldeado por el ingreso pico más reciente de un consumidor–el ingreso más alto recibido en el pasado inmediato.
Para combinar estas dos teorías, un economista recolectó datos sobre el consumo (CONS), ingreso permanente (PERM), e ingreso pico (PICO), y realizó el MCO para desarrollar un modelo. Dados estos datos, ¿Cómo es el modelo? (Todos los valores están en miles de dólares).

Consumo	Ingreso permanente	Pico	Consumo	Ingreso permanente	Pico
12	15	17	14	17	20
22	28	31	20	25	29
15	19	21	17	21	25
17	19	24	15	19	22
19	24	27	16	20	26

a. Evalúe el modelo.
b. ¿La multicolinealidad explicaría la insignificancia del pico? Explique

.46. Los datos que aparecen aquí se recolectaron para explicar los niveles salariales para los trabajadores en una planta local.

Salario (US$1,000)	Años de educación	Sexo
42.2	8	M
58.9	12	M
98,8	16	M
23.5	6	F
12.5	5	M
67.8	12	M
51.9	10	F
81.6	14	F
61.0	12	F

a. Calcule el modelo de regresión utilizando un computador.

b. ¿Existe evidencia de la discriminación de sexo en cuanto a los niveles salariales?

c. ¿La educación es útil para explicar el salario?

d. ¿Hay problemas de autocorrelación y heterosedasticidad?

47. Usted acaba de hacer un modelo de regresión de la retención de un empleado (en años) sobre la edad al momento de la contratación y el género, codificando la variable dummy de género con 1 si es hombre y 0 si es mujer. Los resultados fueron

$$\hat{Y} = 3.2 + 0.65\ EDAD - 1.3\ GÉNERO$$

a. ¿Cuál es la fórmula para hombre? ¿Para mujer?

b. Luego usted se da cuenta que usted quiso codificar 1 si era mujer y 0 si era hombre. ¿Cómo sería la ecuación?

c. ¿Ahora cuál es la fórmula para hombre? ¿Para mujer?

d. Utilizando la fórmula que se vio anteriormente, ¿cuál es el estimado de años de retención si el hombre es contratado a los 23 años de edad? ¿Cuál es utilizando su fórmula revisada?

Ejercicio de computador

Los estudios de finanzas han demostrado que el precio de una acción está directamente relacionado con el nivel de deuda de la empresa emisora y con la tasa de dividendos, pero está inversamente relacionado con el número de acciones en circulación. Ingrese al archivo STOCK de su disco de datos. PRICE (precio) y la tasa de dividendo, DIVRATE, está en dólares, DEBT (deuda), está en millones de dólares y el número de acciones en circulación, OUTSTD, está en millones de acciones. Utilizando PRICE como variable dependiente, evalúe el modelo. Proporcione todas las interpretaciones estadísticas pertinentes y las conclusiones. Prepare su reporte financiero final como se describe en el apéndice I.

PUESTA EN ESCENA

Su asignación como practicante de Griffen Associates, descrita en la sección escenario a comienzos de este capítulo, ilustró brevemente la necesidad de analizar el desempeño de varios fondos mutuos competitivos. Usted debe desarrollar modelos que analizarán los rendimientos de tres años y compararlos con los rendimientos de un año, utilizando como variables explicativas las tasas de rotación (el porcentaje de fondos comprados y vendidos durante cada período en cuestión), los activos totales iniciales en miles de millones de dólares al momento en que se abrió el fondo, y si el fondo tiene una carga por concepto de ventas. Esta última variable se codifica como 1 si tiene una carga, y 0 si no tiene.

Su supervisor en Griffen desea que usted prepare un reporte completo, incluyendo todo el análisis estadístico presentado en este capítulo. Usted debe especificar los modelos para los rendimientos de un año y de tres años para fondos con carga o sin carga. Deben practicarse pruebas de multicolinealidad junto con todo el análisis estadístico relevante. Utilizando los datos suministrados aquí, prepare su informe estadístico como se describió en el apéndice I.

Rendimiento de 3 años	Rendimiento de un año	Carga	Transferencias en 3 años	Transferencias en un año	Activos
5.6	0.1	0	112	58	220.00
4.7	1.9	1	95	62	158.00
4.5	2.6	1	241	65	227.25
4.8	2.0	1	87	61	242.40
5.7	3.5	0	98	57	287.85
4.1	-4.3	1	102	66	207.05
4.7	3.2	1	72	63	237.35
4.1	-4.1	1	96	65	207.05
5.2	2.2	0	78	59	262.60
3.7	2.1	1	118	87	186.85
6.2	5.3	0	98	47	313.10
6.6	11.0	0	87	41	333.30
5.2	0.3	0	117	61	262.60
5.5	-2.1	0	87	46	277.75
5.6	4.7	0	85	35	282.80

Del escenario a la vida real

En el capítulo 3 del ejercicio Del escenario a la vida real, se familiarizó con algunos sitios de información sobre inversionistas en fondos mutuos y con Vanguard Group of Funds en especial. Aquí se analizará una información más detallada que se reporta rutinariamente sobre los fondos y su desempeño. El análisis de la sección Puesta en escena de este capítulo, necesitó información sobre las cargas (costos), activos, y rendimientos de 1 y 3 años. ¿Esta información se encuentra disponible en los sitios web de las empresas de fondos mutuos? Se observan entonces tres empresas importantes de fondos mutuos.

En el Home Page de Vanguard Group (*www.vanguard.com*), seleccione el icono de "mutual funds" (fondos mutuos). Luego, haga clic en "Funds by category" (Fondos por categoría). En Growth funds (Fondos de desarrollo) seleccione "U.S. Growth Portfolio". Haga clic en los rótulos de las carpetas y haga notas sobre dónde obtener datos sobre cargas, activos y rendimientos de 1 y 3 años.

Haga lo mismo para Fidelity Funds (*www.fidelity.com*). En el Home Page vaya a "Fund information" (información sobre el fondo). En el área de búsqueda, digite "Emerging Growth" y haga clic en "go". Observe los datos disponibles.

Repita este proceso nuevamente para Dreyfus Funds (*www.dreyfus.com*). En el Home Page, seleccione "Mutual Funds" (Fondos mutuos) bajo el encabezado de Products (Productos), y luego haga clic en fondos "Growth" (de desarrollo). Observe la información de Emerging Leaders Fund para buscar la disponibilidad de estos mismos datos.