

CAPÍTULO 16



Análisis de regresión: construcción de modelos

CONTENIDO

LA ESTADÍSTICA
EN LA PRÁCTICA:
LA EMPRESA MONSANTO

16.1 EL MODELO LINEAL
GENERAL

- Modelado de relaciones curvilíneas
- Interacción
- Transformaciones a la variable dependiente
- Modelos no lineales que son intrínsecamente lineales

16.2 DETERMINACIÓN
DE CUÁNDO AGREGAR
O QUITAR VARIABLES

- Caso general
- Uso del valor-*p*

16.3 ANÁLISIS DE UN
PROBLEMA MAYOR

16.4 PROCEDIMIENTOS DE
ELECCIÓN DE VARIABLES

- Regresión por pasos
- Selección hacia adelante
- Eliminación hacia atrás
- Regresión de los mejores subconjuntos
- Elección final

16.5 MÉTODO DE REGRESIÓN
MÚLTIPLE PARA EL DISEÑO
DE EXPERIMENTOS

16.6 AUTOCORRELACIÓN
Y LA PRUEBA DE
DURBIN-WATSON

LA ESTADÍSTICA *(en)* LA PRÁCTICA

LA EMPRESA MONSANTO*

SAN LUIS MISSOURI

Las raíces de Monsanto se remontan a una inversión de \$500 de un empresario y a un polvoriento almacén a orillas del Mississippi, donde en 1901 John F. Queeney empezó con la fabricación de sacarina. En la actualidad, Monsanto es una de las mayores empresas químicas de Estados Unidos, produce más de mil productos que van desde productos químicos industriales hasta canchas sintéticas para deportes que se emplean en los estadios modernos. Monsanto es una empresa mundial que cuenta con fábricas, laboratorios, centros técnicos y operaciones de marketing en 65 países.

La división de nutrición química de Monsanto fabrica y comercializa un suplemento de metionina que se usa en alimento para ganado, cerdos y aves de corral. Como en la cría de aves de corral se trabaja con volúmenes altos y márgenes de ganancias reducidos, se necesitan alimentos rentables y con el mayor valor nutricional posible. El alimento de composición óptima es el que produce un crecimiento rápido y un alto peso corporal final con una determinada ingesta de alimento. La industria química trabaja en estrecha colaboración con los criadores de aves de corral para optimizar los alimentos. Por último, el éxito depende de mantener bajo el costo de las aves de corral en comparación con el costo de la carne de res y de otros productos de carne.

Para modelar la relación entre peso corporal y cantidad de metionina x adicionada al alimento para aves de corral, los investigadores de Monsanto emplearon el análisis de regresión. Al principio se obtuvo la siguiente ecuación de regresión lineal estimada.

$$\hat{y} = 0.21 + 0.42x$$

Esta ecuación estimada de regresión resultó estadísticamente significativa; sin embargo, de acuerdo con el análisis de residuales una relación curvilinea parecía ser un modelo más adecuado para la relación entre peso corporal y metionina.

*Los autores agradecen a James R. Ryland y a Robert M. Schisla, especialistas en investigación de Monsanto Nutrition Chemical Division, por proporcionar este artículo para *La estadística en la práctica*.



Los investigadores de Monsanto emplearon el análisis de regresión para obtener un alimento de composición óptima para los criadores de aves de corral.

© PhotoDisc/Getty Images.

Al continuar con las investigaciones encontraron que aunque cantidades pequeñas de metionina tendían a hacer aumentar el peso corporal, en cierto punto el peso corporal se estabilizaba y un aumento en la cantidad de metionina tenía poco o ningún efecto. Peor aún, cuando la cantidad de metionina era mayor que el requerimiento nutrimental, el peso corporal tendía a disminuir. Para modelar la relación curvilinea entre peso corporal y metionina se empleó la siguiente ecuación estimada de regresión múltiple.

$$\hat{y} = -1.89 + 1.32x - 0.506x^2$$

Al aplicar la regresión, los investigadores de Monsanto pudieron encontrar la cantidad óptima de metionina que debía usarse en los productos alimenticios para aves de corral.

En este capítulo el estudio del análisis de regresión se ampliará a la obtención de modelos curvilineos como el usado por los investigadores de Monsanto. Se describirán, además, diversas herramientas que sirven para determinar cuáles son las variables independientes que conducen a una mejor ecuación estimada de regresión.

La construcción de modelos es el proceso que consiste en obtener una ecuación estimada de regresión que describa la relación entre una variable dependiente y una o varias variables independientes. Lo más importante en la construcción de un modelo es hallar la forma funcional adecuada para la relación y seleccionar las variables independientes que se deban incluir en el modelo. En la sección 16.1 se presenta el concepto de modelo lineal general que establece el marco para la construcción de modelos. En la sección 16.2, en la que se presentan las bases para procedimientos más sofisticados que emplean paquetes de software, se enseña un procedimiento

general para determinar cuándo agregar o eliminar variables independientes. En la sección 16.3 se considera un problema más grande de regresión en el que intervienen ocho variables independientes y 25 observaciones; este problema sirve para ilustrar los procedimientos de selección de variables presentados en la sección 16.4, que comprenden la regresión por pasos, el procedimiento de selección hacia adelante, el procedimiento de eliminación hacia atrás y el mejor subconjunto de regresión. En la sección 16.5 se muestra cómo el análisis de regresión múltiple proporciona otro método para la solución de problemas de diseño experimental, y en la sección 16.6 se muestra el uso de la prueba Durbin-Watson para detectar correlación serial o autocorrelación.

16.1 El modelo lineal general

Suponga que se obtienen los datos de una variable independiente y y de k variables independientes x_1, x_2, \dots, x_k . El objetivo es obtener, con estos datos, la ecuación estimada de regresión que mejor exprese la relación entre la variable dependiente y las independientes. Como marco general para el desarrollo de relaciones más complejas entre las variables independientes, se introduce el concepto de **modelo lineal general** con p variables independientes.

Si el modelo de regresión se puede expresar en la forma de la ecuación (16.1), entonces se aplica el procedimiento estándar de regresión múltiple descrito en el capítulo 15.

TABLA 16.1

DATOS DEL EJEMPLO
DE REYNOLDS

Antigüedad en meses	Balanzas vendidas
41	375
106	296
76	317
10	376
22	162
12	150
85	367
111	308
40	189
51	235
9	83
12	112
6	67
56	325
19	189

MODELO LINEAL GENERAL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \quad (16.1)$$

En la ecuación (16.1), cada una de las variables independientes z_j (donde $j = 1, 2, \dots, p$) es función de x_1, x_2, \dots, x_k (las variables para las que se obtuvieron los datos). En algunos casos cada z_j puede ser función de una sola variable x . El caso más sencillo es cuando sólo se obtienen datos de una variable x_1 y se quiere estimar y por medio de una relación lineal. En este caso $z_1 = x_1$ y la ecuación (16.1) se convierte en

$$y = \beta_0 + \beta_1 x_1 + \epsilon \quad (16.2)$$

La ecuación (16.2) es el modelo de regresión lineal simple presentado en el capítulo 14, con la única diferencia de que a la variable independiente se le ha llamado x_1 en lugar de x . En la literatura sobre modelos estadísticos, a este modelo se le llama *modelo simple de primer orden con una variable predictora*.

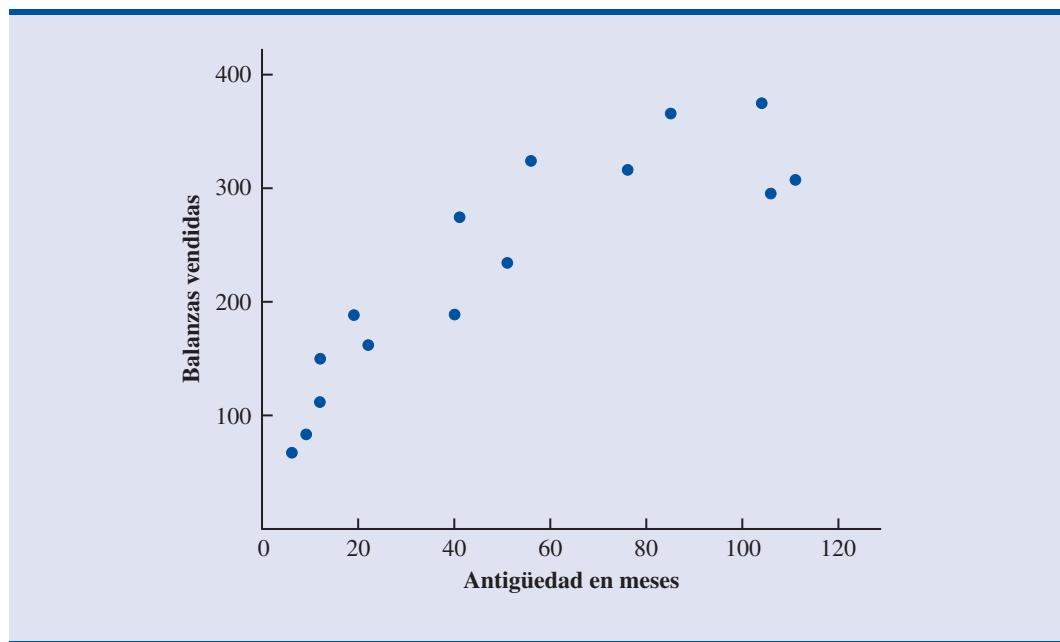
Modelado de relaciones curvilíneas

Con la ecuación (16.1) también se pueden modelar relaciones más complejas. Para ilustrar esto se verá un problema que se le presentó a la empresa Reynolds, Inc., fabricante de balanzas industriales y de equipo para laboratorio. Los gerentes de Reynolds desean investigar la relación que existe entre la antigüedad de sus vendedores y el número de balanzas electrónicas para laboratorio que venden. En la tabla 16.1 se presenta el número de balanzas vendidas por cada uno de 15 vendedores elegidos aleatoriamente y la antigüedad que tiene cada uno de ellos en la empresa. En la figura 16.1 se presenta el diagrama de dispersión de estos datos. En el diagrama de dispersión se observa que es posible que exista una relación curvilínea entre antigüedad de un empleado y número de balanzas que vende. Antes de considerar cómo obtener una relación curvilínea para este problema de Reynolds, se analizarán los resultados de Minitab que se presentan en la figura 16.2 y que corresponden a un modelo simple de primer orden; la ecuación estimada de regresión es

$$\text{Sales (Ventas)} = 111 + 2.38 \text{ Months (Meses)}$$

donde

$$\begin{aligned} \text{Sales (Ventas)} &= \text{número de balanzas electrónicas para laboratorio vendidas} \\ \text{Months (Meses)} &= \text{antigüedad del vendedor, en meses} \end{aligned}$$

FIGURA 16.1 DIAGRAMA DE DISPERSIÓN PARA EL EJEMPLO DE REYNOLDS

La figura 16.3 es la gráfica de residuales estandarizados correspondiente. Aunque los resultados de Minitab indican que la relación sí es significativa (valor-*p* = 0.000) y que una relación lineal explica un porcentaje grande de la variabilidad en las ventas (*R-sq* = 78.1%), la gráfica de residuales estandarizados sugiere que se necesita una relación curvilinea.

Para obtener una relación curvilinea, en la ecuación (16.1) se hace $z_1 = x_1$ y $z_2 = x_1^2$, así resulta el modelo

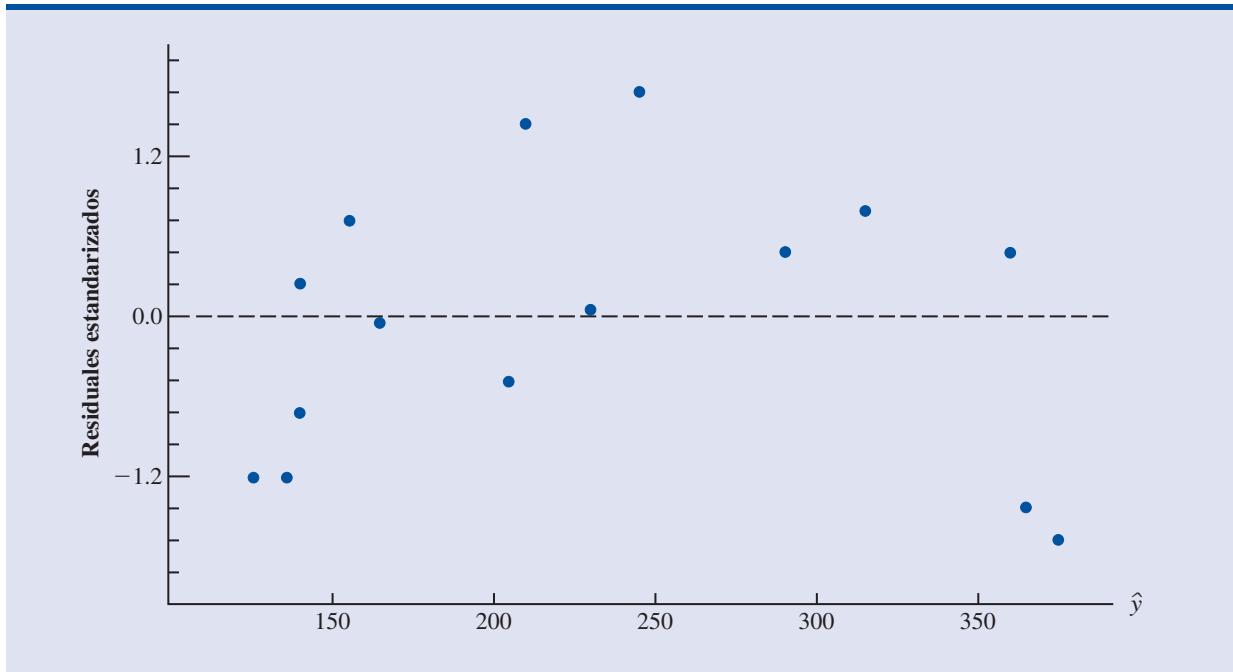
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \quad (16.3)$$

A este modelo se le llama *modelo de segundo orden con una variable predictora*. Para proporcionar la ecuación estimada de regresión correspondiente a este modelo de segundo orden, Minitab

FIGURA 16.2 RESULTADOS DE MINITAB PARA EL EJEMPLO DE REYNOLDS:
MODELO DE PRIMER ORDEN.

The regression equation is Sales = 111 + 2.38 Months
Predictor Coef SE Coef T p
Constant 111.23 21.63 5.14 0.000
Months 2.3768 0.3489 6.81 0.000
S = 49.52 R-sq = 78.1% R-sq(adj) = 76.4%
Analysis of Variance
SOURCE DF SS MS F p
Regression 1 113783 113783 46.41 0.000
Residual Error 13 31874 2452
Total 14 145657

FIGURA 16.3 GRÁFICA DE RESIDUALES ESTANDARIZADOS DEL EJEMPLO DE REYNOLDS: MODELO DE PRIMER ORDEN



necesita los datos originales de la tabla 16.1, así como los datos que corresponden a la segunda variable dependiente que se agrega, que es el cuadrado de los meses de antigüedad que tiene el empleado en la empresa. En la figura 16.4 se presentan los resultados de Minitab correspondientes al modelo de segundo orden; la ecuación estimada de regresión es

Los datos de la variable independiente MonthsSq se obtienen al elevar al cuadrado los valores de Months.

$$\text{Sales (Ventas)} = 45.3 + 6.34 \text{ Months (Meses)} - 0.0345 \text{ MonthsSq (Meses al cuadrado)}$$

donde

MonthsSq = cuadrado del número de meses que ha trabajado el vendedor en la empresa

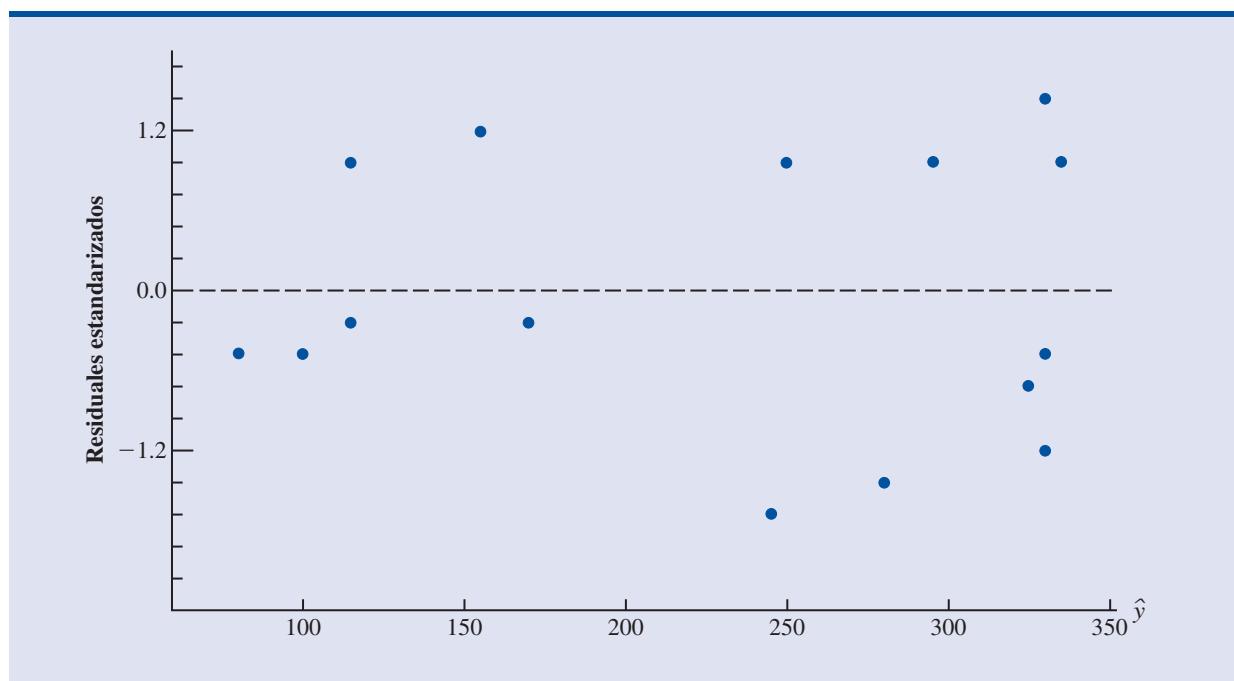
La figura 16.5 es la gráfica de residuales estandarizados correspondiente. En esta gráfica se observa que el patrón curvilíneo anterior ha desaparecido. Al emplear como nivel de significancia 0.05, los resultados de Minitab indican que el modelo general es significativo (el valor-*p* para la prueba *F* es 0.000); observe también que el valor-*p* correspondiente al cociente-*t* de MonthsSq (valor-*p* = 0.002) es menor que 0.05, por lo que se puede concluir que la adición de MonthsSq al modelo es significativa. Como el valor de R-sq(adj) es 88.6%, se puede estar satisfecho con el ajuste que proporciona esta ecuación estimada de regresión. Lo más importante, sin embargo, es ver lo fácil que es tratar las relaciones curvilíneas en el análisis de regresión.

Es claro que por medio de la ecuación (16.1) se pueden modelar muchos tipos de relaciones. Las técnicas de regresión con las que se ha estado trabajando son técnicas que definitivamente no están limitadas a relaciones lineales. En el análisis de regresión múltiple, la palabra *lineal* en el término “modelo lineal general” se refiere únicamente al hecho de que $\beta_0, \beta_1, \dots, \beta_p$, tienen, todos, exponente 1; no implica que la relación entre y y las x_i sea lineal. Es más, en esta sección se ha visto un ejemplo del uso de la ecuación (16.1) para modelar una relación curvilínea.

FIGURA 16.4 RESULTADOS DE MINITAB PARA EL EJEMPLO DE REYNOLDS:
MODELO DE SEGUNDO ORDEN

The regression equation is $Sales = 45.3 + 6.34 \text{ Months} - 0.0345 \text{ MonthsSq}$					
Predictor	Coef	SE Coef	T	p	
Constant	45.35	22.77	1.99	0.070	
Months	6.345	1.058	6.00	0.000	
MonthsSq	-0.034486	0.008948	-3.85	0.002	
$S = 34.45 \quad R-sq = 90.2\% \quad R-sq(\text{adj}) = 88.6\%$					
Analysis of Variance					
SOURCE	DF	SS	MS	F	p
Regression	2	131413	65707	55.36	0.000
Residual Error	12	14244	1187		
Total	14	145657			

FIGURA 16.5 GRÁFICA DE RESIDUALES ESTANDARIZADOS PARA EL EJEMPLO DE REYNOLDS:
MODELO DE SEGUNDO ORDEN



Interacción

Si el conjunto de datos original consta de observaciones para y y para dos variables independientes x_1 y x_2 , y en la ecuación (16.1) se pone $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$ y $z_5 = x_1x_2$ se puede obtener un modelo de segundo orden con dos variables predictoras. El modelo resultante es

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4x_2^2 + \beta_5x_1x_2 + \epsilon \quad (16.4)$$

En este modelo de segundo orden, la variable $z_5 = x_1x_2$ se agrega para tomar en cuenta el posible efecto que pueda tener la acción conjunta de las dos variables. A este tipo de efecto se le llama **interacción**.

Para ver un ejemplo de interacción y de lo que significa, se revisará el estudio de regresión realizado por Tyler Personal Care para un nuevo champú. Se pensó que los dos factores que tenían más influencia sobre las ventas eran el precio de venta por unidad y gasto en publicidad. Para investigar el efecto de estas dos variables sobre las ventas, se formaron parejas con los precios \$2.00, \$2.50 y \$3.00 y los gastos en publicidad \$50 000 y \$100 000 en 24 mercados de prueba. En la tabla 16.2 se presentan las unidades vendidas (en miles).

En la tabla 16.3 se resumen estos datos. Observe que las ventas medias muestrales correspondientes al precio \$2.00 y al gasto en publicidad \$50 000 son 461 000 unidades y que las ventas medias muestrales correspondientes al precio \$2.00 y al gasto en publicidad \$100 000 son 808 000 unidades. Por tanto, cuando el precio se mantiene constante en \$2.00, la diferencia en ventas medias entre los gastos en publicidad de \$50 000 y de \$100 000 es $808\ 000 - 461\ 000 = 347\ 000$ unidades. Cuando el precio del producto es \$2.50, la diferencia en las ventas medias es $646\ 000 - 364\ 000 = 282\ 000$ unidades. Por último, cuando el precio es \$3.00 la diferencia en las ventas medias es $375\ 000 - 332\ 000 = 43\ 000$ unidades. Es claro que la diferencia en ventas medias entre gastos en publicidad de \$50 000 y de \$100 000 depende del precio del producto. En otras palabras, a precios de venta más elevados, el efecto del aumento en los gastos en publicidad disminuye. Estas observaciones hacen evidente la interacción entre las variables precio y gasto en publicidad.

Para proporcionar otra perspectiva de la interacción, en la figura 16.6 se muestran las ventas medias muestrales correspondientes a las seis combinaciones precio-gasto en publicidad. En esta gráfica también se muestra que el efecto de los gastos en publicidad sobre las ventas medias de-

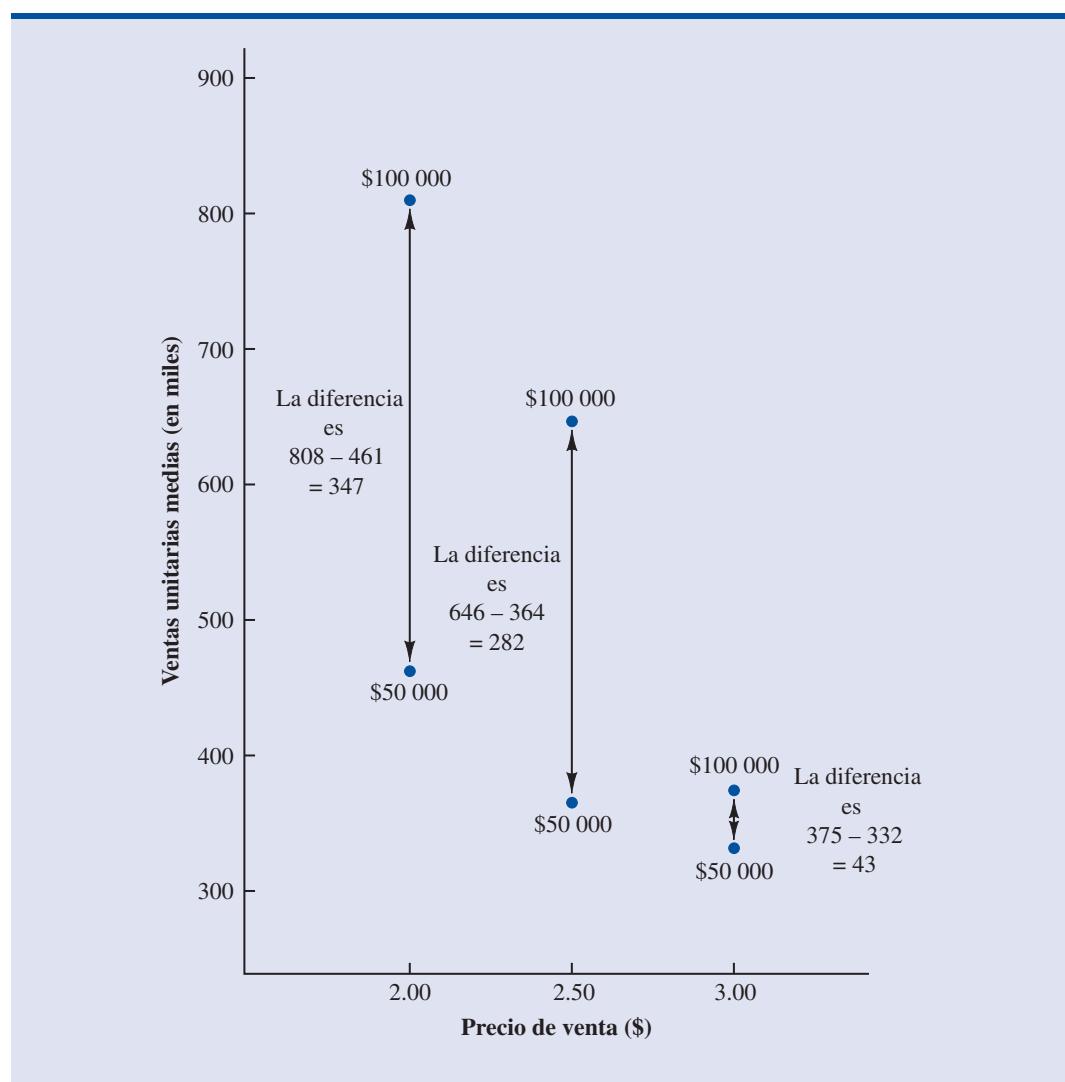
TABLA 16.2 DATOS DEL EJEMPLO TYLER PERSONAL CARE

Precio	Gastos en publicidad (\$ miles)	Ventas (en miles)	Precio	Gasto en publicidad (\$ miles)	Ventas (en miles)
\$2.00	50	478	\$2.00	100	810
\$2.50	50	373	\$2.50	100	653
\$3.00	50	335	\$3.00	100	345
\$2.00	50	473	\$2.00	100	832
\$2.50	50	358	\$2.50	100	641
\$3.00	50	329	\$3.00	100	372
\$2.00	50	456	\$2.00	100	800
\$2.50	50	360	\$2.50	100	620
\$3.00	50	322	\$3.00	100	390
\$2.00	50	437	\$2.00	100	790
\$2.50	50	365	\$2.50	100	670
\$3.00	50	342	\$3.00	100	393

TABLA 16.3 VENTAS UNITARIAS MEDIAS EN EL EJEMPLO DE TYLER PERSONAL CARE

Gasto en publicidad	Precio		
	\$2.00	\$2.50	\$3.00
\$50 000	461	364	332
\$100 000	808	646	375

Ventas medias de 808 000 unidades cuando el precio = \$2.00 y el gasto en publicidad = \$100 000

FIGURA 16.6 VENTAS UNITARIAS MEDIAS (EN MILES) COMO FUNCIÓN DEL PRECIO DE VENTA Y DEL GASTO EN PUBLICIDAD

pende del precio del producto; una vez más se ve el efecto de la interacción. Cuando hay interacción entre dos variables, no se puede estudiar el efecto de una de las variables sobre la respuesta y independientemente de la otra variable. En otras palabras, sólo es posible obtener conclusiones claras si se considera el efecto conjunto que tienen las dos variables sobre la respuesta.

Para tomar en cuenta el efecto de la interacción, se usará el siguiente modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad (16.5)$$

donde

y = ventas unitarias (en miles)

x_1 = precio (\$)

x_2 = gastos en publicidad (\$ miles)

Observe que la ecuación (16.5) refleja la creencia de Tyler de que el número de unidades vendidas depende linealmente del precio de venta y de los gastos en publicidad (representados por los términos $\beta_1 x_1$ y $\beta_2 x_2$) y de que existe interacción entre las dos variables (representada por el término $\beta_3 x_1 x_2$).

Para obtener una ecuación estimada de regresión, se empleó un modelo lineal general con tres variables independientes (z_1 , z_2 y z_3).

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon \quad (16.6)$$

donde

$z_1 = x_1$

$z_2 = x_2$

$z_3 = x_1 x_2$

En la figura 16.7 se presentan los resultados de Minitab correspondientes al modelo de interacción para el ejemplo de Tyler Personal Care. La ecuación estimada de regresión que se obtiene es

$$\begin{aligned} \text{Sales (Ventas)} &= -276 + 175 \text{ Price (Precio)} + \\ &19.7 \text{ AdvExp (GastPubl)} - 6.08 \text{ PriceAdv (PrecioPubl)} \end{aligned}$$

donde

Sales (Ventas) = ventas unitarias (miles)

Price (Precio) = precio del producto (\$)

AdvExp (GastPubl) = gastos en publicidad (\$ miles)

PriceAdv (PrecioPubl) = término de la interacción (precio multiplicado por publicidad)

Los datos de la variable independiente PriceAdv se obtienen multiplicando cada precio por el correspondiente valor de AdvExp.

Como el modelo es significativo (el valor- p en la prueba F es 0.000) y como el valor- p correspondiente a la prueba t para PriceAdv es 0.000, se concluye que la interacción es significativa dado el efecto lineal del precio del producto y de los gastos en publicidad. Por tanto, los resultados de la regresión indican que el efecto de gastos en publicidad sobre las ventas depende del precio.

Transformaciones a la variable dependiente

Al mostrar el uso del modelo lineal general para modelar diversas relaciones que puede haber entre las variables independientes y la variable dependiente, se ha concentrado la atención en transformaciones a una o varias de las variables independientes. Con frecuencia vale la pena utilizar

FIGURA 16.7 RESULTADOS DE MINITAB PARA EL EJEMPLO DE TYLER PERSONAL CARE

The regression equation is Sales = - 276 + 175 Price + 19.7 AdvExpen - 6.08 PriceAdv					
Predictor	Coef	SE Coef	T	p	
Constant	-275.8	112.8	-2.44	0.024	
Price	175.00	44.55	3.93	0.001	
Adver	19.680	1.427	13.79	0.000	
PriceAdv	-6.0800	0.5635	-10.79	0.000	
 $S = 28.17 \quad R-\text{sq} = 97.8\% \quad R-\text{sq}(\text{adj}) = 97.5\%$					
Analysis of Variance					
SOURCE	DF	SS	MS	F	p
Regression	3	709316	236439	297.87	0.000
Residual Error	20	15875	794		
Total	23	725191			

TABLA 16.4

RENDIMIENTO
EN MILLAS POR
GALÓN Y PESOS DE
12 AUTOMÓVILES

Pesos	Millas por galón
2289	28.7
2113	29.2
2180	34.2
2448	27.9
2026	33.3
2702	26.4
2657	23.9
2106	30.5
3226	18.1
3213	19.5
3607	14.3
2888	20.9

transformaciones a la variable dependiente y . Para dar un ejemplo de un caso en el que puede ser útil transformar la variable dependiente, considere los datos de la tabla 16.4, en las que se muestran rendimientos en millas por galón y pesos de 12 automóviles. El diagrama de dispersión de la figura 16.8 indica que entre estas dos variables existe una relación lineal negativa. Por tanto, se usa un modelo simple de primer orden para relacionar estas dos variables. En la figura 16.9 se presentan los resultados que proporciona Minitab; la ecuación estimada de regresión es

$$\text{MPG} = 56.1 - 0.0116 \text{ Weight}$$

donde

MPG = rendimiento en millas por galón

Weight (Peso) = peso del automóvil dado en libras

El modelo es significativo (el valor- p en la prueba F es 0.000) y el ajuste es muy bueno ($R-\text{sq} = 93.5\%$). Sin embargo, en la figura 16.9 se ve que la observación 3 ha sido identificada como una observación cuyo residual estandarizado es grande.

La figura 16.10 es la gráfica de los residuales estandarizados correspondientes al modelo de primer orden. Su forma no parece ser la de la banda horizontal que se esperaría observar si las suposiciones acerca del término del error fueran válidas. La variabilidad de los residuales parece aumentar a medida que aumenta el valor de \hat{y} . En otras palabras, se observa la forma de cuña que, como se dijo en los capítulos 14 y 15, indica una varianza que no es constante. Si las suposiciones para el modelo de esta prueba no parecen satisfacerse, entonces no se justifica sacar conclusiones acerca de la significancia estadística de la ecuación estimada de regresión que se obtiene.

El problema de una varianza no constante suele corregirse al transformar la variable dependiente a otra escala. Por ejemplo, si se trabaja con el logaritmo de la variable dependiente en lugar de la variable dependiente original, los valores de la variable dependiente se comprimirán (estarán más cercanos unos a otros) y con esto disminuirán los efectos de la varianza no constante. La mayor parte de los paquetes de software para estadística proporcionan la posibilidad de aplicar transformaciones logarítmicas mediante logaritmos base 10 (logaritmos comunes) o lo-

FIGURA 16.8 DIAGRAMA DE DISPERSIÓN DE LOS DATOS DEL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN

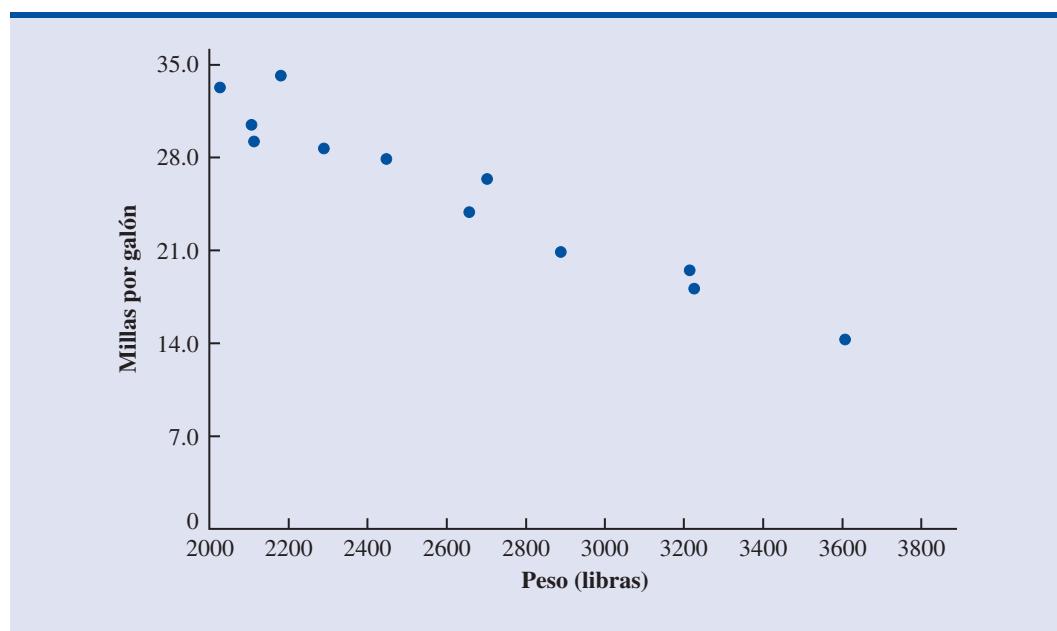


FIGURA 16.9 RESULTADOS DE MINITAB PARA EL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN

The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor	Coef	SE Coef	T	p
Constant	56.096	2.582	21.72	0.000
Weight	-0.0116436	0.0009677	-12.03	0.000

S = 1.671 R-sq = 93.5% R-sq(adj) = 92.9%

Analysis of Variance

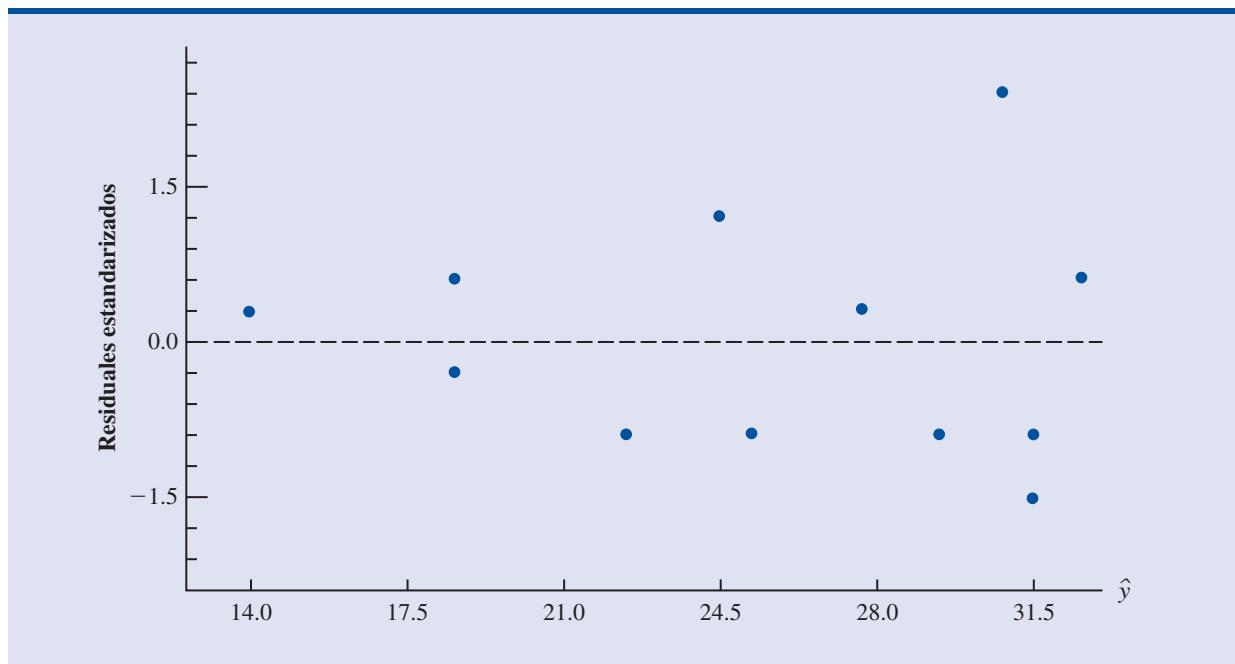
SOURCE	DF	SS	MS	F	p
Regression	1	403.98	403.98	144.76	0.000
Residual Error	10	27.91	2.79		
Total	11	431.88			

Unusual Observations

Obs	Weight	MPG	Fit	SE Fit	Residual	St Resid
3	2180	34.200	30.713	0.644	3.487	2.26R

R denotes an observation with a large standardized residual.

FIGURA 16.10 GRÁFICA DE LOS RESIDUALES ESTANDARIZADOS CORRESPONDIENTES AL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN



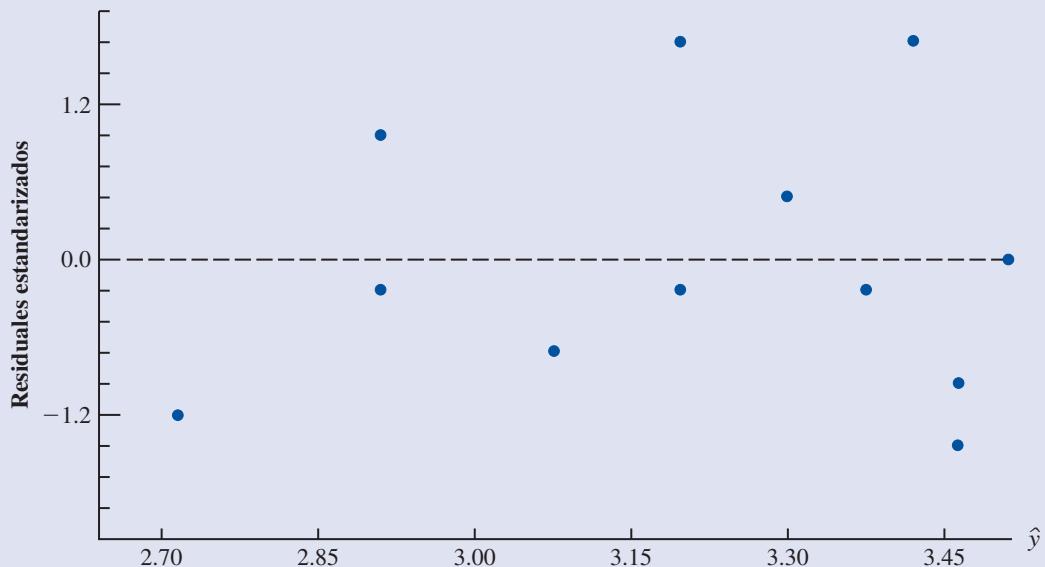
garitmos base $e = 2.71828\dots$ (logaritmos naturales). Aquí se empleará la transformación a logaritmos naturales de los datos de millas por galón y se desarrollará la ecuación estimada de regresión que relaciona el peso con el logaritmo natural de las millas por galón. En la figura 16.11 se muestra la ecuación de regresión que se obtiene al emplear el logaritmo natural de millas por galón como variable dependiente; esta ecuación aparece rotulada como LogeMPG. En la figura 16.12 se presenta la gráfica de los correspondientes residuales estandarizados.

Al observar la gráfica de residuales de la figura 16.12, se ve que la forma de cuña ha desaparecido. Además, ninguna de las observaciones ha sido identificada como una observación cuyo

FIGURA 16.11 RESULTADOS DE MINITAB PARA EL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN: TRANSFORMACIÓN LOGARÍTMICA

The regression equation is LogeMPG = 4.52 -0.000501 Weight
Predictor Coef SE Coef T p
Constant 4.52423 0.09932 45.55 0.000
Weight -0.00050110 0.00003722 -13.46 0.000
S = 0.06425 R-sq = 94.8% R-sq(adj) = 94.2%
Analysis of Variance
SOURCE DF SS MS F p
Regression 1 0.74822 0.74822 181.22 0.000
Residual Error 10 0.04129 0.00413
Total 11 0.78950

FIGURA 16.12 GRÁFICA DE RESIDUALES ESTANDARIZADOS DEL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN: TRANSFORMACIÓN LOGARÍTMICA



residual estandarizado sea grande. El modelo en el que se emplea como variable dependiente el logaritmo de las millas por galón es estadísticamente significativo y proporciona un ajuste excelente a los datos observados. Por tanto, se recomendará usar la ecuación estimada de regresión

$$\text{LogeMPG} = 4.52 - 0.000501 \text{ Weight (Peso)}$$

Para estimar el rendimiento en millas por galón de un automóvil que pese 2 500 libras, se obtiene primero una estimación del logaritmo del rendimiento de millas por galón.

$$\text{LogeMPG} = 4.52 - 0.000501(2500) = 3.2675$$

La estimación de las millas por galón se obtiene al hallar el número cuyo logaritmo natural es 3.2675. Al emplear una calculadora con función exponencial o elevar e a la potencia 3.2675, se obtienen 26.2 millas por galón.

Otro método para problemas con varianza no constante es usar como variable dependiente $1/y$, en lugar de y . A este tipo de transformación se le llama *transformación recíproca*. Por ejemplo, si la variable dependiente se mide en millas por galón, la transformación recíproca dará como resultado una nueva variable dependiente cuyas unidades serán $1/(\text{millas por galón})$ o galones por milla. No hay manera de determinar qué transformación funcionará mejor, si una transformación logarítmica o una transformación recíproca, si no es probándolas.

Modelos no lineales que son intrínsecamente lineales

A los modelos que tienen parámetros $(\beta_0, \beta_1, \dots, \beta_p)$ con exponentes distintos a 1 se les conoce como modelos no lineales. Sin embargo, en el caso de modelos exponenciales, se puede realizar una transformación de las variables que permita realizar el análisis de regresión mediante la

ecuación (16.1), el modelo lineal general. En el modelo general exponencial se tiene la siguiente ecuación de regresión

$$E(y) = \beta_0 \beta_1^x \quad (16.7)$$

Este modelo es adecuado cuando la variable dependiente y aumenta o disminuye en un porcentaje constante, en lugar de en una cantidad fija, a medida que x aumenta.

Por ejemplo, suponga que las ventas y de un producto se relacionan con los gastos en publicidad x (en miles de dólares) de acuerdo con el siguiente modelo exponencial.

$$E(y) = 500(1.2)^x$$

Por tanto, para $x = 1$, $E(y) = 500(1.2)^1 = 600$; para $x = 2$, $E(y) = 500(1.2)^2 = 720$; para $x = 3$, $E(y) = 500(1.2)^3 = 864$. Observe, que en este caso, $E(y)$ no aumenta en una cantidad constante, sino en un porcentaje constante; el porcentaje de aumento es 20%.

Al sacar logaritmos a ambos lados de la ecuación (16.7) se puede transformar este modelo no lineal en un modelo lineal.

$$\log E(y) = \log \beta_0 + x \log \beta_1 \quad (16.8)$$

Ahora, si $y' = \log E(y)$, $\beta'_0 = \log \beta_0$, y $\beta'_1 = \log \beta_1$, la ecuación (16.8) se expresa como

$$y' = \beta'_0 + \beta'_1 x$$

Ahora es claro que se puede emplear la fórmula de la regresión lineal simple para obtener estimaciones de β'_0 y de β'_1 . Al denotar estas estimaciones como b'_0 y b'_1 se obtiene la siguiente ecuación estimada de regresión.

$$\hat{y}' = b'_0 + b'_1 x \quad (16.9)$$

Para obtener predicciones para la variable dependiente original y dado un valor de x , primero se sustituye el valor de x en la ecuación (16.9) y se calcula \hat{y}' . El antilogaritmo de \hat{y}' será la predicción de y o el valor esperado de y .

Muchos modelos no lineales pueden ser transformados a un modelo lineal equivalente. Sin embargo, tales modelos han tenido pocas aplicaciones en el comercio y la economía. Además, los fundamentos matemáticos para el estudio de tales modelos quedan fuera del alcance de este libro.

Ejercicios

Métodos

1. Considere los datos siguientes para las variables x y y .

x	22	24	26	30	35	40
y	12	21	33	35	40	36

- Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1 x$.
- Use los resultados del inciso a para probar si existe una relación significativa entre x y y . Use $\alpha = 0.05$.
- Obtenga el diagrama de dispersión de estos datos. ¿Este diagrama de dispersión sugiere una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1 x + b_2 x^2$? Explique.

- d. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x + b_2x^2$
- e. Remítase al inciso d. ¿La relación entre x , x^2 y y es significativa? Use $\alpha = 0.05$
- f. Prediga el valor de y para $x = 25$.
2. Considere los datos siguientes para las variables x y y .

x	9	32	18	15	26
y	10	20	21	16	22

- a. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x$. Presente un comentario sobre lo adecuado de esta ecuación para predecir y .
- b. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x + b_2x^2$. Dé un comentario sobre lo adecuado de esta relación para predecir y .
- c. Prediga el valor de y para $x = 20$.
3. Considere los datos siguientes para las variables x y y .

x	2	3	4	5	7	7	7	8	9
y	4	5	4	6	4	6	9	5	11

- a. ¿Parece haber una relación lineal entre x y y ? Explique.
- b. Obtenga la ecuación estimada de regresión que relaciona x y y .
- c. Dada la ecuación estimada de regresión obtenida en el inciso b, grafique los residuales estandarizados contra \hat{y} . ¿Las suposiciones del modelo parecen satisfacerse? Explique.
- d. Realice una transformación logarítmica de la variable dependiente y . Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada. ¿Las suposiciones del modelo parecen satisfacerse cuando se usa una variable dependiente transformada? En este caso, ¿la transformación recíproca funciona mejor? Explique.

Aplicaciones

4. El departamento de autopistas estudia la relación entre flujo de tráfico y velocidad. Se considera que el modelo siguiente es el adecuado.

$$y = \beta_0 + \beta_1x + \epsilon$$

donde

y = flujo de tráfico en vehículos por hora

x = velocidad de los vehículos en millas por hora

Los siguientes datos fueron recolectados durante “horas pico” en las seis principales autopistas que salen de la ciudad.

Flujo de tráfico (y)	Velocidad de los vehículos (x)
1256	35
1329	40
1226	30
1335	45
1349	50
1124	25

- a. Obtenga con estos datos una ecuación estimada de regresión.
- b. Use $\alpha = 0.01$ para probar la significancia de la relación.

Autoexamen

5. Para continuar con el problema del ejercicio 4, se sugiere emplear la siguiente ecuación estimada de regresión curvilínea

$$\hat{y} = b_0 + b_1x + b_2x^2$$

- a. Use los datos del problema 4 para estimar los parámetros de esta ecuación estimada de regresión.
 - b. Use $\alpha = 0.01$ para probar la significancia de la relación.
 - c. Estime el flujo de tráfico en vehículos por hora correspondiente a 38 millas por hora.
6. En un estudio sobre instalaciones para servicios de emergencia se investigó la relación entre el número de instalaciones y la distancia promedio a recorrer para dar el servicio de emergencia. En la tabla siguiente se presentan los datos obtenidos.

Número de instalaciones	Distancia promedio (millas)
9	1.66
11	1.12
16	0.83
21	0.62
27	0.51
30	0.47

- a. Trace el diagrama de dispersión de estos datos, considere la distancia promedio a recorrer como la variable dependiente.
 - b. ¿Un modelo lineal simple será apropiado? Explique.
 - c. Con estos datos obtenga la ecuación estimada de regresión que mejor explica la relación entre las dos variables.
7. Un factor importante al comprar un monitor para computadora es el campo de visión. Un monitor que tenga un campo de visión amplio permite tener una imagen aceptable con sólo girar ligeramente la cabeza y una persona de pie cerca del monitor logra ver claramente la imagen de la pantalla. Después de una revisión de monitores LCD de 19 pulgadas, *PC World* encontró que aunque todos los monitores probados aseguraban arcos de 170 grados —tanto horizontal como verticalmente— el rango real de los monitores iba de 108 a 167 grados. En la tabla siguiente se da el ángulo de visión horizontal de ocho monitores de 19 pulgadas y la evaluación dada por *PC World* con base en la calidad de la imagen, el precio y en las políticas de soporte técnico (*PC World*, febrero de 2003).

Monitor	Ángulo	Evaluación
Samsung SyncMaster 191T	167	86
ViewSonic VX900	159	82
Sceptre Technologies X9S-Naga	126	81
Planar PL191M	108	81
Dell UltraSharp 1900FP	153	81
AOC LM914	123	81
KDS USA Radius Rad-9	118	80
NEC MultiSync LCD 1920NX	123	80
Iiyama Pro Lite 4821DT-BK	119	80

- a. Desarrolle un diagrama de dispersión de estos datos, emplee como variable independiente el ángulo de visión horizontal.
- b. ¿Un modelo de regresión lineal simple es apropiado?
- c. Obtenga una ecuación estimada de regresión que explique la relación entre estas dos variables.

8. Corvette, Ferrari y Jaguar fabricaron varios automóviles clásicos con un valor que aún sigue en aumento. En la tabla siguiente, basada en el sistema Martin de evaluación de automóviles de colección, se presenta la evaluación de su rareza (1-20) y el precio (\$ miles) de 15 automóviles clásicos (www.businessweek.com, febrero de 2006).

Año	Fabricante	Modelo	Evaluación	Precio (\$ miles)
1984	Chevrolet	Corvette	18	1600
1956	Chevrolet	Corvette 265/225-hp	19	4000
1963	Chevrolet	Corvette coupe (340-bhp 4-speed)	18	1000
1978	Chevrolet	Corvette coupe Silver Anniversary	19	1300
1960-1963	Ferrari	250 GTE 2+2	16	350
1962-1964	Ferrari	250 GTL Lusso	19	2650
1962	Ferrari	250 GTO	18	375
1967-1968	Ferrari	275 GTB/4 NART Spyder	17	450
1968-1973	Ferrari	365 GTB/4 Daytona	17	140
1962-1967	Jaguar	E-type OTS	15	77.5
1969-1971	Jaguar	E-type Series II OTS	14	62
1971-1974	Jaguar	E-type Series III OTS	16	125
1951-1954	Jaguar	XK 120 roadster (steel)	17	400
1950-1953	Jaguar	XK C-type	16	250
1956-1957	Jaguar	XKSS	13	70

- a. Dé el diagrama de dispersión de estos datos, emplee la evaluación de la rareza como variable independiente y el precio como variable independiente. ¿Es apropiado un modelo de regresión lineal simple?
- b. Obtenga una ecuación estimada de regresión simple en la cual las variables independientes sean $x = \text{evaluación de la rareza}$ y x^2 .
- c. Considere la relación no lineal dada por la ecuación (16.7). Use logaritmos para obtener una ecuación estimada de regresión para este modelo.
- d. ¿Cuál de las ecuaciones estimadas de regresión prefiere, la del inciso b o la del inciso c? Explique.
9. Casi en todo el sistema de trenes ligeros de Estados Unidos se usan vagones eléctricos que corren sobre vías construidas a nivel de la calle. De acuerdo con la Administración de Tránsito Federal de Estados Unidos, el tren ligero es uno de los medios de transporte más seguros, con una tasa de accidentes de 0.99 accidentes por cada millón de millas de pasajeros en comparación con 2.29 en los autobuses. En la tabla siguiente se presenta, para algunos de los sistemas de tren ligero de Estados Unidos, el número de millas de vía y el número de pasajeros, en miles, que utilizan el transporte público en un día entre semana.

Ciudad	Millas	Pasajeros
Los Ángeles	22	70
San Diego	47	75
Portland	38	81
Sacramento	21	31
San Jose	31	30
San Francisco	73	164
Philadelphia	69	84
Boston	51	231
Denver	17	35
Salt Lake City	18	28

(continúa)

Ciudad	Millas	Pasajeros
Dallas	44	39
Nueva Orleans	16	14
San Luis	34	42
Pittsburgh	18	25
Buffalo	6	23
Cleveland	15	15
Newark	9	8

- Dé el diagrama de dispersión de estos datos, emplee como variable independiente el número de millas de vía. ¿Es apropiado emplear un modelo de regresión lineal?
- Use un modelo de regresión lineal simple para obtener una ecuación estimada de regresión que sirva para predecir el número de pasajeros por día entre semana, dado que conoce el número de millas de vía. Construya una gráfica de residuales estandarizados. Con base en la gráfica de residuales estandarizados diga si el modelo de regresión lineal simple es apropiado.
- Realice una transformación logarítmica de la variable dependiente. Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada. ¿Las suposiciones del modelo al usar la variable dependiente transformada se satisfacen?
- Realice una transformación recíproca de la variable dependiente. Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada.
- ¿Cuál de las ecuaciones de regresión estimada recomienda? Explique.

16.2

Determinación de cuándo agregar o quitar variables

En esta sección se mostrará el uso de la prueba F para determinar si es ventajoso agregar una o más variables independientes a un modelo de regresión múltiple. Esta prueba se basa en determinar la disminución del valor de la suma de cuadrados debidos al error al agregar una o más variables independientes al modelo. Primero se ilustrará el uso del modelo en el contexto del ejemplo de Butler Trucking.

En el capítulo 15 se presentó el modelo de Butler Trucking para ilustrar el uso del análisis de regresión múltiple. Como recordará, los directivos de esta empresa deseaban obtener una ecuación estimada de regresión para predecir el tiempo total del recorrido diario de sus camiones repartidores, a partir de dos variables independientes: millas recorridas y número de entregas. Al usar como única variable independiente el número de millas recorridas x_1 , se obtuvo la siguiente ecuación estimada de regresión, mediante el método de mínimos cuadrados.

$$\hat{y} = 1.27 + 0.0678x_1$$

En el capítulo 15 se mostró que la suma de cuadrados debidos al error con este modelo era: $SCE = 8.029$. Cuando se agregó al modelo otra variable independiente, número de entregas x_2 , se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

La suma de cuadrados debidos al error en este modelo fue $SCE = 2.299$. Agregar x_2 dio como resultado una reducción en el valor de SCE . La pregunta es: ¿La adición de la variable x_2 conduce a una reducción *significativa* en el valor de la SCE ?

Se empleará la notación siguiente: $SCE(x_1)$ para denotar la suma de cuadrados debidos al error cuando x_1 es la única variable independiente del modelo; $SCE(x_1, x_2)$ para denotar la suma

de cuadrados debidos al error, cuando tanto x_1 como x_2 son las variables del modelo, y así sucesivamente. Por tanto, la disminución del valor de la SCE que se obtuvo al adicionar x_2 al modelo que sólo tenía como variable independiente a x_1 es

$$\text{SCE}(x_1) - \text{SCE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

Para determinar si esta reducción es significativa se realiza una prueba F .

El numerador del estadístico F es la disminución en el valor de SCE dividida entre la cantidad de variables independientes agregadas al modelo original. En este caso, sólo se agregó una variable, x_2 , por lo que el numerador del estadístico F es

$$\frac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1} = 5.730$$

Lo que se obtiene es una medida de la disminución de SCE por variable independiente añadida al modelo. El denominador del estadístico F es el cuadrado medio debido al error en el modelo que contiene todas las variables independientes. En el caso del ejemplo de Butler Trucking, esto corresponde al modelo que tiene tanto a x_1 como a x_2 ; por tanto $p = 2$ y

$$\text{CME} = \frac{\text{SCE}(x_1, x_2)}{n - p - 1} = \frac{2.299}{7} = 0.3284$$

El siguiente estadístico F es la base para probar si la adición de x_2 es estadísticamente significativa.

$$F = \frac{\frac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\frac{\text{SCE}(x_1, x_2)}{n - p - 1}} \quad (16.10)$$

El número de grados de libertad en el numerador de este estadístico F es igual al número de variables agregadas al modelo y el número de grados en el denominador es igual a $n - p - 1$.

En el caso del problema de Butler Trucking se obtiene

$$F = \frac{\frac{5.730}{1}}{\frac{2.299}{7}} = \frac{5.730}{0.3284} = 17.45$$

Si consulta la tabla 4 del apéndice B, se encuentra que para el nivel de significancia 0.05, $F_{0.05} = 5.59$, por lo que se puede rechazar la hipótesis de que x_2 no sea estadísticamente significativa; en otras palabras, al agregar x_2 al modelo en el que sólo se tiene como variable independiente x_1 , se obtiene una disminución significativa en la suma de los cuadrados debido al error.

Cuando se desea probar la significancia de agregar sólo una variable independiente al modelo, el resultado que se obtiene con la prueba F que se acaba de describir, también se obtiene con la prueba t para la significancia de uno solo de los parámetros (descrita en la sección 15.4). El estadístico F que se acaba de calcular es el cuadrado del estadístico t que se usa para probar la significancia de un solo parámetro.

Puesto que, cuando se agrega una sola variable independiente al modelo, la prueba t es equivalente a la prueba F , esto permite aclarar el uso adecuado de la prueba t para probar la significancia de uno de los parámetros. Si uno de los parámetros no es significativo, la variable correspondiente puede ser eliminada del modelo. Pero, si la prueba t indica que hay dos o más

parámetros que no son significativos, nunca se debe eliminar del modelo más de una variable independiente, con base en la prueba t ; cuando se elimina una variable, puede resultar que una segunda variable, que inicialmente no era significativa, se vuelva significativa.

Ahora cabe considerar si la adición de más de una variable independiente –como conjunto– da como resultado que haya una reducción significativa de la suma de los cuadrados debidos al error.

Caso general

Considere el siguiente modelo de regresión múltiple en el que intervienen q variables independientes, donde $q < p$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \epsilon \quad (16.11)$$

Si a este modelo se le agregan las variables $x_{q+1}, x_{q+2}, \dots, x_p$, se obtiene un modelo con p variables independientes.

$$\begin{aligned} y = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \\ & + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \cdots + \beta_p x_p + \epsilon \end{aligned} \quad (16.12)$$

Para probar si la adición de $x_{q+1}, x_{q+2}, \dots, x_p$, es estadísticamente significativa, las hipótesis nula y alternativa pueden plantearse como sigue.

$$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$$

H_a : Uno o más de los parámetros no es igual a cero

El siguiente estadístico F aporta la base para probar si la adición de estas variables independientes es estadísticamente significativa.

$$F = \frac{\frac{\text{SCE}(x_1, x_2, \dots, x_q) - \text{SCE}(x_1, x_2, \dots, x_q, x_{q+1}, \dots, x_p)}{p - q}}{\frac{\text{SCE}(x_1, x_2, \dots, x_q, x_{q+1}, \dots, x_p)}{n - p - 1}} \quad (16.13)$$

Este valor F calculado se compara con F_α , el valor en la tabla para $p - q$ grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador. Si $F > F_\alpha$ se rechaza H_0 y se concluye que el conjunto de variables independientes agregadas es estadísticamente significativo. Observe que si $q = 1$ y $p = 2$, la ecuación (16.13) se reduce a la ecuación (16.10).

Para muchos estudiantes encontrar la ecuación (16.13) resulta un poco complicado. Para dar una descripción un poco más sencilla de este cociente F , al modelo que tiene la menor cantidad de variables independientes se le denomina modelo reducido y al modelo que tiene la mayor cantidad de variables independientes se le denomina modelo completo. Si $\text{SCE}(\text{reducido})$ denota la suma de los cuadrados debidos al error del modelo reducido y $\text{SCE}(\text{completo})$ denota la suma de los cuadrados debidos al error del modelo completo, el numerador de la ecuación (16.13) se expresa como

$$\frac{\text{SCE}(\text{reducido}) - \text{SCE}(\text{completo})}{\text{número de términos extra}} \quad (16.14)$$

Muchos paquetes de software, como Minitab, proporcionan sumas de cuadrados que corresponden al orden en el que cada variable independiente entra al modelo; en tales casos se simplifican los cálculos de la prueba F para determinar si agregar o eliminar un conjunto de variables.

Observe que “número términos extra” denota la diferencia entre el número de variables independientes en el modelo completo y el número de variables independientes en el modelo reducido. El denominador de la ecuación (16.13) es la suma de los cuadrados debidos al error en el modelo completo dividida entre los correspondientes grados de libertad; en otras palabras, el denominador

nador es el cuadrado medio debido al error en el modelo completo. Si se denota el cuadrado medio debido al error del modelo completo como CME(completo) se puede escribir

$$F = \frac{\frac{\text{SCE(reducido)} - \text{SCE(completo)}}{\text{número de términos extra}}}{\text{CME(completo)}} \quad (16.15)$$

Para ilustrar el uso de este estadístico F , suponga que se tiene un problema de regresión que tiene 30 observaciones. En un modelo en el que intervienen las variables independientes x_1 , x_2 y x_3 la suma de los cuadrados debida al error es 150 y en un segundo modelo en el que las variables independientes son x_1 , x_2 , x_3 , x_4 y x_5 , la suma de los cuadrados debida al error es 100. ¿La adición de las variables x_4 y x_5 produjo una reducción significativa de la suma de los cuadrados debida al error?

Observe, primero, que el número de grados de libertad para STC es $30 - 1 = 29$ y que el número de grados de libertad para la suma de cuadrados debida a la regresión para el modelo completo es cinco (el número de variables independientes en el modelo completo). Por tanto, los grados de libertad para la suma de los cuadrados debida al error en el modelo completo es $29 - 5 = 24$, entonces $\text{CME(completo)} = 100/24 = 4.17$. Así, el estadístico F es

$$F = \frac{\frac{150 - 100}{2}}{4.17} = 6.00$$

Este valor F que se ha calculado se compara con el valor F que se encuentra en la tabla para dos grados de libertad en el numerador y 24 grados de libertad en el denominador. Para el nivel de significancia 0.05, en la tabla 4 del apéndice B se encuentra $F_{0.05} = 3.40$. Como $F = 6.00$ es mayor que 3.40, se concluye que la adición de las variables x_4 y x_5 es estadísticamente significativa.

Uso del valor- p

También puede usarse el criterio del valor- p para determinar si resulta ventajoso agregar una o más variables independientes a un modelo de regresión múltiple. En el ejemplo anterior se mostró cómo realizar la prueba F para determinar si la adición de dos variables independientes, x_4 y x_5 , a un modelo con tres variables independientes, x_1 , x_2 y x_3 , era estadísticamente significativo. En ese ejemplo el valor que se obtuvo para el estadístico F fue 6.00 y se concluyó (por comparación de $F = 6.00$ con el valor crítico $F_{0.05} = 3.40$) que la adición de las variables x_4 y x_5 era significativa. El valor- p que corresponde a $F = 6.00$ (2 grados de libertad en el numerador y 24 grados de libertad en el denominador) es 0.008. Como el valor- $p = 0.008 < \alpha = 0.05$, también se concluye que la adición de las dos variables independientes es significativa. Al emplear las tablas de la distribución F es difícil determinar directamente el valor- p , pero los paquetes de software como Minitab o Excel facilitan este cálculo.

NOTAS Y COMENTARIOS

El cálculo del estadístico F también se basa en las sumas de cuadrados debida a la regresión. Para mostrar esta forma del estadístico F , se nota primero que

$$\begin{aligned} \text{SCE(reducido)} &= \text{STC} - \text{SCR(reducido)} \\ \text{SCE(completo)} &= \text{STC} - \text{SCR(completo)} \end{aligned}$$

Por tanto

$$\begin{aligned} \text{SCE(reducido)} - \text{SCE(completo)} &= [\text{STC} - \text{SCR(reducido)}] - [\text{STC} - \text{SCR(completo)}] \\ &= \text{SCR(completo)} - \text{SCR(reducido)} \end{aligned}$$

Así,

$$F = \frac{\frac{\text{SCR(completo)} - \text{SCR(reducido)}}{\text{número de términos extra}}}{\text{CME(completo)}}$$

Ejercicios

Métodos

10. En un análisis de regresión en el que se emplearon 27 observaciones, se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = 25.2 + 5.5x_1$$

Para esta ecuación estimada de regresión $\text{STC} = 1\,550$ y $\text{SCE} = 520$.

- a. Utilice $\alpha = 0.05$ y pruebe si x_1 es significativa.
Suponga que a este modelo le agrega las variables x_2 y x_3 y obtiene la ecuación de regresión siguiente.

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

Para esta ecuación estimada de regresión $\text{STC} = 1\,550$ y $\text{SCE} = 100$.

- b. Use una prueba F y 0.05 como nivel de significancia para determinar si x_2 y x_3 contribuyen significativamente al modelo.
11. En un análisis de regresión en el que se emplearon 30 observaciones, se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

Para esta ecuación estimada de regresión $\text{STC} = 1\,805$ y $\text{SCR} = 1\,760$.

- a. Con $\alpha = 0.05$, pruebe la significancia de la relación entre las variables.
Suponga que de este modelo elimina las variables x_1 y x_4 y obtiene la siguiente ecuación estimada de regresión.

$$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$$

Para esta ecuación estimada de regresión $\text{STC} = 1\,805$ y $\text{SCR} = 1\,705$.

- b. Calcule $\text{SCE}(x_1, x_2, x_3$ y $x_4)$
c. Calcule $\text{SCE}(x_2, x_3)$
d. Use una prueba F y 0.05 como nivel de significancia para determinar si x_1 y x_4 contribuyen significativamente al modelo.

Aplicaciones

12. La Ladies Professional Golfers Association (LPGA, por sus siglas en inglés) lleva estadísticas sobre el desempeño y las ganancias de sus miembros en la LPGA Tour. En el archivo titulado LPGA del disco compacto se presentan las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que obtuvieron las mayores ganancias en la LPGA Tour de 2005 (www.lpga.com, 2006). Earnings (ganancias) (\$ miles) son los ingresos totales en miles de dólares; Scoring Avg., es la puntuación promedio de una jugadora en todos los eventos; Greens in Reg., es el porcentaje de las veces que una jugadora llega al green en regulación; Putting Avg., es el promedio de putts hechos en el green en regulación, y Sand Saves es el porcentaje de veces que la jugadora

- logra “subir y bajar” (“up and down”) cuando se encuentra en un búnker de arena al lado del green.
- Obtenga una ecuación estimada de regresión que sirva para predecir Scoring Avg. dado Greens in Reg.
 - Obtenga una ecuación estimada de regresión que sirva para predecir Scoring Avg. dados Greens in Reg., Putting Avg. y Sand Saves.
 - Con un nivel de significancia 0.05 pruebe si las dos variables independientes agregadas en el inciso b, Putting Avg. y Sand Saves, contribuyen significativamente a la ecuación estimada de regresión obtenida en el inciso a. Explique.
13. Vaya al ejercicio 12.
- Obtenga una ecuación estimada de regresión que sirva para predecir Earnings, conociendo Putting Avg.
 - Obtenga una ecuación estimada de regresión que sirva para predecir Earnings, conociendo Putting Avg. y Sand Saves.
 - Emplee como nivel de significancia 0.05 y pruebe si las dos variables independientes agregadas en el inciso b, Putting Avg. y Sand Saves, contribuyen significativamente a la ecuación estimada de regresión obtenida en el inciso a. Explique.
 - En general, puntuaciones más bajas llevan a ganancias más altas. Para investigar esta opción para predecir las Earnings, obtenga una ecuación estimada de regresión que sirva para predecir Earnings, Scoring Avg. ¿Preferiría emplear esta ecuación para predecir las ganancias (Earnings) o la ecuación obtenida en el inciso b? Explique.
14. En un estudio realizado a lo largo de 10 años por la American Heart Association se obtuvieron datos acerca de la relación entre edad, presión sanguínea y fumar con el riesgo a sufrir un infarto. Los datos que se presentan a continuación son parte de este estudio. El riesgo se interpreta como la probabilidad (multiplicada por 100) de que el paciente sufra un infarto en los próximos 10 años. Para la variable fumar, defina una variable ficticia que tome el valor 1 si la persona es fumadora y el valor 0 si no es fumadora.

Riesgo	Edad	Presión	Fumador
12	57	152	0
24	67	163	0
13	58	155	0
56	86	177	1
28	59	196	0
51	76	189	1
18	56	155	1
31	78	120	0
37	80	135	1
15	78	98	0
22	71	152	0
36	70	173	1

(continúa)

Riesgo	Edad	Presión	Fumador
15	67	135	1
48	77	209	1
15	60	199	0
36	82	119	1
8	66	166	0
34	80	125	1
3	62	117	0
37	59	207	1

- a. Obtenga una ecuación estimada de regresión que sirva para predecir el riesgo de sufrir un infarto, dados edad y presión sanguínea.
- b. Considere la adición de dos variables independientes al modelo obtenido en el inciso a, una para la interacción entre edad y presión sanguínea y otra que indique si la persona es o no un fumador. Obtenga una ecuación estimada de regresión, emplee estas cuatro variables independientes.
- c. Emplee como nivel de significancia 0.05, realice una prueba para determinar si la adición de la variable de la interacción y la variable fumador contribuyen significantemente a la ecuación estimada de regresión obtenida en el inciso a.
15. La National Football League, NFL, evalúa a sus prospectos con una escala que va del 5 al 9. Estas evaluaciones se interpretan como sigue: 8-9 deberá empezar el año próximo; 7.0-7.9 deberá empezar; 6.0-6.9 servirán de respaldo al equipo y 5.0-5.9 pueden formar parte del club y contribuir. En la tabla siguiente se da posición, peso, tiempo en segundos para correr 40 yardas y la evaluación dada por la NFL a 40 prospectos (*USA Today*, 14 de abril de 2000).

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
1	Peter Warrick	Receptor	194	4.53	9.0
2	Plaxico Burress	Receptor	231	4.52	8.8
3	Sylvester Morris	Receptor	216	4.59	8.3
4	Travis Taylor	Receptor	199	4.36	8.1
5	Laveranues Coles	Receptor	192	4.29	8.0
6	Dez White	Receptor	218	4.49	7.9
7	Jerry Porter	Receptor	221	4.55	7.4
8	Ron Dugans	Receptor	206	4.47	7.1
9	Todd Pinkston	Receptor	169	4.37	7.0
10	Dennis Northcutt	Receptor	175	4.43	7.0
11	Anthony Lucas	Receptor	194	4.51	6.9
12	Darrell Jackson	Receptor	197	4.56	6.6
13	Danny Farmer	Receptor	217	4.60	6.5
14	Sherrod Gideon	Receptor	173	4.57	6.4
15	Trevor Gaylor	Receptor	199	4.57	6.2
16	Cosey Coleman	Guardia	322	5.38	7.4
17	Travis Claridge	Guardia	303	5.18	7.0
18	Kaulana Noa	Guardia	317	5.34	6.8
19	Leander Jordan	Guardia	330	5.46	6.7
20	Chad Clifton	Guardia	334	5.18	6.3
21	Manula Savea	Guardia	308	5.32	6.1
22	Ryan Johanningmeir	Guardia	310	5.28	6.0
23	Mark Tauscher	Guardia	318	5.37	6.0
24	Blaine Saipaia	Guardia	321	5.25	6.0
25	Richard Mercier	Guardia	295	5.34	5.8
26	Damion McIntosh	Guardia	328	5.31	5.3
27	Jeno James	Guardia	320	5.64	5.0

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
28	Al Jackson	Guardia	304	5.20	5.0
29	Chris Samuels	Tackle ofensivo	325	4.95	8.5
30	Stockar McDougle	Tackle ofensivo	361	5.50	8.0
31	Chris McIngosh	Tackle ofensivo	315	5.39	7.8
32	Adrian Klemm	Tackle ofensivo	307	4.98	7.6
33	Todd Wade	Tackle ofensivo	326	5.20	7.3
34	Marvel Smith	Tackle ofensivo	320	5.36	7.1
35	Michael Thompson	Tackle ofensivo	287	5.05	6.8
36	Bobby Williams	Tackle ofensivo	332	5.26	6.8
37	Darnell Alford	Tackle ofensivo	334	5.55	6.4
38	Terrance Beadles	Tackle ofensivo	312	5.15	6.3
39	Tutan Reyes	Tackle ofensivo	299	5.35	6.1
40	Greg Robinson-Ran	Tackle ofensivo	333	5.59	6.0

- Dé una variable ficticia para la posición de los jugadores.
- Obtenga una ecuación estimada de regresión que indique la relación entre evaluación y posición, peso y tiempo para correr 40 yardas.
- Emplee 0.05 como nivel de significancia, pruebe si la ecuación estimada de regresión obtenida en el inciso b representa una relación significante entre las variables independientes y la variable dependiente.
- ¿La posición es un factor significativo para la evaluación del jugador? Use $\alpha = 0.05$. Explique.

16.3

Análisis de un problema mayor

Al introducir el análisis de regresión múltiple, se usó ampliamente el ejemplo de Butler Trucking. Al explorar los conceptos fue una ventaja que este problema fuera pequeño. Sin embargo, este problema pequeño dificulta ilustrar algunas de las cuestiones relacionadas con la selección de variables que intervienen en la construcción de un modelo. Para dar un ejemplo de los procesos de selección de variables que se estudian en la sección siguiente, se introduce un conjunto de datos que consta de 25 observaciones con ocho variables independientes. El doctor David W. Cravens del departamento de marketing de la Texas Christian University otorgó el permiso para emplear estos datos. Por esta razón a este conjunto de datos se le llamará datos de Cravens.*

Los datos de Cravens son de una empresa que tiene varios territorios de ventas, cada uno de los cuales le está asignado a un solo representante de ventas. Para determinar si diversas variables (independientes) predictoras podían explicar las ventas en cada uno de los territorios se realizó un análisis de regresión. A partir de una muestra de 25 territorios se obtuvieron los datos que se muestran en la tabla 16.5; en la tabla 16.6 se presenta la definición de las variables.

Como paso preliminar se considerarán los coeficientes de correlación entre cada par de variables. En la figura 16.13 se presenta la matriz de correlación obtenida con Minitab. Observe que el coeficiente de correlación muestral entre Sales y Time es 0.623, entre Sales y Poten es 0.598 y así sucesivamente.

Si observa los coeficientes de correlación entre las variables independientes, se dará cuenta de que la correlación entre Time y Accounts es 0.758; por tanto, si Accounts se usa como una de las variables independientes, Time no agregaría mucho poder explicatorio al modelo. Recuerde la prueba de la regla práctica que se vio en la sección 15.4, donde la multicolinealidad puede causar problemas si el valor absoluto del coeficiente de correlación muestral, entre cualesquiera dos de las variables independientes, es mayor que 0.7. Por tanto, siempre que sea posible,

*Para más detalles ver David W. Cravens, Robert B. Woodruff y Joe C. Stamper, "An Analytical Approach for Evaluating Sales Territory Performance", *Journal of Marketing*, 36 (enero, 1972): 31-37. Copyright © 1972 American Marketing Association.

TABLA 16.5 DATOS DE CRAVENS

Ventas	Antigüedad	Potencial	GastPubl	Participación	Cambio	Cuentas	Trabajo	Evaluación
3 669.88	43.10	74 065.1	4 582.9	2.51	0.34	74.86	15.05	4.9
3 473.95	108.13	58 117.3	5 539.8	5.51	0.15	107.32	19.97	5.1
2 295.10	13.82	21 118.5	2 950.4	10.91	-0.72	96.75	17.34	2.9
4 675.56	186.18	68 521.3	2 243.1	8.27	0.17	195.12	13.40	3.4
6 125.96	161.79	57 805.1	7 747.1	9.15	0.50	180.44	17.64	4.6
2 134.94	8.94	37 806.9	402.4	5.51	0.15	104.88	16.22	4.5
5 031.66	365.04	50 935.3	3 140.6	8.54	0.55	256.10	18.80	4.6
3 367.45	220.32	35 602.1	2 086.2	7.07	-0.49	126.83	19.86	2.3
6 519.45	127.64	46 176.8	8 846.2	12.54	1.24	203.25	17.42	4.9
4 876.37	105.69	42 053.2	5 673.1	8.85	0.31	119.51	21.41	2.8
2 468.27	57.72	36 829.7	2 761.8	5.38	0.37	116.26	16.32	3.1
2 533.31	23.58	33 612.7	1 991.8	5.43	-0.65	142.28	14.51	4.2
2 408.11	13.82	21 412.8	1 971.5	8.48	0.64	89.43	19.35	4.3
2 337.38	13.82	20 416.9	1 737.4	7.80	1.01	84.55	20.02	4.2
4 586.95	86.99	36 272.0	10 694.2	10.34	0.11	119.51	15.26	5.5
2 729.24	165.85	23 093.3	8 618.6	5.15	0.04	80.49	15.87	3.6
3 289.40	116.26	26 878.6	7 747.9	6.64	0.68	136.58	7.81	3.4
2 800.78	42.28	39 572.0	4 565.8	5.45	0.66	78.86	16.00	4.2
3 264.20	52.84	51 866.1	6 022.7	6.31	-0.10	136.58	17.44	3.6
3 453.62	165.04	58 749.8	3 721.1	6.35	-0.03	138.21	17.98	3.1
1 741.45	10.57	23 990.8	861.0	7.37	-1.63	75.61	20.99	1.6
2 035.75	13.82	25 694.9	3 571.5	8.39	-0.43	102.44	21.66	3.4
1 578.00	8.13	23 736.3	2 845.5	5.15	0.04	76.42	21.46	2.7
4 167.44	58.44	34 314.3	5 060.1	12.88	0.22	136.58	24.78	2.8
2 799.97	21.14	22 809.5	3 552.0	9.14	-0.74	88.62	24.96	3.9

se evitara incluir a las dos variables, Time y Accounts, en el modelo. También el coeficiente de correlación muestral entre Change y Rating, que es 0.549, es elevado y merece ser considerado más cuidadosamente.

Al observar los coeficientes de correlación muestrales entre Sales y cada una de las variables independientes se puede tener una rápida idea de cuáles de las variables independientes son, en sí mismas, buenos predictores. Se encuentra que el mejor predictor de Sales es Accounts, debi-

TABLA 16.6 DEFINICIÓN DE LAS VARIABLES EN LOS DATOS DE CRAVENS

Variable	Definición
Ventas	Total de ventas acreditadas al representante de ventas
Antigüedad (Time)	Antigüedad del empleado en meses
Potencial (Poten)	Potencial de mercado: ventas industriales totales en unidades en el territorio de ventas*
GastPubl (AdvExp)	Gastos del territorio en publicidad
Participación (Share)	Participación en el mercado: promedio ponderado de los últimos cuatro años
Cambio (Change)	Cambio, en los últimos cuatro años en participación en el mercado
Cuentas (Accounts)	Número de cuentas asignadas a los representantes de ventas*
Trabajo (Work)	Carga de trabajo: índice ponderado basado en compras anuales y concentración de cuentas
Evaluación (Rating)	Evaluación general del representante de ventas sobre ocho dimensiones de desempeño: una evaluación agregada en una escala del 1-7

*Estos datos fueron codificados para proteger la confidencialidad.

FIGURA 16.13 COEFICIENTES DE CORRELACIÓN MUESTRAL DE LOS DATOS DE CRAVENS

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.623							
Poten	0.598	0.454						
AdvExp	0.596	0.249	0.174					
Share	0.484	0.106	-0.211	0.264				
Change	0.489	0.251	0.268	0.377	0.085			
Accounts	0.754	0.758	0.479	0.200	0.403	0.327		
Work	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277

do a que su coeficiente de correlación muestral es el más alto (0.754). Recuerde que si sólo se tiene una variable independiente, el cuadrado del coeficiente de correlación muestral es el coeficiente de determinación. Por tanto, Accounts explica $(0.754)^2(100)$, o 56.85%, de la variabilidad en Sales. Las variables independientes que siguen en importancia son Time, Poten y AdvExp, cada una con un coeficiente de correlación muestral de 0.6, aproximadamente.

Aun cuando existen problemas potenciales de multicolinealidad, se va a obtener una ecuación estimada de regresión con estas ocho variables. Con el paquete de software Minitab se obtienen los resultados que se presentan en la figura 16.14. El coeficiente de determinación ajustado para este modelo de regresión múltiple con ocho variables es 88.3%. Observe, sin embargo, que los valores-*p* de las pruebas *t* para cada uno de los parámetros indican que sólo Poten, AdvExp y Share son significativos a un nivel de significancia $\alpha = 0.05$, dado el efecto de todas las demás variables. Por tanto, se deseará investigar los resultados que se obtienen si se usan solamente estas tres

FIGURA 16.14 RESULTADOS DE MINITAB PARA EL MODELO CON OCHO VARIABLES INDEPENDIENTES

The regression equation is					
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share					
+ 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating					
Predictor	Coef	SE Coef	T	p	
Constant	1507.8	778.6	-1.94	0.071	
Time	2.010	1.931	1.04	0.313	
Poten	0.037205	0.008202	4.54	0.000	
AdvExp	0.15099	0.04711	3.21	0.006	
Share	199.02	67.03	2.97	0.009	
Change	290.9	186.8	1.56	0.139	
Accounts	5.551	4.776	1.16	0.262	
Work	19.79	33.68	0.59	0.565	
Rating	8.2	128.5	0.06	0.950	
S = 449.0 R-sq = 92.2% R-sq(adj) = 88.3%					
Analysis of Variance					
SOURCE	DF	SS	MS	F	p
Regression	8	38153568	4769196	23.65	0.000
Residual Error	16	3225984	201624		
Total	24	41379552			

FIGURA 16.15 RESULTADOS DE MINITAB PARA EL MODELO CON TRES VARIABLES, POTEN, ADVEXP Y Share

The regression equation is Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share					
Predictor Coef SE Coef T p					
Constant	-1603.6	505.6	-3.17	0.005	
Poten	0.054286	0.007474	7.26	0.000	
AdvExp	0.16748	0.04427	3.78	0.001	
Share	282.75	48.76	5.80	0.000	
S = 545.5 R-sq = 84.9% R-sq(adj) = 82.7%					
Analysis of Variance					
SOURCE DF SS MS F p					
Regression	3	35130240	11710080	39.35	0.000
Residual Error	21	6249310	297586		
Total	24	41379552			

variables. En la figura 16.15 se muestran los resultados proporcionados por Minitab para la ecuación estimada de regresión con estas tres variables. Se ve que el coeficiente de determinación ajustado para esta ecuación estimada de regresión es 82.7%, el cual, aunque no es tan bueno como el de la ecuación estimada de regresión con ocho variables, es alto.

¿Cómo se puede encontrar una ecuación estimada de regresión que dé mejores resultados, dada la información de que se dispone? Una posibilidad es calcular todas las regresiones posibles. Es decir, obtener ocho ecuaciones estimadas de regresión con una sola variable (cada una de las cuales corresponde a una de las variables independientes), 28 ecuaciones estimadas de regresión con dos variables independientes (que es el número de combinaciones de ocho variables tomadas de dos en dos), y así sucesivamente. Para los datos de Cravens se necesitan, en total, 255 ecuaciones estimadas de regresión conteniendo una o más de las variables independientes.

Con los excelentes paquetes de software de que se dispone en la actualidad, se pueden calcular todas estas regresiones. Sin embargo, hacerlo representa una gran cantidad de cálculos y requiere que se revise una gran cantidad de resultados de computadora, la mayor parte de los cuales corresponderán a modelos obviamente pobres. En lugar de hacer esto se prefiere seguir un método más sistemático para elegir el subconjunto de variables independientes que proporcione la mejor ecuación estimada de regresión. En la sección siguiente se presentan algunos de los métodos más conocidos.

16.4

Procedimientos de elección de variables

Los procedimientos de selección de variables son especialmente útiles en las primeras etapas de la construcción de un modelo, pero no pueden sustituir la experiencia y el criterio del analista.

En esta sección se verán cuatro **procedimientos de selección de variables**: la regresión por pasos, la selección hacia adelante, la selección hacia atrás y la regresión del mejor subconjunto. Dado un conjunto de datos en el que hay varias variables independientes, estos procedimientos permiten determinar con qué variables independientes se obtiene el mejor modelo. Los tres primeros procedimientos son iterativos; en cada paso del procedimiento se agrega o se elimina una variable independiente y se evalúa el nuevo modelo. El procedimiento continúa hasta que un criterio de detención indica que el procedimiento ya no puede hallar un modelo mejor. El último procedimiento (mejores subconjuntos) no es un procedimiento que evalúe las variables de una en una, sino que evalúa modelos de regresión en los que intervienen distintos subconjuntos de variables independientes.

En los procedimientos regresión por pasos, selección hacia adelante y eliminación hacia atrás, en cada paso, el criterio para elegir una variable independiente para agregarla o eliminarla del modelo, se basa en el estadístico F presentado en la sección 16.2. Suponga, por ejemplo, que se desea considerar si agregar x_2 a un modelo en el que interviene x_1 o eliminar x_2 de un modelo en el que intervienen x_1 y x_2 . Para probar si la adición o la eliminación de x_2 es estadísticamente significativa, las hipótesis nula y alternativa pueden establecerse como sigue:

$$\begin{aligned} H_0: \beta_2 &= 0 \\ H_a: \beta_2 &\neq 0 \end{aligned}$$

En la sección 16.2 (ver ecuación (16.10)) se mostró que

$$F = \frac{\frac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\frac{\text{SCE}(x_1, x_2)}{n - p - 1}}$$

es usado como criterio para determinar si la presencia de x_2 en el modelo causa una reducción significativa de la suma de los cuadrados debidos al error. El valor- p correspondiente a este estadístico F es el criterio que se emplea para determinar si se debe agregar o eliminar una variable independiente del modelo de regresión. Para el rechazo se emplea la regla usual: rechazar H_0 si valor- $p \leq \alpha$.

Regresión por pasos

El procedimiento de regresión por pasos empieza por determinar en cada paso si alguna de las variables *que ya se encuentran en el modelo* debe ser eliminada. Para esto primero se calcula el estadístico F y el correspondiente valor- p para cada una de las variables independientes que intervienen en el modelo. Minitab le llama al nivel de significancia α que se emplea para determinar si una variable independiente debe ser eliminada del modelo *Alpha to remove* (*Alpha para eliminar*). Si el valor- p de alguna de las variables independientes es mayor que *Alpha to remove*, la variable independiente que tenga el mayor valor- p se elimina del modelo y el proceso de regresión por pasos empieza un nuevo paso.

Si ninguna de las variables independientes puede ser eliminada del modelo, el procedimiento trata de ingresar otra variable independiente al modelo. Para hacer esto primero se calcula el estadístico F y el valor- p de cada variable independiente que no está en el modelo. Minitab le llama al nivel de significancia α que emplea para determinar si una variable independiente debe agregarse al modelo, *Alpha to enter* (*Alpha para ingresar*). La variable independiente que tiene el menor valor- p es ingresada al modelo siempre que su valor- p sea menor que *Alpha to enter*. Este procedimiento continúa de la misma forma hasta que no haya ninguna variable independiente que pueda ser eliminada o agregada al modelo.

En la figura 16.16 se muestran los resultados obtenidos por Minitab con el procedimiento de regresión por pasos aplicado a los datos de Cravens, con 0.05 como *Alpha to remove* y 0.05 como *Alpha to enter*. Este procedimiento por pasos terminó en cuatro pasos. La ecuación estimada de regresión obtenida con el procedimiento de regresión por pasos de Minitab es

$$\hat{y} = -1441.93 + 9.2 \text{ Accounts (Cuentas)} + 0.175 \text{ AdvExp (GastPubl)} + 0.0382 \text{ Poten} + 190 \text{ Share (Participación)}$$

En la figura 16.16, observe también que, después de cuatro pasos, $s = \sqrt{\text{CME}}$ se ha reducido de 881 en el mejor modelo con una variable [Cuentas (Accounts)] a 454. El valor de R-sq ha aumentado de 56.85 a 90.04% y el R-sq(adj) de la ecuación estimada de regresión recomendada es 88.05%.

En resumen, en cada paso del procedimiento de regresión por pasos, lo primero que se considera es si alguna de las variables independientes puede ser eliminada del modelo que se tiene. Si

Dado que los procedimientos de una en una variable no consideran todos los subconjuntos posibles de una cantidad dada de variables independientes, estos procedimientos no necesariamente eligen el modelo con el que se obtenga el valor mayor R-sq.

FIGURA 16.16 RESULTADO DE LA REGRESIÓN POR PASOS DE MINITAB PARA LOS DATOS DE CRAVENS

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05				
Response is Sales on 8 predictors, with N = 25				
Step	1	2	3	4
Constant	709.32	50.29	-327.24	-1441.93
Accounts	21.7	19.0	15.6	9.2
T-Value	5.50	6.41	5.19	3.22
P-Value	0.000	0.000	0.000	0.004
AdvExp		0.227	0.216	0.175
T-Value		4.50	4.77	4.74
P-Value		0.000	0.000	0.000
Poten			0.0219	0.0382
T-Value			2.53	4.79
P-Value			0.019	0.000
Share				190
T-Value				3.82
P-Value				0.001
S	881	650	583	454
R-Sq	56.85	77.51	82.77	90.04
R-Sq(adj)	54.97	75.47	80.31	88.05
C-p	67.6	27.2	18.4	5.4

ninguna de las variables independientes puede ser eliminada del modelo, el procedimiento verifica si alguna de las variables independientes que no intervienen en el modelo puede ser ingresada al modelo. Debido a la naturaleza del procedimiento de regresión por pasos, puede ser que una variable independiente sea ingresada al modelo en un paso, en un paso subsiguiente eliminada y después ingresada al modelo en un paso posterior. El procedimiento se detiene cuando no hay ya ninguna variable independiente que pueda ser eliminada del modelo ni agregada al modelo.

Selección hacia adelante

En el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente y se van agregando variables de una en una con el mismo procedimiento que se usa en la regresión por pasos para determinar si una variable independiente debe ser ingresada al modelo. Pero, en el procedimiento de selección hacia adelante no se permite que se elimine del modelo una variable que ha sido ingresada. El procedimiento se detiene cuando el valor-*p* de cada una de las variables independientes que no están en el modelo es mayor que *Alpha to enter*.

La ecuación estimada de regresión obtenida mediante el procedimiento de selección hacia adelante de Minitab es

$$\hat{y} = -1441.93 + 9.2 \text{ Accounts (Cuentas)} + 0.175 \text{ AdvExp (GastPubl)} + 0.0382 \text{ Poten} + 190 \text{ Share (Participación)}$$

Por tanto, en el caso de los datos de Cravens, con el procedimiento de selección hacia adelante (con 0.05 como *Alpha to enter*) se llega a la misma ecuación estimada de regresión que con el procedimiento por pasos.

Eliminación hacia atrás

En el procedimiento de eliminación hacia atrás se empieza con un modelo en el que se incluyen todas las variables independientes. Después, de una en una, se van eliminando variables independientes mediante el mismo procedimiento que en la regresión por pasos. Sin embargo, en el procedimiento de eliminación hacia atrás no se permite que una variable que ya ha sido eliminada vuelva a ser ingresada al modelo. El procedimiento se detiene cuando ninguna de las variables independientes del modelo tenga un valor-*p* mayor que *Alpha to remove*.

La ecuación estimada de regresión obtenida con el procedimiento de eliminación hacia atrás de Minitab aplicado a los datos de Cravens (con 0.05 como *Alpha to remove*) es

$$\hat{y} = -1312 + 3.8 \text{ Time (Antigüedad)} + 0.0444 \\ \text{Potencial} + 0.152 \text{ AdvExp (GastPubl)} + 259 \text{ Share (participación)}$$

Al comparar la ecuación estimada de regresión obtenida mediante el procedimiento de eliminación hacia atrás con la ecuación estimada de regresión obtenida con el procedimiento de selección hacia adelante, se ve que hay tres variables independientes comunes a los dos procedimientos: AdvExp, Poten y Share. Pero, en el procedimiento de eliminación hacia atrás se incluyó Time en lugar de Accounts.

La selección hacia adelante y la eliminación hacia atrás son dos extremos en la construcción de modelos; en el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente en el modelo y, una por una, se van agregando variables independientes, mientras que en el procedimiento de eliminación hacia atrás se empieza teniendo todas las variables independientes en el modelo y, de una en una, se eliminan variables. Con los dos procedimientos se puede llegar a la misma ecuación estimada de regresión. Sin embargo, también es posible que se llegue a ecuaciones estimadas de regresión diferentes, como ocurre en el caso de los datos de Cravens. ¿Por cuál de las ecuaciones estimadas de regresión decidirse? Esto es algo que queda a discusión. Al final el analista tiene que aplicar su propio criterio. El procedimiento de los mejores subconjuntos para la construcción de modelos que se estudia a continuación proporciona más información para la construcción de modelos, información que debe ser considerada antes de tomar la decisión final.

La selección hacia adelante y la eliminación hacia atrás pueden llevar a modelos distintos.

Regresión de los mejores subconjuntos

La regresión por pasos, la selección hacia adelante y la eliminación hacia atrás son métodos para elegir un modelo de regresión que agrega o elimina, una por una, variables independientes. Ninguno de estos métodos garantiza que dado un determinado número de variables se encuentre el mejor modelo. Por tanto, estos métodos de una por una suelen ser considerados como heurísticos para la selección de un buen modelo de regresión.

Algunos paquetes de software usan un procedimiento llamado regresión de los mejores subconjuntos que permite al usuario, hallar el mejor modelo de regresión para un número determinado de variables independientes. Minitab cuenta con este procedimiento. En la figura 16.17 se presenta parte de los resultados obtenidos mediante el procedimiento de los mejores subconjuntos de Minitab aplicado a los datos de Cravens.

En estos resultados aparecen las dos mejores ecuaciones de regresión estimada con una sola variable, las dos mejores ecuaciones con dos variables, las dos mejores ecuaciones con tres variables, etc. El criterio que se emplea para determinar cuáles son las mejores ecuaciones estimadas de regresión con un determinado número de predictores es el valor del coeficiente de determinación (*R-sq*). Por ejemplo, Accounts proporciona la mejor ecuación estimada de regresión con una sola variable independiente, *R-sq* = 56.8%; al usar AdvExp y Accounts se obtiene la mejor ecuación estimada de regresión con dos variables independientes, *R-sq* = 77.5%, y con Poten, AdvExp y Share se obtiene la mejor ecuación estimada de regresión con tres variables independientes, *R-sq* = 84.9%. Para los datos de Cravens, el mayor coeficiente de determinación ajustado (*Adj.R-sq* = 89.4%) es el del modelo con seis variables independientes, Time, Poten, AdvExp, Share, Change y Accounts. Sin embargo, el coeficiente de determinación ajustado del mejor modelo con cuatro variables independientes (Poten, AdvExp, Share y Accounts) es casi igual de alto (88.1%). Por lo general, se prefiere el modelo más sencillo con el menor número de variables.

FIGURA 16.17 PARTE DE LOS RESULTADOS OBTENIDOS CON LA REGRESIÓN DE LOS MEJORES SUBCONJUNTOS DE MINITAB

Vars	R-sq	Adj. R-sq	s	A			
				C	c	R	a
1	56.8	55.0	881.09				X
1	38.8	36.1	1049.3	X			
2	77.5	75.5	650.39		X	X	
2	74.6	72.3	691.11		X	X	
3	84.9	82.7	545.52		X	X	X
3	82.8	80.3	582.64		X	X	X
4	90.0	88.1	453.84		X	X	X
4	89.6	87.5	463.93		X	X	X
5	91.5	89.3	430.21		X	X	X
5	91.2	88.9	436.75		X	X	X
6	92.0	89.4	427.99		X	X	X
6	91.6	88.9	438.20		X	X	X
7	92.2	89.0	435.66		X	X	X
7	92.0	88.8	440.29		X	X	X
8	92.2	88.3	449.02		X	X	X

Elección final

El análisis de los datos de Cravens hecho hasta ahora es una buena preparación para tomar una decisión por un modelo, pero antes habrá que hacer también otros análisis. Como se indicó en los capítulos 14 y 15, es necesario hacer un cuidadoso análisis de los residuales. Se desea que la gráfica de los residuales del modelo elegido parezca una banda horizontal. Suponga que en los residuales no se encuentre ningún problema y que se desee emplear los resultados del procedimiento de los mejores subconjuntos para decidirse por un modelo.

El procedimiento de los mejores subconjuntos indica que el mejor modelo con cuatro variables es el que contiene las variables independientes Poten, AdvExp, Share y Accounts. Este modelo resulta ser también el modelo con cuatro variables encontrado mediante el procedimiento de regresión por pasos. La tabla 16.7 ayuda a tomar la decisión final. En esta tabla se muestran varios modelos que contienen algunas, o las cuatro, de estas cuatro variables independientes.

TABLA 16.7 MODELOS SELECCIONADOS CON Accounts, Poten, AdvExp Y Share

Modelo	Variables independientes	Adj. R-sq
1	Accounts	55.0
2	AdvExp, Accounts	75.5
3	Poten, Share	72.3
4	Poten, AdvExp, Accounts	80.3
5	Poten, AdvExp, Share	82.7
6	Poten, AdvExp, Share, Accounts	88.1

En la tabla 16.7 se ve que el modelo que sólo tiene AdvExp y Accounts es bueno. Su coeficiente de determinación ajustado es 75.5%, y con el modelo con las cuatro variables sólo se logra un aumento de 12.6 puntos porcentuales. El modelo más sencillo que sólo tiene dos variables puede preferirse si, por ejemplo, es difícil medir el potencial de mercado (Poten). Sin embargo, si ya se cuenta con los datos y se requiere gran precisión en la predicción de las ventas, es claro que se preferirá el modelo con las cuatro variables.

NOTAS Y COMENTARIOS

1. En el procedimiento por pasos se requiere que *Alpha to remove* sea mayor o igual que *Alpha to enter*. Este requerimiento evita que, en un mismo paso, una misma variable sea eliminada y reingresada.
2. Para crear nuevas variables independientes que puedan ser usadas con los procedimientos de esta sección se usan funciones de las variables independientes. Por ejemplo, si se desea tener en el modelo x_1x_2 para reflejar la interacción, se usan los datos de x_1 y de x_2 para crear una variable $z = x_1x_2$.
3. Ninguno de los procedimientos que agregan o eliminan variables de una en una garantizan que se encuentre el mejor modelo de regresión. Pero estos procedimientos son excelentes para hallar buenos modelos, en especial cuando hay poca multicolinealidad.

Ejercicios

Aplicaciones

16. En un estudio se obtuvieron datos de variables que pueden estar relacionadas con el número de semanas que está desempleado un trabajador de la industria. La variable dependiente de este estudio (semanas) se definió como el número de semanas que un empleado está desempleado debido a despido. En este estudio se usaron las siguientes variables independientes.

Age (edad)	Edad del trabajador
Educ (educación)	Número de años de estudio
Married (casado)	Variable ficticia; 1 si está casado, 0 si no es así
Head (cabeza)	Variable ficticia; 1 si es cabeza de familia, 0 si no es así
Tenure (ocupación)	Número de años en el trabajo anterior
Manager (administrativo)	Variable ficticia; 1 si su ocupación es en administración , 0 si no es así
Sales (ventas)	Variable ficticia; 1 si su ocupación es en ventas, 0 si no es así

Estos datos se encuentran en el archivo Layoffs del disco compacto que se distribuye con el libro.

- a. Obtenga la mejor ecuación estimada de regresión que tenga una variable.
 - b. Emplee el procedimiento por pasos para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to enter* y *Alpha to remove*.
 - c. Use el procedimiento de selección hacia adelante para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to enter*.
 - d. Use el procedimiento de eliminación hacia atrás para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to remove*.
 - e. Use el procedimiento de regresión de los mejores subconjuntos para obtener la mejor ecuación estimada de regresión.
17. La Ladies Professional Golfers Association (LPGA) lleva estadísticas sobre el desempeño y las ganancias de sus miembros en la LPGA Tour. En el archivo titulado LPGATour 2 del disco compacto se presentan las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que tuvieron las mejores ganancias en la LPGA Tour de 2005 (www.lpga.com, 2006). Earnings (ga-

nancias) (miles) son las ganancias totales en miles de dólares en todos los eventos de la gira; Scoring Avg., es la puntuación promedio de la jugadora en todos los eventos; Drive Average es la distancia media alcanzada en el drive por el jugador en yardas; Greens in Reg., es el porcentaje de veces que una jugadora llega al green en regulación; Putting Avg., es el promedio de putts realizados en el green en regulación, y Sand Saves es el porcentaje de veces que la jugadora logra “subir y bajar” (“up and down”) cuando se encuentra en un búnker de arena al lado del green. Sea Drive Greens una nueva variable independiente que represente la interacción entre la distancia media alcanzada en el drive por el jugador y Greens in Reg. Use los métodos de esta sección para obtener la mejor ecuación estimada de regresión múltiple para estimar Scoring Avg de un jugador.

18. Jeff Sagarin proporciona, desde 1985, evaluaciones deportivas para *USA Today*. En el béisbol sus pronósticos RPG (runs/game) estadísticos toman en cuenta todas las estadísticas de ofensiva del jugador y, se asegura, que es la mejor medida del verdadero valor de la ofensiva de un jugador. En los datos que se presentan a continuación se da el RPG y varios estadísticos de ofensiva de la temporada de la Liga Mayor de Béisbol correspondientes a 20 miembros de los Yankees de Nueva York (www.usatoday.com, 3 de marzo de 2006). Los rótulos de las columnas se definen como sigue: RPG, estadístico que predice número de carreras por juego; H, batazos buenos; 2B, dobles; 3B, triples; HR, cuadrangulares; RBI, carreras bateadas; BB, bases por bola; SO, ponchadas; SB, bases robadas; CS, atrapado en robo de base; OBP, porcentaje en base; SLG, porcentaje de potencia de bateo; AVG, promedio de bateo.

Jugador	RPG	H	2B	3B	HR	RBI	BB	SO	SB	CS	OBP	SLG	AVG
D Jeter	6.51	202	25	5	19	70	77	117	14	5	0.389	0.45	0.309
H Matsui	6.32	192	45	3	23	116	63	78	2	2	0.367	0.496	0.305
A Rodriguez	9.06	194	29	1	48	130	91	139	21	6	0.421	0.61	0.321
G Sheffield	6.93	170	27	0	34	123	78	76	10	2	0.379	0.512	0.291
R Cano	5.01	155	34	4	14	62	16	68	1	3	0.32	0.458	0.297
B Williams	4.14	121	19	1	12	64	53	75	1	2	0.321	0.367	0.249
J Posada	5.36	124	23	0	19	71	66	94	1	0	0.352	0.43	0.262
J Giambi	9.11	113	14	0	32	87	108	109	0	0	0.44	0.535	0.271
T Womack	2.91	82	8	1	0	15	12	49	27	5	0.276	0.28	0.249
T Martinez	5.08	73	9	0	17	49	38	54	2	0	0.328	0.439	0.241
M Bellhorn	4.07	63	20	0	8	30	52	112	3	0	0.324	0.357	0.21
R Sierra	3.27	39	12	0	4	29	9	41	0	0	0.265	0.371	0.229
J Flaherty	1.83	21	5	0	2	11	6	26	0	0	0.206	0.252	0.165
B Crosby	3.48	27	0	1	1	6	4	14	4	1	0.304	0.327	0.276
M Lawton	5.15	6	0	0	2	4	7	8	1	0	0.263	0.25	0.125
R Sanchez	3.36	12	1	0	0	2	2	3	0	1	0.326	0.302	0.279
A Phillips	2.13	6	4	0	1	4	1	13	0	0	0.171	0.325	0.15
M Cabrera	1.19	4	0	0	0	0	0	2	0	0	0.211	0.211	0.211
R Johnson	3.44	4	2	0	0	0	1	4	0	0	0.3	0.333	0.222
F Escalona	5.31	4	1	0	0	2	1	4	0	0	0.375	0.357	0.286

Considere que la variable dependiente es la estadística RPG.

- Obtenga la mejor ecuación estimada de regresión con una variable.
 - Emplee los métodos de esta sección para obtener la mejor ecuación estimada de regresión múltiple que estime el RPG de un jugador.
19. Vaya al ejercicio 14. Mediante la edad, la presión sanguínea, si la persona es o no fumadora y cualquier interacción entre estas variables, obtenga una ecuación estimada de regresión que sirva para predecir riesgo. Haga una descripción breve del proceso que utilice para obtener esta ecuación estimada de regresión para estos datos.

16.5

Método de regresión múltiple para el diseño de experimentos

En la sección 15.7 se vio el uso de las variables ficticias en el análisis de regresión múltiple. En esta sección se muestra cómo el uso de variables ficticias en una ecuación de regresión múltiple puede proporcionar otro método para resolver problemas de diseño experimental (o diseño de experimentos). El uso de la regresión múltiple en el diseño experimental se demostrará con el ejemplo del diseño completamente aleatorizado presentado en el capítulo 13.

Recuerde que Chemitech elaboró un nuevo sistema de filtración para el suministro público de agua. Chemitech compraría los componentes del sistema de filtración a diversos proveedores y los armaría en sus instalaciones en Columbia, Carolina del Sur. Se tenían tres métodos de ensamblados, identificados como método A, B y C. Los gerentes de Chemitech deseaban saber qué método de ensamblado producía mayor número de sistemas de filtración por semana.

Se tomó una muestra aleatoria de 15 empleados y cada uno de los tres métodos de ensamblado le fue asignado aleatoriamente a 5 de estos empleados. En la tabla 16.8 se presenta el número de unidades ensambladas por cada empleado. Las medias muestrales del número de unidades producidas con cada uno de los tres métodos son las siguientes:

Método de ensamblado	Número medio producido
A	62
B	66
C	52

Aunque el método B parece ser el que proporciona una tasa de producción más alta, lo que interesa saber es si las tres medias muestrales observadas son suficientemente diferentes como para poder concluir que las medias poblacionales correspondientes a los tres métodos de ensamblado son diferentes.

En el método de regresión aplicado a este problema se empieza por definir las variables ficticias que se usarán para indicar cuál de los métodos de ensamblado fue usado. Como en el problema de Chemitech hay tres métodos de ensamblado, o tratamientos, se necesitan dos variables ficticias. En general, si el factor que se va a investigar tiene k niveles, o tratamientos, se necesita definir $k - 1$ variables ficticias. Para el experimento de Chemitech se definen las variables ficticias A y B de la manera que se muestra en la tabla 16.9.

TABLA 16.8 NÚMERO DE UNIDADES PRODUCIDAS POR LOS 15 TRABAJADORES

Método		
A	B	C
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

TABLA 16.9 VARIABLES FICTICIAS PARA EL EXPERIMENTO DE CHEMITECH

A	B	
1	0	Observación relacionada con el método de ensamblado A
0	1	Observación relacionada con el método de ensamblado B
0	0	Observación relacionada con el método de ensamblado C

Las variables ficticias se pueden usar para relacionar el número de unidades, y , producidas por semana con el método de ensamblado usado por el empleado.

$$\begin{aligned} E(y) &= \text{Valor esperado del número de unidades producidas por semana} \\ &= \beta_0 + \beta_1 A + \beta_2 B \end{aligned}$$

Por tanto, si interesa el valor esperado del número de unidades ensambladas por semana por un empleado mediante el método C, de acuerdo con el procedimiento para asignar valores numéricos a las variables ficticias se tendrá $A = B = 0$. La ecuación de regresión múltiple se reduce entonces a

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

La interpretación es que β_0 se puede interpretar como el valor esperado de la cantidad de unidades ensambladas por semana por un empleado que use el método C. En otras palabras, β_0 es la media de la cantidad de unidades ensambladas por semana mediante el método C.

A continuación se considera la forma de la ecuación de regresión múltiple correspondiente a cada uno de los otros métodos. Los valores de las variables ficticias correspondientes al método A son $A = 1$ y $B = 0$, y entonces

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

Los correspondientes al método B son $A = 0$ y $B = 1$, y entonces

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Como se ve, $\beta_0 + \beta_1$ representa la media del número de unidades ensambladas por semana mediante el método A, y $\beta_0 + \beta_2$ representa la media del número de unidades ensambladas por semana con el método B.

Ahora se desea obtener estimaciones para los coeficientes β_0 , β_1 y β_2 y, de esta manera, obtener una estimación del número medio de unidades ensambladas por semana con cada uno de los métodos. En la tabla 16.10 se presentan los datos muestrales que consisten en 15 observaciones de A, B y y . En la figura 16.18 se presentan los resultados correspondientes obtenidos usando la regresión múltiple de Minitab. Como se ve, las estimaciones de β_0 , β_1 y β_2 son $b_0 = 52$, $b_1 = 10$ y $b_2 = 14$. De esta manera, las mejores estimaciones de las medias del número de unidades ensambladas por semana con cada uno de los métodos de ensamblado son:

Método de ensamblado	Estimación de $E(y)$
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 + b_2 = 52 + 14 = 66$
C	$b_0 = 52$

Observe que estas estimaciones, de los números medios de unidades producidas con cada uno de estos tres métodos de ensamblado, obtenidas mediante el análisis de regresión son las mismas que las medias muestrales presentadas previamente.

TABLA 16.10 DATOS PARA EL DISEÑO COMPLETAMENTE ALEATORIZADO DE CHEMITECH

A	B	y
1	0	58
1	0	64
1	0	55
1	0	66
1	0	67
0	1	58
0	1	69
0	1	71
0	1	64
0	1	68
0	0	48
0	0	57
0	0	59
0	0	47
0	0	49

Ahora se va a ver cómo usar los resultados del análisis de regresión múltiple para realizar la prueba ANOVA de la diferencia entre las medias de estos tres métodos. Primero, se observa que si las medias no difieren

$$E(y) \text{ para el método A} - E(y) \text{ para el método C} = 0$$

$$E(y) \text{ para el método B} - E(y) \text{ para el método C} = 0$$

Como β_0 es igual a $E(y)$ al emplear el método C y $\beta_0 + \beta_1$ es igual a $E(y)$ al emplear el método A, la primera diferencia es igual a $(\beta_0 + \beta_1) - \beta_0 = \beta_1$. Y como $\beta_0 + \beta_2$ es igual a $E(y)$ al emplear el método B, la segunda diferencia es igual a $(\beta_0 + \beta_2) - \beta_0 = \beta_2$. Se concluye que entre

FIGURA 16.18 RESULTADOS DE LA REGRESIÓN MÚLTIPLE PARA EL DISEÑO COMPLETAMENTE ALEATORIZADO DE CHEMITECH

The regression equation is y = 52.0 + 10.0 A + 14.0 B
Predictor Coef SE Coef T P
Constant 52.000 2.380 21.84 0.000
A 10.000 3.367 2.97 0.012
B 14.000 3.367 4.16 0.001
S = 5.32291 R-Sq = 60.5% R-Sq(adj) = 53.9%
Analysis of Variance
SOURCE DF SS MS F P
Regression 2 520.00 260.00 9.18 0.004
Residual Error 12 340.00 28.33
Total 14 860.00

los tres métodos no hay diferencia si $\beta_1 = 0$ y $\beta_2 = 0$. Por tanto, la hipótesis nula en una prueba para la diferencia entre las medias se puede expresar como

$$H_0 : \beta_1 = \beta_2 = 0$$

Tome el nivel de significancia $\alpha = 0.05$. Recuerde que para probar este tipo de hipótesis nula acerca de la significancia de la relación de regresión se emplea la prueba F de significancia general. En el resultado de Minitab que se presenta en la figura 16.18 se observa que el valor- p correspondiente a $F = 9.18$ es 0.004. Como valor- $p = 0.004 < \alpha = 0.05$, $H_0 : \beta_1 = \beta_2 = 0$ se rechaza y se concluye que las medias de los tres métodos de ensamblado no son iguales. Como la prueba F indica que la relación de regresión múltiple es significativa, se puede realizar una prueba t para determinar la significancia de cada uno de los parámetros, β_1 y β_2 . Con $\alpha = 0.05$, los valores- p , 0.012 y 0.001, que aparecen en los resultados de Minitab, indican que las hipótesis nulas $H_0 : \beta_1 = 0$ y $H_0 : \beta_2 = 0$, se pueden rechazar. Por tanto, ambos parámetros son estadísticamente significativos. Así, se concluye que las medias de los parámetros A y C son diferentes y que también las medias de los parámetros B y C son diferentes.

Ejercicios

Métodos

- Autoexamen**
- 20. Considere un diseño completamente aleatorizado en el que haya cuatro tratamientos: A, B, C y D. Escriba la ecuación de regresión múltiple que sirva para analizar estos datos. Defina todas las variables.
 - 21. Dé una ecuación de regresión múltiple que sirva para analizar los datos de un diseño de bloque aleatorizado que tenga tres tratamientos y dos bloques. Defina todas las variables.
 - 22. Dé una ecuación de regresión múltiple que sirva para analizar los datos de un diseño bifactorial que tenga dos niveles para el factor A y tres niveles para el factor B. Defina todas las variables.

Aplicaciones

- Autoexamen**
- 23. La empresa Jacobs Chemical desea estimar el tiempo promedio (en minutos) necesario para mezclar un lote de un material empleando máquinas provenientes de tres fabricantes diferentes. Para limitar los costos de la prueba se mezclaron cuatro lotes de material en las máquinas producidas por cada uno de los fabricantes. A continuación se presentan los tiempos requeridos.

Fabricante 1	Fabricante 2	Fabricante 3
20	28	20
26	26	19
24	31	23
22	27	22

- a. Dé una ecuación de regresión múltiple que sirva para analizar estos datos.
 - b. Dé las mejores estimaciones de los coeficientes en su ecuación.
 - c. En términos de los coeficientes de las ecuaciones de regresión cuáles son las hipótesis a probar para ver si los tiempos son iguales con las máquinas de los tres fabricantes.
 - d. ¿Cuál es la conclusión que se obtiene con un nivel de significancia 0.05?
- 24. En la publicidad de cuatro pinturas diferentes se asegura que todas tienen el mismo tiempo de secado. Para comprobar esto se probaron cinco muestras de cada pintura.

Los tiempos de secado de cada muestra se presentan a continuación

Pintura 1	Pintura 2	Pintura 3	Pintura 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

- a. Use $\alpha = 0.05$ para probar si existe una diferencia significativa entre los tiempos de secado.
- b. Dé una estimación del tiempo medio de secado de la pintura 2. ¿Cómo se obtiene de los resultados de un paquete de software?
25. Un comerciante de automóviles realizó una prueba para determinar si el tiempo requerido para ajustar un motor dependía de si se empleaba un analizador computarizado o un analizador electrónico. Como el tiempo que se necesita para ajustar un motor depende de si se trata de un auto pequeño, mediano o grande, se usaron los tres tipos de automóviles como bloques del experimento. Los datos que se obtuvieron (en minutos) son los que se presentan a continuación.

		Automóvil		
		Pequeño	Mediano	Grande
Analizador	Computarizado	50	55	63
	Electrónico	42	44	46

- Emplee $\alpha = 0.05$ para probar si hay diferencias significativas.
26. Una empresa de ventas por catálogo diseñó un experimento factorial para probar los efectos del tamaño de un anuncio publicitario y su diseño sobre el número (en miles) de catálogos solicitados. Se consideraron tres diseños y dos tamaños diferentes del anuncio publicitario. De éstos se obtuvieron los datos siguientes. Pruebe si hay efectos significativos debido al diseño, al tamaño o a interacciones. Use $\alpha = 0.05$.

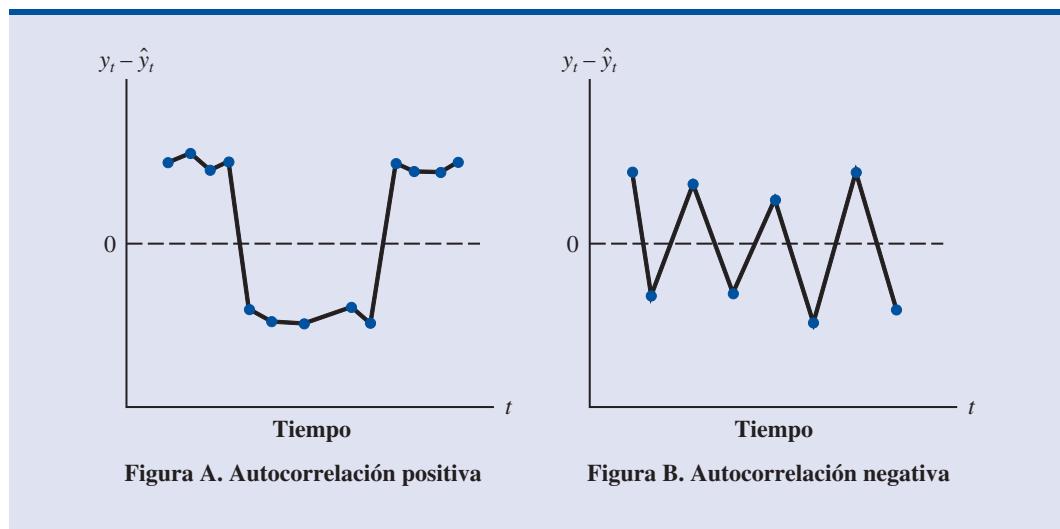
		Tamaño del anuncio publicitario	
		Pequeño	Grande
Diseño	A	8 12	12 8
	B	22 14	26 30
C	C	10 18	18 14

16.6

Autocorrelación y la prueba de Durbin-Watson

En los negocios y en la economía suele ocurrir que los datos que se usan en estudios de regresión estén correlacionados a lo largo del tiempo. No es raro que el valor de y en el periodo t , que se denota y_t , esté relacionado con el valor de y en un periodo anterior. En tales casos se dice que existe **autocorrelación** en los datos (o **correlación serial**). Si el valor de y en el periodo t está rela-

FIGURA 16.19 DOS CONJUNTOS DE DATOS CON CORRELACIÓN DE SEGUNDO ORDEN



cionado con su valor en el periodo $t - 1$, existe correlación de primer orden. Si el valor de y en el periodo t está relacionado con su valor en el periodo $t - 2$, existe correlación de segundo orden y así sucesivamente.

Cuando hay autocorrelación se viola una de las suposiciones del modelo de regresión: los términos del error no son independientes. En el caso de la autocorrelación de primer orden, el error en el periodo t que se denota ϵ_t , estará relacionado con el error en el periodo $t - 1$, que se denota ϵ_{t-1} . En la figura 16.19 se ilustran dos casos de autocorrelación de primer orden. En la gráfica A se presenta el caso de una autocorrelación positiva; en la gráfica B el de una autocorrelación negativa. En la autocorrelación positiva se espera que el residual positivo de un periodo vaya seguido de un residual positivo en el periodo siguiente, que el residual negativo de un periodo vaya seguido de un residual negativo en el periodo siguiente y así sucesivamente. En la autocorrelación negativa se espera que el residual positivo de un periodo vaya seguido de un residual negativo en el periodo siguiente, después un residual positivo y así sucesivamente.

Si existe autocorrelación, se pueden cometer errores serios cuando se realizan pruebas de significancia estadística basadas en el modelo de regresión supuesto. Por tanto, es importante poder detectar la autocorrelación y tomar medidas correctivas. A continuación se mostrará cómo usar el estadístico de Durbin-Watson para detectar autocorrelación de primer orden.

Suponga que los valores de ϵ no sean independientes sino que estén relacionados de la manera siguiente:

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \quad (16.16)$$

donde ρ es un parámetro cuyo valor absoluto es menor que 1 y z_t es una variable aleatoria distribuida normal e independientemente, que tienen media cero y varianza σ^2 . En la ecuación 16.16 se ve que si $\rho = 0$, los términos del error no están relacionados y cada uno tiene media cero y varianza σ^2 . En este caso no hay autocorrelación y se satisfacen las suposiciones de la regresión. Si $\rho > 0$, existe autocorrelación positiva; si $\rho < 0$, existe autocorrelación negativa. En cualquiera de estos casos, se violan las suposiciones de la regresión acerca del término del error.

En la prueba de Durbin-Watson para autocorrelación se usan los residuales para determinar si $\rho = 0$. Para simplificar la notación para el estadístico de Durbin-Watson el residual i se denota $e_i = y_i - \hat{y}_i$. El estadístico de prueba Durbin-Watson se calcula como sigue.

ESTADÍSTICO DE PRUEBA DURBIN-WATSON

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (16.17)$$

Si valores sucesivos de los residuales se encuentran cercanos unos de otros (autocorrelación positiva), el valor del estadístico de prueba Durbin-Watson será pequeño. Si valores sucesivos de los residuales se encuentran alejados unos de otros (autocorrelación negativa), el valor del estadístico de prueba Durbin-Watson será grande.

El estadístico de prueba Durbin-Watson va de cero a cuatro, si su valor es dos, esto indica que no existe autocorrelación. Durbin y Watson elaboraron tablas para determinar cuándo su estadístico indica la existencia de autocorrelación. En la tabla 16.11 se presentan límites inferiores y superiores (d_L y d_U) para las pruebas de hipótesis con $\alpha = 0.05$; n denota el número de observaciones. Siempre, la hipótesis nula a probar es que no existe autocorrelación

$$H_0: \rho = 0$$

La hipótesis alternativa que se prueba en la autocorrelación positiva es

$$H_a: \rho > 0$$

La hipótesis alternativa que se prueba en la autocorrelación negativa es

$$H_a: \rho < 0$$

TABLA 16.11 VALORES CRÍTICOS PARA LA PRUEBA DE DURBIN-WATSON PARA AUTOCORRELACIÓN

Nota: Los valores que se presentan en la tabla son los valores críticos para la prueba de Durbin-Watson de una cola para autocorrelación. En pruebas de dos colas, se duplica el nivel de significancia.

Puntos de significancia de d_L y d_U : $\alpha = 0.05$
Número de variables independientes

n^*	1		2		3		4		5	
	d_L	d_U								
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

* Para valores intermedios de n , interpolar linealmente.

FIGURA 16.20 PRUEBA DE HIPÓTESIS PARA AUTOCORRELACIÓN MEDIANTE LA PRUEBA DE DURBIN-WATSON

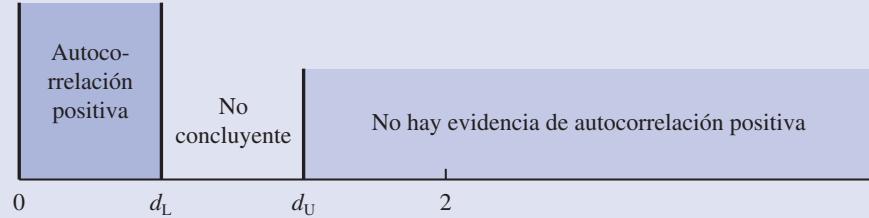


Diagrama A: Prueba para autocorrelación positiva

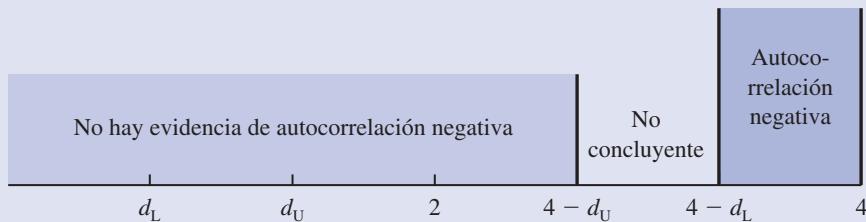


Diagrama B: Prueba para autocorrelación negativa

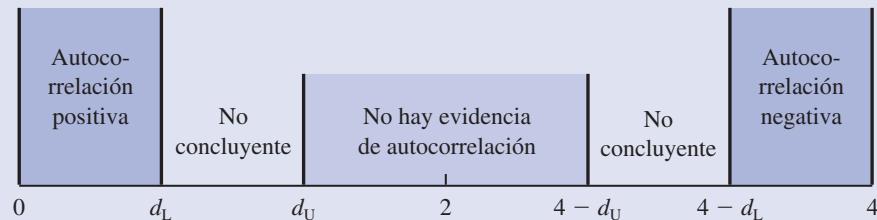


Diagrama C: Prueba para autocorrelación de dos colas

También se puede hacer una prueba de dos colas. En este caso la hipótesis alternativa es

$$H_a: \rho \neq 0$$

En la figura 16.20 se muestra el uso de los valores d_L y d_U de la tabla 16.11 para probar si existe autocorrelación. En el diagrama A se ilustra la prueba para autocorrelación positiva: Si $d < d_L$, se concluye que existe correlación positiva. Si $d_L \leq d \leq d_U$, se dice que la prueba no es concluyente. Si $d > d_U$, se concluye que no hay evidencia de autocorrelación positiva.

En el diagrama B se ilustra la prueba para autocorrelación negativa. Si $d > 4 - d_L$, se concluye que existe autocorrelación negativa. Si $4 - d_U \leq d \leq 4 - d_L$, se dice que la prueba no es concluyente. Si $d < 4 - d_U$, se concluye que no existe evidencia de autocorrelación negativa.

El diagrama C ilustra la prueba de dos colas. Si $d < d_L$ o $d > 4 - d_L$, se rechaza H_0 y se concluye que existe correlación. Si $d_L \leq d \leq d_U$ o si $4 - d_U \leq d \leq 4 - d_L$, se dice que la prueba no es concluyente. Si $d_U < d < 4 - d_U$, se concluye que no hay evidencia de autocorrelación.

Si se determina que hay una autocorrelación significativa, se debe verificar si se omitieron una o varias variables independientes importantes que tengan un efecto de orden temporal sobre la variable dependiente. Si no se encuentran tales variables, incluir una variable independiente que mida el tiempo en el que se hace la observación (el valor de esta variable, puede ser, por ejemplo, 1 para la primera observación, 2 para la segunda, etc.) algunas veces ayuda para eliminar o reducir la autocorrelación. Cuando no funcionan estos intentos para eliminar o reducir la autocorrelación, hacer transformaciones de las variables independientes resulta útil; un estudio sobre esas transformaciones puede encontrarse en libros más avanzados sobre análisis de regresión.

Observe que en las tablas de Durbin-Watson el menor valor para el tamaño de la muestra es 15. La razón es que para muestras menores, la prueba suele ser no concluyente; en realidad, se suele creer que el tamaño de la muestra debe ser de por lo menos 50 para que con la prueba se obtengan resultados que valgan la pena.

Ejercicios

Aplicaciones

27. En los datos siguientes se muestran los precios diarios de cierre (en dólares por acción) de IBM desde el 3 de noviembre de 2005, hasta el 1 de diciembre de 2005 (*Compustat*, 26 de febrero de 2006).

Fecha	Precio (\$)
Nov. 3	82.87
Nov. 4	83.00
Nov. 7	83.61
Nov. 8	83.15
Nov. 9	82.84
Nov. 10	83.99
Nov. 11	84.55
Nov. 14	84.36
Nov. 15	85.53
Nov. 16	86.54
Nov. 17	86.89
Nov. 18	87.77
Nov. 21	87.29
Nov. 22	87.99
Nov. 23	88.80
Nov. 25	88.80
Nov. 28	89.11
Nov. 29	89.10
Nov. 30	88.90
Dic. 1	89.21



- Defina la variable independiente Periodo, donde Periodo = 1 corresponda al dato del 3 de noviembre, Periodo = 2 corresponda al dato del 4 de Noviembre, etc. Obtenga una ecuación estimada de regresión que sirva para predecir el precio del cierre dado el valor del Periodo.
 - Emplee como nivel de significancia 0.05 y pruebe si existe autocorrelación positiva en estos datos.
28. Remítase al conjunto de datos de Craven de la tabla 16.5. En la sección 16.3 se mostró que el coeficiente de determinación ajustado de la ecuación estimada de regresión que contenía Accounts-(Cuentas), AdvExp (GastPubl), Poten y Share (Participación) era 88.1%. Use 0.05 como nivel de significancia y aplique la prueba de Durbin-Watson para determinar si existe autocorrelación positiva.

Resumen

En este capítulo se analizaron varios de los conceptos que se usan en la construcción de modelos para hallar la ecuación estimada de regresión. Primero se presentó el concepto de modelo lineal general para mostrar cómo pueden extenderse los métodos estudiados en los capítulos 14 y 15 a las relaciones curvilíneas y a los efectos de interacción. Después, se vio cómo emplear transformaciones a la variable dependiente cuando se presentan problemas como el de una varianza no constante en los términos del error.

En muchas aplicaciones del análisis de regresión se emplea un gran número de variables independientes. Para agregar o eliminar variables a un modelo de regresión se vio un método general basado en el estadístico F . Después se presentó un problema más grande en el que se tenían 25 observaciones y ocho variables independientes. También se vio que cuando se tienen problemas más grandes, uno de los asuntos a resolver es hallar el mejor subconjunto de variables independientes; para esto existen varios procedimientos de selección de variables: regresión por pasos, selección hacia adelante, eliminación hacia atrás y regresión de los mejores subconjuntos.

En la sección 16.5 se amplió el estudio para ver cómo obtener modelos de regresión múltiple que sirven como otro método para la solución de problemas de análisis de varianza y de diseño de experimentos. El capítulo concluyó con una aplicación del análisis de residuales mediante la prueba de Durbin-Watson para autocorrelación.

Glosario

Modelo lineal general Modelo de la forma $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon$, en donde cada una de las variables independientes z_j ($j = 1, 2, \dots, p$) es función de x_1, x_2, \dots, x_k , las variables para las que se han recolectado datos.

Interacción Efecto de dos variables independientes cuando actúan juntas.

Procedimientos de selección de variables Métodos para la selección de un subconjunto de variables independientes para un modelo de regresión.

Autocorrelación Correlación en los errores, que se presenta cuando los términos del error pertenecientes a puntos sucesivos de tiempo están relacionados.

Correlación serial Es lo mismo que autocorrelación.

Prueba de Durbin-Watson Prueba para determinar si existe autocorrelación de primer orden.

Fórmulas clave

Modelo lineal general

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \quad (16.1)$$

Estadístico de prueba F para agregar o eliminar $p - q$ variables

$$F = \frac{\frac{\text{SCE}(x_1, x_2, \dots, x_q) - \text{SCE}(x_1, x_2, \dots, x_q, x_{q+1}, \dots, x_p)}{p - q}}{\frac{\text{SCE}(x_1, x_2, \dots, x_q, x_{q+1}, \dots, x_p)}{n - p - 1}} \quad (16.13)$$

Autocorrelación de primer orden

$$\epsilon_t = \rho \epsilon_{t-1} + z_t \quad (16.16)$$

Estadístico de prueba de Durbin-Watson

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (16.17)$$

Ejercicios complementarios

29. La disminución en los precios de las impresoras láser a color, hacen de ellas una muy buena alternativa frente a las impresoras de inyección de tinta. *PC World* examinó y evaluó 10 impresoras a color. En los datos siguientes se presentan el precio, la velocidad de impresión en páginas por minuto (ppm) y la evaluación de *PC World* de estas impresoras (*PC World*, diciembre de 2005).

Fabricante y modelo	Velocidad (ppm)	Evaluación
Dell 3000cn	3.4	83
Oki Data C5200n	5.2	81
Konica Minolta MagiColor 2430DL	2.7	79
Brother HL-2700CN	3.1	78
Lexmark C522n	3.8	77
HP Color LaserJet 3600n	5.6	74
Xerox Phaser 6120n	1.6	73
Konica Minolta MagiColor 2450	1.6	71
HP Color LaserJet 2600n	2.6	70
HP Color LaserJet 2550L	1.1	61

- a. Elabore un diagrama de dispersión, use como variable independiente la velocidad de impresión. ¿Un modelo de regresión simple parece apropiado?
- b. Obtenga una ecuación estimada de regresión múltiple en la que las variables independientes sean x = velocidad y x^2 .
- c. Considere el modelo no lineal indicado por la ecuación (16.7). Use logaritmos para transformar este modelo no lineal en un modelo lineal equivalente y obtenga la ecuación estimada de regresión correspondiente. ¿Esta ecuación estimada de regresión proporciona un mejor ajuste que la ecuación estimada de regresión obtenida en el inciso b?
30. Muchos fondos internacionales ofrecen tasas más razonables que en Estados Unidos. Como los mercados internacionales suelen moverse en direcciones distintas a los mercados de Estados Unidos, las inversiones en mercados extranjeros pueden reducir el riesgo de un inversionista. En la tabla siguiente se presentan 20 fondos internacionales dando tipo de fondo (con comisión o sin comisión), coeficiente de gastos (%), seguridad (0 = la más riesgosa, 10 = la más segura) y su desempeño en un año al 10 de diciembre de 1999 (*Mutual Funds*, febrero de 2000).

Tipo de fondo	Coeficiente de gastos (%)	Seguridad	Desempeño (%)
ABN AMRO Int'l Equity "Com"	Sin comisión	1.38	6.9
Accessor Int'l Equity "Adv"	Sin comisión	1.59	7.1
Artisan International	Sin comisión	1.45	6.8
Columbia Int'l Stock	Sin comisión	1.56	7.1
Concert Inv. "A" Int'l Equity	Con comisión	2.16	6.3
Diversified Invstr Int'l Eqty	Sin comisión	1.40	7.3

(continúa)

	Tipo de fondo	Coeficiente de gastos (%)	Seguridad	Desempeño (%)
Driehaus Int'l Growth	Sin comisión	1.88	6.5	92
Founders Passport	Sin comisión	1.52	7.0	86
Guardian Baillie Fifford Int'l "A"	Con comisión	1.62	7.1	37
Jamestown Int'l Equity	Sin comisión	1.56	7.1	35
Julius Baer Int'l Equity	Sin comisión	1.79	6.9	71
Aetna "I" Int'l	Sin comisión	1.35	7.3	46
Pilgrim Int'l Value "A"	Con comisión	1.80	7.1	42
Fidelity Diversified Int'l	Sin comisión	1.48	7.5	42
Putnam "A" Int'l Growth	Con comisión	1.59	6.9	55
Sit Int'l Growth	Sin comisión	1.50	6.9	49
Touchstone Int'l Equity "A"	Con comisión	1.60	7.5	35
United Int'l Growth "A"	Con comisión	1.28	7.1	47
Vontobel Int'l Equity	Sin comisión	1.50	7.0	43
Waddell & Reed Int'l Growth "B"	Con comisión	2.46	7.0	75

- a. Utilice los métodos de este capítulo para obtener una ecuación estimada de regresión que sirva para estimar el desempeño de un fondo con base en los datos proporcionados.
- b. ¿La ecuación estimada de regresión obtenida en el inciso a proporciona un buen ajuste? Explique.
- c. Acorn International es un fondo sin comisión cuyo coeficiente de gastos es 1.12% y cuya seguridad es 7.6. Use la ecuación estimada de regresión obtenida en el inciso a para estimar el desempeño en un año de Acorn International.
31. En un estudio se investigó la relación entre el retraso en la auditoría, tiempo transcurrido desde el fin del año fiscal de una empresa hasta la fecha del informe del auditor, y variables que describen al cliente y al auditor. A continuación se presentan algunas de las variables independientes incluidas en este estudio.

Industria	Variable ficticia que toma el valor 1 si se trata de una industria y 0 si se trata de un banco, de una institución de ahorro, de préstamo o de seguros.
Pública	Variable ficticia que toma el valor 1 si la empresa es comercializada en la bolsa o es extra bursátil; 0 si no es así.
Calidad	Medida de la calidad general de los controles internos, a juicio del auditor, con una escala de cinco puntos que van desde "prácticamente ninguna" (1) hasta "excelente" (5).
Terminado	Una medida que va de 1 a 4, a juicio del auditor, donde 1 indica "todo el trabajo realizado después del fin de año" y 4 indica "la mayor parte del trabajo realizado antes del fin de año".

En una muestra de 40 empresas se obtuvieron los datos siguientes.

Retraso	Industria	Pública	Calidad	Terminado
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3

Retraso	Industria	Pública	Calidad	Terminado
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2
75	1	1	4	3
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
80	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

- a. Obtenga una ecuación estimada de regresión con todas las variables independientes.
- b. ¿La ecuación estimada de regresión obtenida en el inciso a proporciona un buen ajuste?
- c. Trace un diagrama de dispersión en el que se presente la variable retraso en función de la variable terminado.
- d. Con base en sus observaciones acerca de la relación entre retraso y terminado obtenga otra ecuación estimada de regresión, distinta a la dada en el inciso a, que explique la mayor proporción posible de la variabilidad de retraso.
32. Remítase a los datos del ejercicio 31. Considere un modelo en el que para predecir retraso se use únicamente industria. Emplee como nivel de significancia 0.01 y pruebe si existe alguna autocorrelación en los datos.
33. Remítase a los datos del ejercicio 31.
 - a. Obtenga una ecuación estimada de regresión para predecir retraso empleando industria y calidad.
 - b. Grafique los residuales obtenidos con la ecuación estimada de regresión obtenida en el inciso a en función del orden en que están presentados los datos. ¿Parece existir alguna autocorrelación en los datos? Explique.
 - c. Con un nivel de significancia 0.05, pruebe si existe alguna autocorrelación en los datos.
34. Se realizó un estudio para investigar la actividad de los compradores cuando buscan y miran cosas dentro de una tienda, y de acuerdo con esto se les clasificó como inactivos, poco activos y muy activos. También se midió qué tan cómodo se sentía cada comprador en la tienda; puntuaciones más altas correspondían a mayor comodidad. Los datos siguientes provienen de este estudio. Emplee como nivel de significancia 0.05 y realice una prueba para determinar las diferencias que existen en la comodidad dentro de la tienda entre los tres tipos de compradores.

archivo en CD
Browsing

	Inactivos	Poco activos	Muy activos
	4	5	5
	5	6	7
	6	5	5
	3	4	7
	3	7	4
	4	4	6
	5	6	5
	4	5	7

35. La revista *Money* publicó precios y diversos datos de los 418 vehículos más populares entre los modelos del 2003. Una de estas variables fue el valor de reventa del vehículo, expresado como porcentaje del precio de reventa sugerido por el fabricante. Estos datos se clasificaron de acuerdo con el tamaño y tipo de vehículo. En la tabla siguiente se presentan los valores de reventa de 10 automóviles pequeños elegidos aleatoriamente, de 10 automóviles medianos elegidos aleatoriamente, de 10 automóviles de lujo elegidos aleatoriamente y de 10 automóviles deportivos elegidos aleatoriamente (*Money*, marzo de 2003).

archivo en CD
Resale

	Pequeño	Mediano	De lujo	Deportivo
	26	26	36	41
	31	29	38	39
	41	41	38	30
	32	27	39	34
	27	26	35	40
	34	33	26	43
	31	27	40	42
	38	29	47	39
	27	35	41	44
	42	39	32	50

Use $\alpha = 0.05$ para determinar si existe alguna diferencia significativa entre los valores medios de reventa de los cuatro tipos de automóviles.

Caso problema 1 Análisis de las estadísticas de la PGA Tour

archivo en CD
PGATour

La Professional Golfers Association (PGA) lleva un registro sobre ganancias y datos de desempeño de sus miembros en la PGA Tour. En el archivo PGA Tour del disco compacto se presentan los datos de fin de año sobre el desempeño de los 125 jugadores que tuvieron los mejores ingresos en los eventos de la PGA Tour de 2005 (www.pgatour.com, 2006). Cada renglón del conjunto de datos corresponde a un jugador de la PGA Tour, y los datos han sido ordenados con base en las ganancias totales. A continuación se presenta la descripción de los datos.

Earnings	Ganancias totales en los eventos de la PGA Tour
Scoring Avg.	Puntuación promedio de un jugador en todos los eventos
Yards/Drive	Promedio de yardas por salto p. 740
Driving Acc.	Porcentaje de veces que el jugador llega a la calle con un tee shot
Greens in Reg.	Porcentaje de veces que el jugador llega al green en regulación se considera como golpe a un green en regulación, si cualquier parte de la bola está tocando la superficie y la diferencia entre el valor del par para el hoyo y el número de los movimientos para golpear el green es por lo menos 2.

Putting Avg.	Promedio de putts (toques) realizados en el green en regulación
Save Pct.	Porcentaje de veces que el jugador logra “subir y bajar” (“up and down”) cuando se encuentra en un búnker de arena al lado del green

Informe administrativo

Suponga que un representante de la PGA Tour lo contrata para analizar los datos para una presentación que se realizará en la reunión anual de la PGA Tour. Este representante le pregunta si es posible usar estos datos para determinar una medida del desempeño que sea el mejor predictor de la puntuación promedio de un jugador. Use los métodos presentados en este capítulo y en los capítulos anteriores para analizar estos datos. Formule un informe para el representante de la PGA Tour en el que resuma su análisis y en el que incluya los resultados estadísticos más importante, sus conclusiones y recomendaciones. En un apéndice presente todo el material técnico que considere adecuado.

Caso problema 2 Rendimiento de combustible en los automóviles

archivo
en
CD
Cars

En todos los automóviles nuevos que se venden en Estados Unidos, viene una etiqueta sobre el consumo de combustible indicando el rendimiento en millas por galón que se espera del automóvil tanto en ciudad como en carretera. En la *Fuel Economy Guide* del Departamento de Energía de Estados Unidos se encuentra esta información para cualquier automóvil o camión. En el archivo Cars del disco compacto que se distribuye con el libro se encuentra parte de estos datos para 230 automóviles (www.fueleconomy.gov, 21 de marzo de 2003). A continuación se presenta una descripción de los datos que vienen en el disco compacto.

Class	Tipo de automóvil (compacto, mediano, grande)
Manufacturer	Empresa fabricante del automóvil
carline name	Nombre del automóvil
displ	Desplazamiento del motor en litros
cyl	Cilindros que tiene el motor (4, 6, 8)
trans	Tipo de transmisión (automática, manual)
cty	Consumo de combustible en la ciudad en millas por galón
hwy	Consumo de combustible en carretera en millas por galón

Informe administrativo

Emplee los métodos presentados en este capítulo y en los anteriores y analice este conjunto de datos. El objetivo es obtener una ecuación estimada de regresión que sirva para estimar el consumo de combustible en la ciudad y una ecuación estimada de regresión que sirva para estimar el consumo de combustible en carretera. De su análisis, presente un resumen, en el que incluya los resultados estadísticos más importantes, sus conclusiones y recomendaciones. En un apéndice incluya cualquier material técnico que considere adecuado (resultados de computadora, gráficas de residuales, etc.).

Caso problema 3 Predicción de las tasas de alumnos que llegan a titularse en las universidades

Para los administradores universitarios, el porcentaje de alumnos que ingresan a una universidad y que llegan hasta su titulación es un dato estadístico importante. Algunos de los factores que están relacionados con el porcentaje de alumnos que llegan hasta la titulación es el porcentaje de

clases en las que hay menos de 20 alumnos, el porcentaje de clases en las que hay más de 50 estudiantes, la proporción de estudiantes por facultad, el porcentaje de estudiantes que solicitan ingresar a la universidad y que son admitidos, el porcentaje de estudiantes de primer ingreso que estuvieron en el 10% más alto de sus clases de bachillerato y la reputación académica de la universidad. Para estudiar el efecto de estos factores sobre el porcentaje de alumnos que llegan a la titulación, se recolectaron datos de 48 universidades de Estados Unidos (*America's Best Colleges*, Edición del año 2000). Estos datos se encuentran en el archivo GradeRate del disco compacto. A continuación se presenta una descripción de los datos que aparecen en el disco.



Region	Región del país en donde se encuentra la universidad
Graduation Rate	Porcentaje de estudiantes que entran a la universidad y que se titulan
% of classes under 20	Porcentaje de las clases en las que hay menos de 20 alumnos
% of classes of 50 or more	Porcentaje de las clases en las que hay más de 50 alumnos
Student-Faculty Ratio	Cociente del número de estudiantes inscritos dividido entre el número de profesores
Acceptance rate	Porcentaje de estudiantes que solicitan inscripción a la universidad y que son aceptados
1st-Year students in top 10% of HS class	De los estudiantes admitidos a la universidad, porcentaje que estuvo en el 10% más alto de sus clases de bachillerato
Academic Reputation Score	Una medida de la reputación de la universidad determinada mediante una revisión a los administradores en otras universidades: medida en una escala del 1 (marginal) al 5 (distinguida)

Informe administrativo

Use los métodos presentados en este capítulo y en los capítulos anteriores para analizar este conjunto de datos. Presente un resumen de su análisis en el que dé los principales resultados estadísticos, sus conclusiones y recomendaciones, en un informe administrativo. En un apéndice presente cualquier material técnico (resultados de computadora, gráficas de residuales, etc.) que considere adecuados.

Apéndice 16.1 Procedimientos de selección de variables con Minitab

En la sección 16.4 se vio el uso de los procedimientos de selección de variables para la solución de problemas de regresión múltiple. En la figura 16.16 se mostraron los resultados que da la regresión por pasos de Minitab aplicada a los datos de Cravens y en la figura 16.17 los resultados que da el procedimiento de los mejores subconjuntos de Minitab. En este apéndice se describen los pasos necesarios para obtener los resultados que se muestran en esas dos figuras, así como los pasos que se requieren en los procedimientos de selección hacia adelante y eliminación hacia atrás. Primero, en una hoja de cálculo de Minitab se ingresan los datos de la tabla 16.5. Los valores de Sales, Time, Poten, AdvExp, Share, Change, Accounts y Rating se ingresan en las columnas C1-C9 de la hoja de cálculo de Minitab.

Uso del procedimiento por pasos de Minitab

Mediante los pasos siguientes se obtienen los resultados de la regresión por pasos de Minitab para los datos de Cravens.

- Paso 1.** Seleccionar el menú Stat
- Paso 2.** Seleccionar el menú Regression
- Paso 3.** Elegir Stepwise

Paso 4. Cuando aparezca el cuadro de diálogo **Stepwise Regression:**Ingresar Sales en el cuadro de diálogo **Response**Ingresar Time, Poten, AdvExp, Share, Change, Accounts y Rating en el cuadro **Predictors**Seleccionar el botón **Methods****Paso 5.** Cuando aparezca el cuadro de diálogo **Stepwise Method:**Seleccionar **Stepwise (forward and backward)**Ingresar 0.05 en el cuadro **Alpha to enter**Ingresar 0.05 en el cuadro **Alpha to remove**Clic en **OK****Paso 6.** Cuando aparezca el cuadro de diálogo **Stepwise Regression:**Clic en **OK****Uso del procedimiento de selección hacia adelante de Minitab**

Para usar el procedimiento de selección hacia adelante de Minitab, sólo hay que modificar el paso 5 del procedimiento de regresión por pasos, como se indica a continuación:

Paso 5. Cuando aparezca el cuadro de diálogo **Stepwise-Methods:**Seleccionar **Forward Selection**Ingresar 0.05 en el cuadro de diálogo **Alpha to enter**Clic en **OK****Uso del procedimiento de eliminación hacia atrás de Minitab**

Para usar el procedimiento de eliminación hacia atrás de Minitab, sólo hay que modificar el paso 5 del procedimiento de regresión por pasos, como se indica a continuación:

Paso 5. Cuando aparezca el cuadro de diálogo **Stepwise-Methods:**Seleccionar **Backward elimination**Ingresar 0.05 en el cuadro de diálogo **Alpha to remove**Clic en **OK****Uso del procedimiento de los mejores subconjuntos de Minitab**

Mediante los pasos siguientes se obtienen los resultados para los datos de Cravens que da la regresión de los mejores subconjuntos de Minitab.

Paso 1. Seleccionar el menú **Stat****Paso 2.** Seleccionar el menú **Regression****Paso 3.** Elegir **Best Subsets****Paso 4.** Cuando aparezca el cuadro de diálogo **Best Subsets Regression**Ingresar Sales en el cuadro **Response**Ingresar Time, Poten, AdvExp, Share, Change, Accounts y Rating en el cuadro **Predictors**Clic en **OK**