CAPÍTULO 6

Análisis de regresión múltiple: temas adicionales

En este capítulo se reúnen varios temas sobre el análisis de regresión múltiple que no pudieron ser vistos de forma adecuada en los capítulos anteriores. Estos temas no son tan fundamentales como el material visto en los capítulos 3 y 4, pero son importantes para la aplicación de la regresión múltiple a una amplia gama de problemas empíricos.

6.1 Efectos del escalamiento de datos sobre los estadísticos de MCO

En el capítulo 2 sobre regresión bivariada se analizaron de forma breve los efectos que la modificación de las unidades de medición tienen sobre el intercepto y la pendiente estimadas por MCO. También se mostró que la modificación de las unidades de medición no afecta la *R*-cuadrada. Ahora se vuelve al tema del escalamiento de datos y se examinan los efectos que escalar las variables dependiente o independiente tiene sobre los errores estándar, los estadísticos *t,* los estadísticos *F* y los intervalos de confianza.

Se encontrará que todo lo que se espera que ocurra, en efecto ocurre. Cuando se reescalan las variables, tanto los coeficientes como los errores estándar, los intervalos de confianza, los estadísticos *t* y los estadísticos *F* se modifican de una manera tal que se preservan todos los efectos medidos y los resultados de las pruebas. Aunque esto no es ninguna sorpresa —preocuparía que no fuera así— es útil ver de forma explícita lo que ocurre. Con frecuencia, el escalamiento de datos se emplea con fines cosméticos, por ejemplo para reducir la cantidad de ceros después del punto decimal en un coeficiente estimado. Eligiendo de manera adecuada las unidades de medición puede mejorarse la apariencia de una ecuación estimada sin modificar nada esencial.

Este problema podría tratarse de manera general, pero se ilustra mucho mejor con ejemplos. Asimismo, no tiene objeto introducir aquí una notación abstracta.

Se comienza con una ecuación en la que se relaciona el peso de los niños al nacer con la cantidad de cigarros fumados y el ingreso familiar:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc, \qquad \textbf{6.1}$$

donde

bwght = peso del niño al nacer, en onzas.
cigs = cantidad de cigarros por día que fumó la madre durante el embarazo.
faminc = ingreso familiar anual, en miles de dólares.

TABLA 6.1

Efectos del escalamiento de datos

Variable dependiente	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Variables independientes			
cigs	−.4634 (.0916)	−.0289 (.0057)	——
packs	——	——	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercepto	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observaciones	1,388	1,388	1,388
R-cuadrada	.0298	.0298	.0298
SRC	557,485.51	2,177.6778	557,485.51
EER	20.063	1.2539	20.063

En la primera columna de la tabla 6.1 se dan las estimaciones de esta ecuación obtenidos empleando los datos del archivo BWGHT.RAW. Los errores estándar se dan entre paréntesis. El estimado de *cigs* indica que si una mujer fuma cinco cigarros más por día, se pronostica que el peso del niño al nacer será aproximadamente .4634(5) = 2.317 onzas menos. El estadístico *t* para *cigs* es −5.06, de manera que esta variable es estadísticamente muy significativa.

Supóngase que ahora el peso al nacer se mide en libras, y no en onzas. Sea *bwghtlbs* = *bwght*/16 el peso al nacer en libras. ¿Qué pasa con los estadísticos de MCO si en la ecuación se emplea ésta como variable dependiente? El efecto sobre los coeficientes estimados se encuentra con facilidad mediante una sencilla manipulación de la ecuación (6.1). Toda la ecuación se divide entre 16:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)faminc.$$

Como el lado izquierdo de la ecuación es el peso al nacer dado en libras, resulta que cada uno de los coeficientes nuevos será el correspondiente coeficiente anterior dividido entre 16. Para comprobar esto, en la columna (2) de la tabla 6.1 se da la regresión de *bwghtlbs* sobre *cigs* y *faminc* A cuatro cifras decimales, el intercepto y las pendientes de la columna (2) son las de la columna (1) divididas entre 16. Por ejemplo, ahora el coeficiente de *cigs* es −.0289; esto significa que si *cigs* fuera mayor en cinco unidades, el peso al nacer sería .0289(5) = .1445 *libras* menor. En

términos de onzas, se tiene .1445(16) = 2.312, lo cual, debido al error de redondeo, es un poco diferente del 2.317 obtenido antes. El punto es que una vez que los efectos se transforman a las mismas unidades, se obtiene exactamente el mismo resultado, al margen de la manera en que se mida la variable dependiente.

¿Qué ocurre con la significancia estadística? Como era de esperarse, modificar la variable dependiente de onzas a libras no tiene ningún efecto sobre la importancia estadística de las variables independientes. En la columna (2) los errores estándar son 16 veces menores que en la columna (1). Con unos cuantos cálculos rápidos se ve que los estadísticos *t* de la columna (2) son, en realidad, idénticos a los estadísticos *t* de la columna (1). Los extremos de los intervalos de confianza en la columna (2) son exactamente los extremos de los intervalos de confianza en la columna (1) divididos entre 16. Esto se debe a que los IC se modifican por el mismo factor que los errores estándar. [Recuerde que aquí el IC de 95% es $\hat{\beta}_j \pm 1.96 \text{ ee}(\hat{\beta}_j)$.]

En términos de la bondad de ajuste, las *R*-cuadradas de las dos regresiones son idénticas, como debería ser. Observe que la suma de residuales cuadrados, SRC, y el error estándar de la regresión, EER, difieren entre estas ecuaciones. Estas diferencias se explican con facilidad. Sea $\hat{u}_i$ el residual de la observación *i* en la ecuación original (6.1). Entonces, cuando la variable independiente es *bwghtlbs* el residual es simplemente $\hat{u}_i/16$. De manera que en la segunda ecuación el residual *cuadrado* es $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$. A esto se debe que en la columna (2) la suma de los residuales cuadrados sea igual a la SRC de la columna (1) dividida entre 256.

Dado que el EER $= \hat{\sigma} = \sqrt{\text{SR}/(n - k - 1)} = \sqrt{\text{SSR}/1{,}385}$, en la columna (2) el EER es 16 veces menor que en la columna (1). Otra manera de ver esto es que el error en la ecuación, en la que como variable dependiente se emplea *bwghtlbs,* tiene una desviación estándar 16 veces menor que la desviación estándar del error original. Esto no significa que al cambiar la manera en que se mide el peso al nacer se reduzca el error; el que el EER sea menor sólo refleja la diferencia en las unidades de medición.

A continuación se regresa la variable dependiente a sus unidades originales: *bwght* se mide en onzas. Ahora se modifican las unidades de medición de una de las variables independientes, *cigs*. Se define la variable *packs* como la cantidad de cajetillas de cigarros fumadas por día. De manera que, *packs* = *cigs*/20. ¿Qué pasa ahora con los coeficientes y con los demás estadísticos de MCO? Bueno, se puede escribir

$$\widehat{bwght} = \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2 faminc = \hat{\beta}_0 + (20\hat{\beta}_1)packs + \hat{\beta}_2 faminc.$$

Por tanto, el intercepto y el coeficiente de pendiente de *faminc* no cambia, pero el coeficiente de *packs* es 20 veces el de *cigs*. Esto es intuitivamente interesante. En la columna (3) de la tabla 6.1 se presentan los resultados de la regresión de *bwght* sobre *packs* y *faminc*. A propósito, recuerde que no tendría caso incluir en una misma ecuación *cigs* y *packs* esto induciría una multicolinealidad perfecta y no tendría ningún significado.

Además del coeficiente de *packs*, en la columna (3) no hay otro estadístico que sea diferente de los estadísticos de la columna (1): el error estándar de *packs* es 20 veces mayor que el de *cigs,* que aparece en la columna (1). Esto significa que el estadístico *t* para probar la significancia de los cigarros fumados es el mismo ya sea que éstos se midan en términos de cigarros individuales o de cajetillas. Esto es natural.

Pregunta 6.1

Suponga que en la ecuación original del peso al nacer (6.1), *faminc* se mide en dólares y no en miles de dólares. Por tanto, se define la variable *fincdol* = 1,000·*faminc*. ¿Cómo se modificarán los estadísticos de MCO al emplear *fincdol* en lugar de *faminc*? Para presentar los resultados de la regresión, ¿considera usted que es mejor medir el ingreso en dólares o en miles de dólares?

El ejemplo anterior explica con detalle la mayoría de las posibilidades que surgen cuando se modifican las unidades de medición de las variables dependiente e independiente. En economía suelen cambiarse las unidades de medición cuando se trata de cantidades de dólares, en especial cuando son cantidades muy grandes.

En el capítulo 2 se dijo que, si la variable dependiente aparece de forma logarítmica, modificar las unidades de medición no afecta el coeficiente de pendiente. Lo mismo es válido aquí: modificar las unidades de medición de la variable dependiente, cuando ésta aparece de forma logarítmica, no afecta a ninguna de las estimaciones de pendiente. Esto es consecuencia del sencillo hecho de que $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ para cualquier constante $c_1 > 0$. El nuevo intercepto será $\log(c_1) + \hat{\beta}_0$. De manera similar, modificar las unidades de medición de cualquiera de las x_j, cuando en la regresión aparece $\log(x_j)$, sólo afecta el intercepto. Esto corresponde a lo que ya se sabe acerca de cambios porcentuales y, en particular, de elasticidades: son invariantes a las unidades de medición ya sea de y o de x_j. Por ejemplo, si en (6.1) se hubiera especificado log(*bwght*), como la variable dependiente, se hubiera estimado la ecuación y después vuelta a estimar con log(*bwghtlbs*) como variable dependiente, en ambas regresiones los coeficientes de *cigs* y de *faminc* hubieran sido los mismos, sólo el intercepto hubiera sido diferente.

Coeficientes beta

En econometría, algunas veces una variable clave se mide en una escala que es difícil de interpretar. Los economistas laborales suelen incluir en las ecuaciones de salario puntuaciones de pruebas, y las escalas que se usan para las puntuaciones de estas pruebas suelen ser arbitrarias y difíciles de interpretar (¡por lo menos para los economistas!). En casi todos los casos lo que interesa es comparar la puntuación de un determinado individuo con la de la población. De esta manera, en lugar de preguntar por el efecto sobre el salario por hora si, por ejemplo, la puntuación en una prueba es 10 puntos superior, es más sutil preguntar qué pasa cuando la puntuación de la prueba es una *desviación estándar* superior.

Nada impide que se vea lo que ocurre con la variable dependiente cuando en un modelo estimado una variable independiente aumenta cierto número de desviaciones estándar, suponiendo que ya se haya obtenido la desviación estándar muestral (lo cual es fácil con la mayoría de los paquetes para regresión). Esto suele ser una buena idea. Así, por ejemplo, cuando se ve el efecto de la puntuación obtenida en un examen estándar, como en el SAT (examen de admisión estándar, en Estados Unidos) sobre el GPA (promedio general en la universidad), puede determinarse la desviación estándar del SAT y ver lo que ocurre cuando la puntuación en el SAT aumenta una o dos desviaciones estándar.

Algunas veces es útil obtener resultados de la regresión cuando *todas* las variables que intervienen, tanto la variable dependiente como las variables independientes, han sido *estandarizadas*. Una variable está estandarizada cuando se le resta su media y se divide entre su desviación estándar (vea el apéndice C). Esto significa que para cada una de las variables de la muestra se calcule su *valor-z*. Después se corre la regresión empleando los valores-z.

¿Por qué es útil la estandarización? Lo más fácil es partir de la ecuación original de MCO, con las variables en sus formas originales:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i. \qquad \textbf{6.2}$$

Para hacer hincapié en que la estandarización se aplica a todos los valores muestrales se han incluido los subíndices correspondientes a la observación i. Ahora, si se obtiene el promedio de (6.2), se usa el hecho de que $\hat{u}_i$ tiene media muestral cero, y el resultado se sustrae de (6.2), se obtiene

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

Ahora, sea $\hat{\sigma}_y$ la desviación estándar muestral de la variable dependiente, sea $\hat{\sigma}_1$ la *ds* muestral de x_1, sea $\hat{\sigma}_2$ la *ds* muestral de x_2, etc. Entonces, mediante álgebra sencilla se obtiene la ecuación

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + \ldots + (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{u}_i/\hat{\sigma}_y). \qquad \textbf{6.3}$$

Cada variable en (6.3) ha sido estandarizada sustituyéndola por su valor-*z*, lo que ha dado como resultado nuevos coeficientes de pendiente. Por ejemplo, el coeficiente de pendiente de $(x_{i1} - \bar{x}_1)/\hat{\sigma}_1$ es $(\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1$. Este coeficiente es simplemente el coeficiente original, $\hat{\beta}_1$, multiplicado por el cociente de la desviación estándar de x_1 entre la desviación estándar de y. El intercepto ha desaparecido por completo.

Es útil reescribir (6.3), eliminando los subíndices *i*, como

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + error, \qquad \textbf{6.4}$$

donde z_y denota el valor-*z* de y, z_1 denota el valor-*z* de x_1, y así sucesivamente. Los nuevos coeficientes son

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \text{ para } j = 1, \ldots, k. \qquad \textbf{6.5}$$

A los $\hat{b}_j$ se les conoce como **coeficientes estandarizados** o **coeficientes beta**. (El último nombre es el más común, lo cual no es muy afortunado debido a que beta gorro es lo que se ha empleado para denotar las estimaciones *usuales* de MCO.)

La ecuación (6.4) confiere a los coeficientes beta un significado interesante: si x_1 aumenta en una desviación estándar, entonces $\hat{y}$ se modifica en $\hat{b}_1$ desviaciones estándar. De esta manera, los efectos se miden no en términos de las unidades originales de y o de x_j, sino de unidades de desviaciones estándar. Como esto vuelve irrelevantes las escalas de los regresores, esta ecuación coloca a las variables explicativas en igualdad de condiciones. En una ecuación de MCO usual, no es posible que simplemente mirando las distintas magnitudes de los coeficientes se concluya que la variable explicativa con el coeficiente mayor sea "la más importante". Se acaba de ver que cambiando las unidades de medición de las x_j pueden modificarse las magnitudes de los coeficientes según se desee. Pero, cuando se ha estandarizado cada una de las x_j, comparar las magnitudes de los coeficientes beta obtenidos es más informativo.

Aun cuando los coeficientes tengan una interpretación sencilla —por ejemplo, cuando la variable dependiente y las variables independientes de interés se encuentran en forma logarítmica, de manera que los coeficientes de MCO de interés son elasticidades estimadas— sigue siendo útil calcular los coeficientes beta. Aunque las elasticidades no tienen unidades de medición, una variación de una determinada variable explicativa de, por ejemplo, 10% puede representar una variación mayor o menor sobre el rango de esa variable que una variación de otra variable explicativa de 10%. Por ejemplo, en un estado en el que el ingreso sea grande pero la variación en el gasto por estudiante sea relativamente pequeña, no tendrá mucho sentido comparar la elasticidad del ingreso con la del gasto. Comparar las magnitudes de los coeficientes beta sí puede ser útil.

Para obtener los coeficientes beta, siempre pueden estandarizarse y, x_1, …, x_k y después correr la regresión de MCO de y sobre los valores-*z* de x_1, …, x_k —caso en el que no es necesario incluir el intercepto, ya que éste será cero—. Cuando son muchas las variables independientes esto puede ser tedioso. Algunos paquetes para regresión proporcionan los coeficientes beta mediante un sencillo comando. El ejemplo siguiente muestra el empleo de los coeficientes beta.

Ejemplo 6.1

[Efectos de la contaminación sobre el precio de la vivienda]

Para ejemplificar el uso de los coeficientes beta se emplearán los datos del ejemplo 4.5 (del archivo HPRICE2.RAW). Recuerde que la variable independiente clave es *nox*, una medida de la cantidad de óxido de nitrógeno en el aire de cada comunidad. Una manera de entender la magnitud del efecto de la contaminación —sin profundizar en los conocimientos subyacentes al efecto del óxido de nitrógeno en la calidad del aire— es calcular los coeficientes beta. (Otro método es el que se dio en el ejemplo 4.5: se obtuvo la elasticidad del precio respecto a *nox* empleando *price* y *nox* en forma logarítmica.)

La ecuación poblacional es el modelo nivel-nivel

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u,$$

donde todas las variables excepto *crime* ya fueron definidas en el ejemplo 4.5; *crime* es la cantidad de delitos reportado *per cápita*. En la ecuación siguiente se dan los coeficientes beta (así que cada variable ha sido transformada en su valor-*z*):

$$\widehat{zprice} = -.340\ znox - .143\ zcrime + .514\ zrooms - .235\ zdist - .270\ zstratio.$$

Esta ecuación indica que un aumento de *nox* en una desviación estándar hace que el precio disminuya .34 desviaciones estándar; un aumento de *crime* en una desviación estándar hace que el precio disminuya .14 desviaciones estándar. De manera que una variación relativa de la contaminación poblacional tiene un efecto mayor sobre el precio de la vivienda que la misma variación relativa de la delincuencia. El tamaño de la vivienda, medido por el número de habitaciones (*rooms*), es lo que tiene el mayor efecto estandarizado. Para saber cuál es el efecto de cada una de las variables independientes sobre el valor en dólares del precio mediano de la vivienda, es necesario emplear las variables no estandarizadas.

Emplear las variables estandarizada o no estandarizada no afecta a la significancia estadística: en ambos casos los estadísticos *t* son los mismos.

6.2 Más acerca de la forma funcional

En varios de los ejemplos anteriores ha aparecido el recurso más empleado en econometría para tomar en cuenta relaciones no lineales entre la variable explicada y las variables explicativas: el empleo de logaritmos para las variables, dependiente o independiente. También se han encontrado modelos que contienen términos cuadráticos para algunas de las variables explicativas; ahora se desea proporcionar un tratamiento sistemático de estos casos. En esta sección se verán algunas variaciones y extensiones de las formas funcionales, las cuales suelen encontrarse en el trabajo aplicado.

Más acerca del empleo de las formas funcionales logarítmicas

Se comenzará viendo de nuevo la interpretación de los parámetros en el modelo

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \qquad \textbf{6.6}$$

en el que las variables han sido tomadas del ejemplo 4.5. Recuerde que en todo el libro $\log(x)$ es el logaritmo *natural* de x. El coeficiente β_1 es la elasticidad de *price* respecto a *nox* (contaminación). El coeficiente β_2 es la variación que hay en $\log(price)$, cuando $\Delta rooms = 1$; como se ha

visto varias veces, multiplicado por 100 es el cambio porcentual aproximado de *price*. Recuerde que $100\cdot\beta_2$ suele conocerse como la semielasticidad de *price* respecto a *rooms*.

Al estimar empleando los datos del archivo HPRICE2.RAW, se obtiene

$$\widehat{\log(price)} = 9.23 - .718\log(nox) + .306\,rooms \quad \textbf{6.7}$$
$$(0.19)\ \ (.066)\qquad\qquad (.019)$$
$$n = 506,\ R^2 = .514.$$

Por tanto, cuando *nox* aumenta 1%, *price* disminuye .718%, manteniendo *rooms* constante. Cuando *rooms* aumenta en uno, *price* aumenta aproximadamente 100(.306) = 30.6%.

La estimación de que una habitación más incremente el precio aproximadamente 30.6% resulta ser un poco inexacto en esta aplicación. El error de aproximación se debe a que a medida que la variación de log(*y*) es mayor, la aproximación $\%\Delta y \approx 100\cdot\Delta\log(y)$ se hace cada vez más inexacta. Por fortuna, existe un cálculo sencillo que permite calcular el cambio porcentual exacto.

Para describir este procedimiento, se considerará el modelo general estimado

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1\log(x_1) + \hat{\beta}_2 x_2.$$

(Agregar más variables independientes no modifica el procedimiento.) Ahora, fijando x_1, se tiene $\Delta\widehat{\log(y)} = \hat{\beta}_2\Delta x_2$. Al emplear propiedades algebraicas sencillas de las funciones exponencial y logarítmica se obtiene la variación porcentual *exacta* pronosticada para *y*

$$\%\Delta\hat{y} = 100\cdot[\exp(\hat{\beta}_2\Delta x_2) - 1], \quad \textbf{6.8}$$

donde la multiplicación por 100 convierte la variación proporcional en una variación porcentual. Cuando $\Delta x_2 = 1$,

$$\%\Delta\hat{y} = 100\cdot[\exp(\hat{\beta}_2) - 1]. \quad \textbf{6.9}$$

Aplicado al ejemplo del precio de la vivienda con $x_2 = rooms$ y $\hat{\beta}_2 = .306$, $\%\Delta\widehat{price} = 100[\exp(.306) - 1] = 35.8\%$, que es claramente mayor que la variación porcentual aproximada, 30.6%, que se obtuvo directamente con la ecuación (6.7). {A propósito, éste no es un estimador insesgado porque exp(·) es una función no lineal; sin embargo, es un estimador consistente de $100[\exp(\beta_2) - 1]$. Esto se debe a que el límite de probabilidad pasa por funciones continuas, mientras que el operador del valor esperado no. Vea el apéndice C.}

El ajuste en la ecuación (6.8) no es tan crucial con cambios porcentuales pequeños. Por ejemplo, cuando en la ecuación (6.7) se incluye el cociente estudiante-profesor (*stratio*), su coeficiente estimado es −.052, lo que significa que si *stratio* aumenta en uno, *price* disminuye aproximadamente 5.2%. La variación proporcional exacta es $\exp(-.052) - 1 \approx -.051$, o sea −5.1%. Por otro lado, si *stratio* aumenta en cinco, entonces la variación porcentual aproximada del precio (*price*) es −26%, mientras que la variación exacta que se obtiene con la ecuación (6.8) es $100[\exp(-.26) - 1] \approx -22.9\%$.

La aproximación logarítmica para las variaciones porcentuales tiene una ventaja que justifica el proporcionarla en los resultados aun cuando la variación porcentual sea grande. Para describir esta ventaja, considere de nuevo el efecto sobre el precio de una variación de uno en la cantidad de habitaciones. La aproximación logarítmica es simplemente el coeficiente de *rooms*, en la ecuación (6.7), multiplicado por 100, es decir, 30.6%. Se calculó también que una estimación de la variación porcentual exacta al *aumentar* la cantidad de habitaciones en uno *es* 35.8%. Pero, ¿qué ocurre si se desea estimar el cambio porcentual cuando la cantidad de habitaciones

disminuye en uno? En la ecuación (6.8) se tendrán $\Delta x_2 = -1$ y $\hat{\beta}_2 = .306$, y de esta manera $\%\widehat{\Delta price} = 100[\exp(-.306) - 1] = -26.4$, es decir, una disminución de 26.4%. Observe que la aproximación basada en el empleo del coeficiente de *rooms* está entre 26.4 y 35.8 —un resultado que siempre se presenta—. En otras palabras, el uso simple del coeficiente (multiplicado por 100) da una estimación que está siempre entre los valores absolutos de las estimaciones correspondientes a un aumento y a una disminución. Si interesa en específico un aumento o una disminución, se emplean los cálculos basados en la ecuación (6.8).

Lo que se acaba de señalar acerca del cálculo de variaciones porcentuales es en esencia lo que se señala en la introducción a la economía cuando se trata del cálculo, por ejemplo, de elasticidades precio de la demanda con base en variaciones grandes del precio: el resultado depende de si se emplea el precio y la cantidad iniciales o finales en el cálculo de los cambios porcentuales. Emplear la aproximación logarítmica es similar en esencia a calcular la elasticidad arco de la demanda, en donde para calcular las variaciones porcentuales se emplean en los denominadores precios y cantidades promedio.

Se ha visto que emplear logaritmos naturales conduce a coeficientes con una interpretación interesante, y que pueden ignorarse las unidades de medición de las variables que aparecen en forma logarítmica porque los coeficientes de pendiente no varían ante un cambio de unidades. Existen algunas otras razones a las que se debe que los logaritmos sean tan empleados en las aplicaciones. Primero, cuando $y > 0$, los modelos en los que como variable dependiente se emplea $\log(y)$ suelen satisfacer mejor los supuestos del MLC que los modelos en los que se emplea y en forma lineal. Variables estrictamente positivas suelen tener distribuciones condicionales que son heterocedásticas o asimétricas; empleando logaritmos, ambos problemas pueden atenuarse, o incluso eliminarse.

Además, en algunos casos, el empleo de logaritmos suele estrechar el rango de la variable, en una cantidad considerable. Esto hace las estimaciones menos sensibles a observaciones atípicas (o extremas) de las variables dependiente o independiente. En el capítulo 9 se verá el tema de las observaciones atípicas.

Para el uso de logaritmos existen algunas reglas prácticas, aunque ninguna está escrita en piedra. Cuando una variable es una cantidad positiva en dólares, suele emplearse su logaritmo. Esto se ha visto en el caso de variables tales como sueldos, salarios, ventas y valor de mercado de las empresas. Variables como población, cantidad de empleados y matrícula escolar suelen aparecer en forma logarítmica; éstas tienen la característica común de ser cantidades enteras grandes.

Variables que se miden en años —como educación, experiencia, antigüedad, edad, etc.— usualmente aparecen en su forma original. Variables que son una proporción o un porcentaje —como tasa de desempleo, tasa de participación en un plan de pensiones, porcentaje de estudiantes que aprueban un examen estandarizado y tasa de detención por delitos reportados— pueden aparecer ya sea en su forma original o en forma logarítmica, aunque hay una tendencia a usarlas en su forma lineal. Esto se debe a que cualquier coeficiente de regresión relacionado con la variable *original* —sea la variable dependiente o independiente— tendrá una interpretación como variación en *puntos porcentuales*. (Vea en el apéndice A un repaso de la diferencia entre variación porcentual y variación en puntos porcentuales.) Si en una regresión se usa, por ejemplo, $\log(unem)$ donde *unem* es el porcentaje de individuos desempleados, se debe tener mucho cuidado de distinguir entre una variación de un punto porcentual y una variación porcentual. Recuerde, si *unem* varía de 8 a 9, este es un aumento de un punto porcentual, pero un aumento de 12.5% a partir del nivel inicial de desempleo. Usar el logaritmo significa que se está poniendo atención a la variación porcentual de la tasa de desempleo: $\log(9) - \log(8) \approx .118$ es decir 11.8%, que es la aproximación logarítmica del aumento real de 12.5 por ciento.

Pregunta 6.2

Suponga que la cantidad anual de detenciones de personas que conducen bajo los efectos del alcohol se determina mediante

$$\log(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2 age16\_25 + other\ factors,$$

donde *age 16_25* es la proporción de la población entre 16 y 25 años de edad. Demuestre que β_2 tiene la interpretación (*ceteris paribus*) siguiente: es el cambio porcentual en la variable *arrests* (arrestos) al aumentar el porcentaje de personas de 16 a 25 años de edad en un *punto porcentual*.

Una limitación de los logaritmos en que no pueden usarse si una variable toma valores negativos o cero. Cuando una variable y es no negativa, pero puede tomar el valor 0 suele emplearse $\log(1 + y)$. Las interpretaciones de la variación porcentual suelen preservarse, salvo cuando se trata de variaciones que comienzan en $y = 0$ (donde la variación porcentual incluso no está definida). En general, usar $\log(1 + y)$ y después interpretar las estimaciones como si la variable fuera $\log(y)$ es aceptable cuando los datos de y contienen relativamente poco ceros. Un ejemplo puede ser aquel en el que y es horas de capacitación por empleado en la población de empresas manufactureras, si una proporción grande de las empresas proporciona capacitación al menos a un trabajador. Sin embargo, técnicamente, $\log(1 + y)$ no puede estar distribuida normalmente (aunque puede ser menos heterocedástica que y). Otras posibilidades útiles, aunque más avanzadas, son los modelos Tobit y Poisson del capítulo 17.

Una desventaja de usar la variable dependiente de forma logarítmica es que es más difícil pronosticar la variable original. El modelo original permite pronosticar $\log(y)$, no y. Sin embargo, es bastante sencillo convertir un pronóstico para $\log(y)$ en un pronóstico para y (ver la sección 6.4). Un punto relacionado con esto es que *no* es correcto comparar las R-cuadradas de modelos en los que y es la variable dependiente en un caso y $\log(y)$ es variable dependiente en otro. Estas mediciones explican variaciones de variables diferentes. En la sección 6.4 se verá cómo calcular medidas comparables de bondad de ajuste.

Modelos con funciones cuadráticas

Las **funciones cuadráticas** se emplean también con bastante frecuencia en economía para captar efectos marginales crecientes o decrecientes. En el apéndice A se encuentra un repaso de las propiedades de las funciones cuadráticas.

El caso más simple, es aquel en el que y depende de un solo factor observado x, pero lo hace de forma cuadrática:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Por ejemplo, tómese $y = wage$ (salario) y $x = exper$ (experiencia). Como se vio en el capítulo 3, este modelo cae fuera del análisis de regresión simple pero puede ser tratado con facilidad empleando regresión múltiple.

Es importante recordar que β_1 no mide la variación en y respecto a x; no tiene ningún sentido mantener x^2 constante mientras se varía x. Si la ecuación estimada se expresa como

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \qquad \textbf{6.10}$$

entonces se tiene la aproximación

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ de manera que } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad \textbf{6.11}$$

Esto indica que la pendiente de la relación entre x y y dependen del valor de x; la pendiente estimada es $\hat{\beta}_1 + 2\hat{\beta}_2 x$. Si se sustituye con $x = 0$, se ve que $\hat{\beta}_1$ puede interpretarse como la

pendiente aproximada al pasar de $x = 0$ a $x = 1$. Después de eso, el segundo término, $2\hat{\beta}_2 x$, debe ser tomado en cuenta.

Si lo único que interesa es calcular el cambio predicho en y dado un valor inicial de x y una variación de x, puede emplearse directamente (6.10): no hay ninguna razón para emplear el cálculo de aproximación. Sin embargo, en general lo que más interesa es resumir con rapidez el efecto de x sobre y, y, en este caso, la interpretación de $\hat{\beta}_1$ y $\hat{\beta}_2$ en (6.11) proporciona este resumen. Típicamente, x se sustituye por su valor promedio en la muestra o por otros valores interesantes como la mediana o los valores de los cuartiles inferior o superior.

En muchas de las aplicaciones, $\hat{\beta}_1$ es positivo y $\hat{\beta}_2$ es negativo. Por ejemplo, empleando los datos de salario del archivo WAGE1.RAW, se obtiene

$$\widehat{wage} = 3.73 + .298\, exper - .0061\, exper^2 \quad \textbf{6.12}$$
$$(.35)\ (.041)\qquad (.0009)$$
$$n = 526,\ R^2 = .093.$$

La ecuación estimada indica que *exper* tiene un efecto decreciente sobre *wage*. El primer año de experiencia vale 30¢ por hora ($.298). El segundo año de experiencia vale menos [aproximadamente $.298 - 2(.0061)(1) \approx .286$, es decir 28.6¢, de acuerdo con la aproximación en (6.11) con $x = 1$]. Al pasar de 10 a 11 años de experiencia, el aumento predicho en *wage* es aproximadamente $.298 - 2(.0061)(10) = .176$, es decir 17.6¢. Y así sucesivamente.

Cuando el coeficiente de x es positivo y el de x^2 es negativo la cuadrática tiene forma parabólica. Siempre existe un valor positivo de x para el que el efecto de x sobre y es cero; antes de este punto, x tiene un efecto positivo sobre y; después, x tiene un efecto negativo sobre y. En la práctica, es importante saber dónde se encuentra este punto de inflexión.

En la ecuación estimada (6.10) si $\hat{\beta}_1 > 0$ y $\hat{\beta}_2 < 0$, el punto de inflexión (o máximo de la función) siempre se alcanzará en el punto correspondiente al coeficiente de x sobre *el doble* del valor absoluto del coeficiente de x^2:

$$x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|. \quad \textbf{6.13}$$

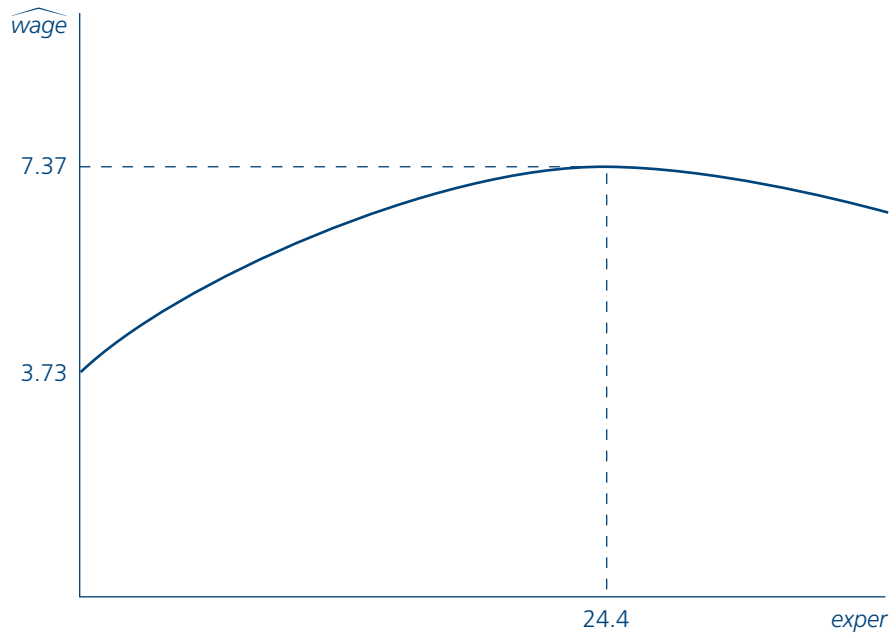
En el ejemplo del salario, $x^* = exper^*$ es $.298/[2(.0061)] \approx 24.4$. (Observe cómo al hacer este cálculo simplemente se ha eliminado el signo menos de $-.0061$.) Esta relación cuadrática se ilustra en la figura 6.1.

En la ecuación (6.12) del salario, el rendimiento de la experiencia se vuelve cero aproximadamente a los 24.4 años. ¿Cómo debe entenderse esto? Existen al menos tres explicaciones posibles. Primero, puede ser que en la muestra haya pocas personas que tengan más de 24 años de experiencia y por esto la parte de la curva a la derecha de 24 puede ignorarse. El costo de emplear ecuaciones cuadráticas para captar efectos decrecientes es que llega un momento en el que la forma cuadrática cambia de dirección. Si este punto está más allá de todas excepto un porcentaje pequeño de las personas de la muestra, entonces esto no debe preocupar mucho. Pero en la base de datos WAGE1.RAW, aproximadamente 28% de las personas de la muestra tienen más de 24 años de experiencia; este es un porcentaje muy alto como para ignorarlo.

Es posible que el rendimiento de *exper* realmente se vuelva negativo en algún punto, pero es difícil creer que esto ocurra a los 24 años de experiencia. Una posibilidad más creíble es que el efecto estimado de *exper* sobre *wage* esté sesgado debido a que no se hayan controlado otros factores, o debido a que la relación funcional entre *wage* y *exper* dada por la ecuación (6.12)

FIGURA 6.1

Relación cuadrática entre $\widehat{wage}$ y *exper*.



no sea del todo correcta. En el ejercicio para computadora C6.2 se pide al lector explorar estas posibilidades controlando la educación y empleando log(*wage*) como variable dependiente.

Cuando en un modelo hay una variable dependiente de forma logarítmica y una variable explicativa que aparece de forma cuadrática, debe tenerse cuidado al reportar los efectos parciales. El ejemplo siguiente muestra que las funciones cuadráticas pueden tener también forma de U, en lugar de la forma parabólica. En la ecuación (6.10) la forma de U surge cuando $\hat{\beta}_1$ es negativa y $\hat{\beta}_2$ es positiva; esto capta un efecto creciente de x sobre y.

Ejemplo 6.2

[Efectos de la contaminación sobre el precio de la vivienda]

Ahora se modificará el modelo del ejemplo 4.5 para el precio de la vivienda incluyendo un término cuadrático en *rooms*:

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \qquad \textbf{6.14}$$

El modelo estimado con los datos del archivo HPRICE2.RAW es

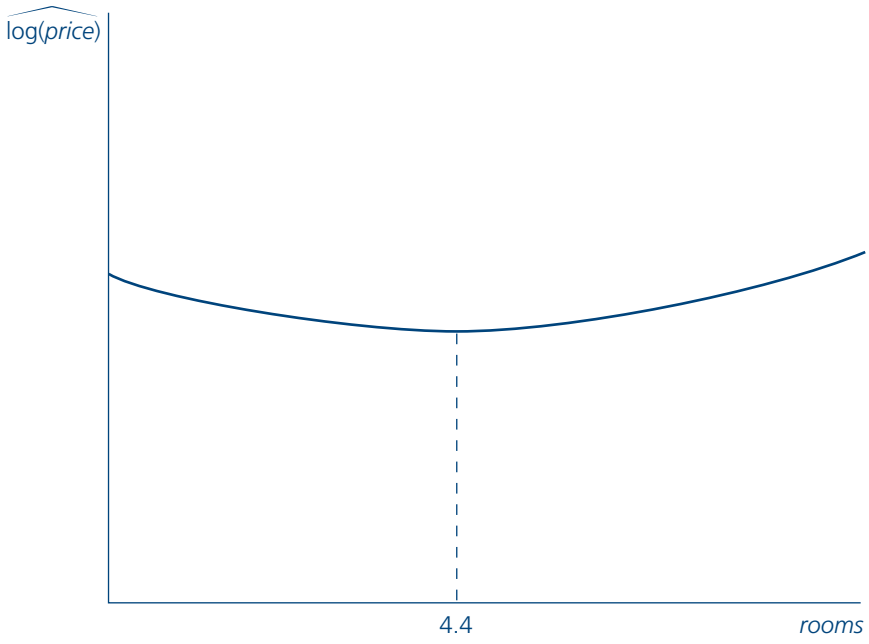
$$\widehat{\log(price)} = 13.39 - .902\log(nox) - .087\log(dist)$$
$$(.57)\quad(.115)\qquad(.043)$$
$$-.545\,rooms + .062\,rooms^2 - .048\,stratio$$
$$(.165)\qquad(.013)\qquad(.006)$$
$$n = 506,\ R^2 = .603.$$

El estadístico t para el término cuadrático $rooms^2$ es aproximadamente 4.77, de manera que este término es estadísticamente muy significativo. Pero, ¿cuál es la interpretación del efecto de *rooms* sobre log(*price*)? En un principio, el efecto parece ser extraño. Como el coeficiente de *rooms* es negativo y el de $rooms^2$ es positivo, esta ecuación literalmente implica que, para valores bajos de *rooms*, una habitación más tiene un efecto *negativo* sobre log(*price*). En algún punto el efecto se vuelve positivo, y la forma cuadrática significa que la semielasticidad de *price* respecto a *rooms* aumenta a medida que *rooms* aumenta. Esta situación se muestra en la figura 6.2.

El valor del punto de inflexión de *rooms* se obtiene empleando la ecuación (6.13) (aun cuando $\hat{\beta}_1$ sea negativo y $\hat{\beta}_2$ sea positivo). El valor absoluto del coeficiente de *rooms*, .545, dividido entre el doble del coeficiente de $rooms^2$, .062, da $rooms^* = .545/[2(.062)] \approx 4.4$; este punto se indica en la figura 6.2.

FIGURA 6.2

$\widehat{\log(price)}$ como función cuadrática de *rooms*.



¿Puede creerse que partiendo de tres habitaciones y aumentando a cuatro en realidad se reduzca el valor esperado de una casa? Lo más seguro es que no. Resulta que sólo en cinco de las 506 comunidades de la muestra hay casas en las que la cantidad promedio de habitaciones sea 4.4 o menos, aproximadamente 1% de la muestra. Esta cantidad es tan pequeña que para fines prácticos la función cuadrática a la izquierda de 4.4 puede ignorarse. A la derecha de 4.4, se ve que agregar una habitación tiene un efecto creciente sobre el cambio porcentual del precio:

$$\Delta\widehat{\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

y de esta manera

$$\begin{aligned}\%\Delta\widehat{price} &\approx 100\{[-.545 + 2(.062)]rooms\}\Delta rooms \\ &= (-54.5 + 12.4\ rooms)\Delta rooms.\end{aligned}$$

Así, el aumento de *rooms* de, por ejemplo, cinco a seis, hace que el precio aumente aproximadamente $-54.5 + 12.4(5) = 7.5\%$; el aumento de seis a siete hace que el precio aumente aproximadamente $-54.5 + 12.4(6) = 19.9\%$. Este es un efecto creciente muy fuerte.

En general, ¿qué pasa si los coeficientes de los términos lineal y cuadrado tienen el *mismo* signo (ya sea ambos positivos o ambos negativos) y la variable explicativa sea necesariamente no negativa (como en el caso de *rooms* o de *exper*)? En cualquier caso, no hay un punto de inflexión en un valor de $x > 0$. Por ejemplo, si tanto β_1 como β_2 son positivos, el menor valor esperado de y se encuentra en $x = 0$ y el aumento de x siempre tiene un efecto positivo y creciente sobre y. (Esto ocurre también si $\beta_1 = 0$ y $\beta_2 > 0$, lo que significa que el efecto parcial en $x = 0$ es cero y crece a medida que crece x.) De manera similar, si tanto β_1 como β_2 son negativos, el mayor valor esperado de y se encuentra en $x = 0$, y aumentos en x tienen un efecto negativo sobre y, y la magnitud del efecto aumenta a medida que x es más grande.

Hay muchas otras posibilidades del empleo de funciones cuadráticas junto con logaritmos. Por ejemplo, una extensión de (6.14) que permite una elasticidad no constante entre *price* y *nox* es

$$\begin{aligned}\log(price) &= \beta_0 + \beta_1\log(nox) + \beta_2[\log(nox)]^2 \\ &+ \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u.\end{aligned} \qquad \textbf{6.15}$$

Si $\beta_2 = 0$, entonces β_1 es la elasticidad de *price* respecto a *nox*. De otro modo, esta elasticidad depende del nivel de *nox*. Para ver esto, pueden combinarse los argumentos de los efectos parciales en los modelos cuadráticos y logarítmico para mostrar que

$$\%\Delta price \approx [\beta_1 + 2\beta_2\log(nox)]\%\Delta nox; \qquad \textbf{6.16}$$

por tanto, la elasticidad de *price* respecto a *nox* es $\beta_1 + 2\beta_2\log(nox)$, de manera que dependen de $\log(nox)$.

Por último, en los modelos de regresión pueden incluirse otros términos polinomiales. Desde luego, el cuadrático es el que se encuentra con más frecuencia, pero de vez en cuando hay un término cúbico o incluso uno cuártico. Una forma funcional que suele ser razonable para una función de costo total es

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u.$$

Estimar un modelo así no ofrece complicaciones. Interpretar los parámetros es más complicado (aunque sencillo empleando el cálculo); aquí estos modelos no se estudiarán más.

Modelos con términos de interacción

Algunas veces es natural que el efecto parcial, la elasticidad o la semielasticidad de la variable dependiente respecto a una variable explicativa dependa de la magnitud de *otra* variable explicativa. Por ejemplo, en el modelo

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u,$$

el efecto parcial de *bdrms* (cantidad de recámaras) sobre *price* (precio) (manteniendo constantes todas las demás variables) es

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft. \qquad \textbf{6.17}$$

Si $\beta_3 > 0$, entonces (6.17) implica que, en casas más grandes, una recámara más produce un aumento mayor en el precio. En otras palabras, existe un **efecto de interacción** entre la superficie en pies cuadrados y la cantidad de recámaras. Para sintetizar el efecto de *bdrms* sobre *price*, es necesario evaluar (6.17) en valores útiles de *sqrft*, como por ejemplo en el valor de la media, o en los cuartiles inferior o superior en la muestra. Si β_3 es o no cero es sencillo de probar.

Cuando se incluye un término de interacción puede ser complicado interpretar los parámetros de las variables originales. Por ejemplo, la ecuación anterior para el precio de las casas, (6.17), indica que β_2 es el efecto de *bdrms* sobre *price* para una casa con ¡cero pies cuadrados! Este efecto, por supuesto, no tiene mucho interés. En cambio, se debe tener cuidado de elegir valores útiles de *sqrft*, valores tales como la media o la mediana muestrales, en la versión estimada de la ecuación (6.17).

Con frecuencia es útil volver a parametrizar un modelo de manera que los coeficientes de las variables originales tengan un significado útil. Considere un modelo que tenga dos variables explicativas y una interacción:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

Como se acaba de decir, β_2 es el efecto parcial de x_2 sobre y cuando $x_1 = 0$. Con frecuencia esto no tiene ningún interés. En cambio, se puede parametrizar de nuevo el modelo y obtener

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u,$$

donde μ_1 es la media poblacional de x_1 y μ_2 es la media poblacional de x_2. Ahora, puede verse con facilidad que el coeficiente de x_2, δ_2, es el efecto parcial de x_2 sobre y en la media de x_1. (Multiplicando en la segunda ecuación la interacción y comparando los coeficientes se puede demostrar fácilmente que $\delta_2 = \beta_2 + \beta_3 \mu_1$. El parámetro δ_1 tiene una interpretación similar.) Por tanto, si se sustrae la media de las variables —en la práctica, esta es la media muestral— antes de crear el término de interacción, los coeficientes de las variables originales tienen una interpretación útil. Además, se obtienen de inmediato los errores estándar de los efectos parciales en los valores medios. Nada impide sustituir μ_1 o μ_2 por otros valores de las variables explicativas que puedan ser de interés. El ejemplo siguiente muestra cómo se pueden usar los términos de interacción.

Ejemplo 6.3

[Efectos de la asistencia a clases sobre el desempeño en el examen final]

Un modelo para explicar el resultado estandarizado de un examen final (*stndfnl*) en términos del porcentaje de asistencia a clases, el anterior promedio general de calificaciones y la puntuación en el ACT (examen de admisión a la universidad) es

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u. \quad \textbf{6.18}$$

(Se emplea la puntuación estandarizada obtenida en el examen por las razones vistas en la sección 6.1: facilita interpretar el desempeño de un estudiante en relación con el resto del grupo.) Además de los términos cuadráticos de *priGPA* y *ACT*, este modelo contiene un término de interacción entre *priGPA* y la tasa de asistencia a clases. La idea es que la asistencia a clases puede tener efectos diferentes en estudiantes con distinto desempeño en el pasado, medido mediante el *priGPA*. Lo que interesa es el efecto de la asistencia a clases sobre la puntuación en el examen final: $\Delta stndfnl/\Delta atndrte = \beta_1 + \beta_6 priGPA$.

Empleando las 680 observaciones del archivo ATTEND.RAW, de los estudiantes de una materia en principios de microeconomía, la ecuación estimada es

$$\widehat{stndfnl} = 2.05 - \underset{(1.36)}{.0067}\, atndrte - \underset{(.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT + \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA \cdot atndrte \quad \textbf{6.19}$$

$$n = 680,\ R^2 = .229,\ \bar{R}^2 = .222.$$

Esta ecuación debe interpretarse con extremo cuidado. Si sólo se observa el coeficiente de *atndrte* se concluirá, de manera incorrecta, que la asistencia a clases tiene efecto *negativo* sobre la puntuación obtenida en el examen final. Pero se supone que este coeficiente mide el efecto cuando $priGPA = 0$, lo cual no tiene ningún interés (el *priGPA* menor de la muestra es .86). También se debe tener cuidado de no considerar por separado las estimaciones de β_1 y β_6 y concluir que, como cada uno de los estadísticos *t* es insignificante, no se puede rechazar H_0: $\beta_1 = 0, \beta_6 = 0$. En realidad, el valor-*p* de la prueba *F* de esta hipótesis conjunta es .014, de manera que con toda seguridad se rechaza H_0 al nivel de 5%. Este es un buen ejemplo de que considerar por separado estadísticos *t* al probar una hipótesis conjunta puede conducir a conclusiones equivocadas.

¿Cómo debe estimarse el efecto parcial de *atndrte* sobre *stndfnl*? Para obtener el efecto parcial, en la ecuación deben sustituirse valores útiles de *priGPA*. En la muestra, el valor medio de *priGPA* es 2.59, de manera que en la media de *priGPA*, el efecto de *atndrte* sobre *stndfnl* es $-.0067 + .0056(2.59) \approx .0078$. ¿Qué significa esto? Como *atndrte* se mide como un porcentaje, esto significa que si *atndrte* aumenta 10 puntos porcentuales $\widehat{stndfnl}$ aumentará .078 desviaciones estándar a partir de la puntuación media en el examen final.

¿Cómo puede decirse si la estimación .0078 es estadísticamente distinta de cero? Se necesita volver a correr la regresión, sustituyendo *priGPA*·*atndrte* por (*priGPA* − 2.59)·*atndrte*. Esto da, como nuevo coeficiente de *atndrte*, el efecto estimado cuando $priGPA = 2.59$, junto con su error estándar; en la regresión no cambia nada más. (En la sección 4.4 se describe este mecanismo.) Corriendo esta regresión se obtiene que el error estándar de $\hat{\beta}_1 + \hat{\beta}_6(2.59) = .0078$ es .0026, con lo que $t = .0078/.0026 = 3$. Por tanto, en el *priGPA*, promedio, se concluye que la asistencia a clases tiene un efecto positivo, estadísticamente significativo sobre la puntuación en el examen final.

Pregunta 6.3

Si a la ecuación (6.18) se le agrega el término $\beta_7 ACT \cdot atndrte$, ¿cuál será el efecto parcial de *atndrte* sobre *stndfnl*?

Cuando se trata de determinar el efecto de *priGPA* sobre *stndfnl* las cosas se vuelven un poco más complicadas debido al término cuadrático $priGPA^2$. Para determinar el efecto en el valor medio de *priGPA* y en la tasa media de asistencia, 82, se sustituye $priGPA^2$ por $(priGPA - 2.59)^2$ y $priGPA \cdot atndrte$ por $priGPA \cdot (atndrte - 82)$. El coeficiente de *priGPA* será el efecto parcial en los valores medios y se tendrá su error estándar. (Vea el ejercicio para computadora C6.7.)

6.3 Más sobre bondad de ajuste y selección de los regresores

Hasta ahora, al evaluar los modelos de regresión no se ha dado mucha atención a la magnitud de R^2 esto se debe a que los estudiantes principiantes tienden a darle demasiada importancia a *R*-cuadrada. Como se verá en breve, elegir un conjunto de variables explicativas con base en la magnitud de la *R*-cuadrada puede conducir a modelos sin sentido. En el capítulo 10 se verá que las *R*-cuadradas obtenidas de regresiones de series de tiempo pueden ser artificialmente altas y, como resultado, llevar a conclusiones equivocadas.

En los supuestos del modelo lineal clásico no hay nada que requiera que R^2 sea mayor que algún valor particular; R^2 es simplemente una buena estimación de cuánto de la variación de y es explicada por $x_1, x_2, \ldots, x_k$ en la población. Aquí se han visto varias regresiones que han tenido *R*-cuadradas bastante pequeñas. Aunque esto significa que hay diversos factores que afectan a y, que no han sido tomados en cuenta, esto no significa que los factores en u estén correlacionados con las variables independientes. El supuesto RLM.4 de media condicional cero es lo que determina que se obtengan estimadores insesgados de los efectos *ceteris paribus* de las variables independientes, y la magnitud de la *R*-cuadrada no tiene relación directa con esto.

Que *R*-cuadrada sea pequeña implica que la varianza del error es relativamente grande en relación con la varianza de y, lo cual significa que es posible que sea difícil estimar con precisión las β_j. Pero recuerde que en la sección 3.4 se vio que una varianza grande del error puede ser contrarrestada con un tamaño de muestra grande: si se tienen suficientes datos, es posible que se puedan estimar con precisión los efectos parciales, aun cuando no se hayan controlado muchos de los factores no observados. El que se puedan o no obtener estimaciones suficientemente precisas depende de la aplicación. Por ejemplo, suponga que a algunos estudiantes nuevos de una universidad grande se les otorga financiamiento, de forma aleatoria, para comprar equipos de cómputo. Si los montos del financiamiento en realidad son determinados de forma aleatoria, se puede determinar el efecto *ceteris paribus* del monto del financiamiento sobre el subsecuente promedio general de calificaciones en la universidad empleando el análisis de regresión simple. [Debido a la asignación aleatoria, todos los demás factores que afectan el promedio general de calificaciones (GPA) no estarán correlacionados con el monto del financiamiento.] Es probable que el monto del financiamiento explique poco de la variación del GPA, de manera que la *R*-cuadrada de esta regresión tal vez será muy pequeña. Sin embargo, si se obtiene una muestra de tamaño grande, será posible obtener un estimador razonablemente preciso del efecto del financiamiento.

Otro buen ejemplo en el que el poder explicativo no tiene nada que ver con la estimación insesgada de las β_j es el que se obtiene mediante el análisis de la base de datos del archivo APPLE.RAW. A diferencia de otras bases de datos empleadas aquí, las variables explicativas clave del archivo APPLE.RAW se obtuvieron de manera experimental —es decir, sin atender a otros factores que pudieran afectar a la variable dependiente—. La variable que se desea explicar, *ecolbs*, es la cantidad (hipotética) de libras de manzanas "no dañinas para la ecología" ("ecoetiquetadas") que demanda una familia. A cada familia (en realidad, a cada cabeza de familia) se le presentó una descripción de las manzanas ecoetiquetadas, junto con los precios

de las manzanas normales (o regulares) (*regprc*) y los precios de las manzanas ecoetiquetadas hipotéticas (*ecoprc*). Dado que a las familias se les asignaron los pares de precios de forma aleatoria, éstos no están relacionados con otros factores observados (tales como el ingreso familiar) ni con factores no observados (tales como el interés por un ambiente limpio). Por tanto, la regresión de *ecolbs* sobre *ecoprc*, *regprc* (a través de las muestras generadas de esta manera) produce estimadores insesgados de los efectos del precio. Sin embargo, la *R*-cuadrada de esta regresión es sólo .0364: las variables precio sólo explican cerca de 3.6% de la variación total de *ecolbs*, de manera que aquí hay un caso en el que se explica muy poco de la variación de *y* aun cuando se está en la rara situación en que se sabe que los datos han sido generados de manera que se puede obtener una estimación insesgada de las β_j. (Dicho sea de paso, la adición de características observadas en la familia tiene un efecto muy pequeño sobre el poder explicativo. Vea el ejercicio para computadora C6.11.)

Recuerde, sin embargo, que cuando a una ecuación se le agregan variables la *variación* relativa de *R*-cuadrada es muy útil: en (4.41) el estadístico *F* para probar la significancia conjunta depende de manera crucial de la diferencia entre las *R*-cuadradas del modelo no restringido y el restringido.

R-cuadrada ajustada

La mayoría de los paquetes para regresión dan, junto con la *R*-cuadrada, un estadístico llamado ***R*-cuadrada ajustada**. Dado que esta última se reporta en muchas de las aplicaciones y tiene algunas características útiles, se verá en esta subsección.

Para ver cómo puede ajustarse la *R*-cuadrada usual, es útil escribirla de la manera siguiente:

$$R^2 = 1 - (\text{SRC}/n)/(\text{STC}/n), \qquad \textbf{6.20}$$

donde SRC es la suma de los residuales cuadrados y STC es la suma total de cuadrados; comparando con la ecuación (3.28), lo único que se ha hecho es dividir entre *n* tanto la SRC como la STC. Esta expresión muestral lo que en realidad estima la R^2. Defínase σ_y^2 como la varianza poblacional de *y* y sea σ_u^2 la varianza poblacional del término del error, *u*. (Hasta ahora, σ^2 se ha empleado para denotar σ_u^2, pero aquí es útil ser más específico.) La ***R*-cuadrada poblacional** se define como $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$; que es la proporción de la variación de *y* en la población explicada por las variables independientes. Esto es lo que se supone que estima R^2.

R^2 estima σ_u^2 mediante SRC/*n*, que se sabe es sesgado, de manera que, ¿por qué no sustituir SRC/*n* por SRC/$(n - k - 1)$? Además, también se puede usar STC/$(n - 1)$ en lugar de STC/*n*, ya que el primero es un estimador insesgado de σ_y^2. Empleando estos estimadores, se llega a la *R*-cuadrada ajustada:

$$\begin{aligned}\bar{R}^2 &= 1 - [\text{SRC}/(n - k - 1)]/[\text{STC}/(n - 1)] \\ &= 1 - \hat{\sigma}^2/[\text{STC}/(n - 1)],\end{aligned} \qquad \textbf{6.21}$$

ya que $\hat{\sigma}^2 = \text{SRC}/(n - k - 1)$. Debido a la notación empleada para denotar la *R*-cuadrada ajustada, también se le suele llamar *R-barra cuadrada*.

A la *R*-cuadrada ajustada se le llama también *R-cuadrada corregida*, pero este nombre no resulta muy adecuado porque implicaría que $\bar{R}^2$ de alguna manera es un mejor estimador de la *R*-cuadrada poblacional que R^2. Por desgracia, en general $\bar{R}^2$ *no* se sabe que sea un mejor estimador. Se podría pensar que $\bar{R}^2$ corrigiera el sesgo de R^2 al estimar la *R*-cuadrada poblacional, ρ^2, pero no lo hace: el cociente de dos estimadores insesgados no es un estimador insesgado.

La característica más atractiva de $\bar{R}^2$ es que impone una sanción a la adición de más variables independientes a un modelo. Se sabe que R^2 nunca disminuye cuando se agrega una variable independiente a la ecuación de regresión: esto se debe a que SRC nunca aumenta (y en general disminuye) a medida que se agregan más variables independientes. Pero la fórmula de $\bar{R}^2$ muestra

que ésta depende de manera explícita de k, la cantidad de variables independientes. Si se agrega una variable independiente a la regresión, SRC disminuye, pero lo mismo ocurre con los gl en la regresión, $n - k - 1$. SRC/$(n - k - 1)$ puede aumentar o disminuir cuando se agrega una nueva variable independiente a una regresión.

Un hecho algebraico interesante es el siguiente: si se agrega una variable independiente a una ecuación de regresión, $\bar{R}^2$ aumenta si, y sólo si, el valor absoluto del estadístico t de la nueva variable es mayor que uno. (Una extensión de esto es que cuando a una regresión se agrega un grupo de variables $\bar{R}^2$ aumenta si, y sólo si, el estadístico F de la significancia conjunta de las nuevas variables es mayor que la unidad.) Por tanto, de inmediato se ve que emplear $\bar{R}^2$ para decidir si una determinada variable independiente (o un conjunto de variables) pertenece a un modelo proporciona una respuesta distinta a la de las pruebas estándar t o F (debido a que a los niveles de significancia tradicionales un estadístico t o F igual a uno no es estadísticamente significativo).

Algunas veces es útil tener una fórmula de $\bar{R}^2$ en términos de R^2. Mediante álgebra sencilla se obtiene

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1). \qquad \textbf{6.22}$$

Por ejemplo, si $R^2 = .30$, $n = 51$ y $k = 10$, entonces $\bar{R}^2 = 1 - .70(50)/40 = .125$. Por tanto, cuando n es pequeña y k es grande, $\bar{R}^2$ puede ser sustancialmente menor a R^2. En realidad, si la R-cuadrada usual es pequeña y $n - k - 1$ es pequeño, ¡$\bar{R}^2$ puede ser negativa! Por ejemplo, se puede sustituir $R^2 = .10$, $n = 51$ y $k = 10$ para comprobar que $\bar{R}^2 = -.125$. Una $\bar{R}^2$ negativa indica un ajuste muy pobre del modelo en relación con los grados de libertad.

En las regresiones, algunas veces se reporta la R-cuadrada ajustada junto con la R-cuadrada usual, y algunas veces se reporta $\bar{R}^2$ en lugar de R^2. Es importante recordar que la que aparece en el estadístico F en (4.41) es R^2 y no $\bar{R}^2$. La misma fórmula con $\bar{R}^2_r$ o $\bar{R}^2_{ur}$ *no* es válida.

Uso de la R-cuadrada ajustada para elegir entre modelos no anidados

En la sección 4.5 se vio cómo calcular un estadístico F para probar la significancia conjunta de un grupo de variables; esto permite decidir, a un determinado nivel de significancia, si por lo menos una de las variables del grupo afecta a la variable dependiente. Esta prueba no permite decidir *cuál* de las variables es la que tiene algún efecto. En algunos casos, se desea elegir un modelo que no tenga variables independientes redundantes, y para esto puede servir la R-cuadrada ajustada.

En el ejemplo del sueldo en la liga mayor de béisbol en la sección 4.5 se vio que ni *hrunsyr* ni *rbisyr* eran significativas individualmente. Estas dos variables están muy correlacionadas, de manera que habrá que elegir entre los modelos

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$$

y

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

Estas dos ecuaciones son **modelos no anidados**, ya que ninguna de ellas es un caso especial de la otra. El estadístico F visto en el capítulo 4 sólo permite probar modelos *anidados*: un modelo (el restringido) es un caso especial del otro (el no restringido). Como ejemplos de modelos

restringidos y no restringidos vea las ecuaciones (4.32) y (4.28). Una posibilidad es crear un modelo compuesto que contenga *todas* las variables explicativas de los modelos originales y después probar cada modelo contra el general empleando la prueba *F*. El problema de este proceso es que puede que ambos se rechacen o que ninguno se rechace (como ocurre en el ejemplo del sueldo en la liga mayor de béisbol en la sección 4.5). De manera que este proceso no siempre proporciona una manera de distinguir entre modelos con regresores no anidados.

En la regresión del sueldo de los jugadores de béisbol, $\bar{R}^2$ en la regresión que contiene *hrunsyr* es .6211 y $\bar{R}^2$ en la regresión que contiene *rbisyr* es .6226. Por tanto, con base en la *R*-cuadrada ajustada, hay una muy ligera preferencia por el modelo que tiene *rbisyr*. Pero, prácticamente la diferencia es muy pequeña y es posible que se obtengan respuestas diferentes controlando algunas de las variables del ejercicio para computadora C4.5. (Como los dos modelos no anidados contienen cinco parámetros, puede emplearse la *R*-cuadrada usual para llegar a la misma conclusión.)

Comparar $\bar{R}^2$ para elegir entre diversos conjuntos no anidados de variables independientes puede ser valioso cuando las variables representan formas funcionales diferentes. Considere dos modelos que relacionan la intensidad de la Investigación y Desarrollo (I&D) con las ventas de una empresa:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u. \qquad \textbf{6.23}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \qquad \textbf{6.24}$$

El primer modelo capta el rendimiento decreciente mediante la inclusión de *sales* en forma logarítmica; el segundo modelo hace lo mismo empleando un término cuadrático y, por tanto, contiene un parámetro más que el primero.

Cuando se estima la ecuación (6.23) empleando las 32 observaciones de empresas químicas del archivo RDCHEM.RAW, R^2 es .061 y para la ecuación (6.24) R^2 es .148. Por tanto, parece que la ecuación cuadrática tiene un ajuste mucho mejor. Pero la comparación de las *R*-cuadradas usuales no es justa para el primer modelo porque éste contiene un parámetro menos que la ecuación (6.24). Es decir, (6.23) es un modelo más parsimonioso que (6.24).

Permaneciendo todo lo demás igual, los modelos más sencillos son mejores. Dado que la *R*-cuadrada usual no sanciona modelos más complicados, es mejor emplear $\bar{R}^2$. Para (6.23) $\bar{R}^2$ es .030, mientras que para (6.24) $\bar{R}^2$ es .090. Por tanto, aun después del ajuste para la diferencia en los grados de libertad, el modelo cuadrático gana. Tal modelo también se prefiere cuando a cada regresión se agregan márgenes de ganancia.

Existe una limitación importante al usar $\bar{R}^2$ para elegir entre modelos no anidados: no se puede usar para escoger entre distintas formas funcionales de la variable dependiente. Esto es lamentable, porque con frecuencia se desea decidir entre si usar *y* o log(*y*) (o alguna otra transformación) como variable dependiente con base en la bondad de ajuste. Pero para este propósito no se pueden usar ni R^2 ni $\bar{R}^2$. La razón es sencilla: estas *R*-cuadradas miden la proporción explicada de la variación total en la variable dependiente que se emplee en la regresión, y funciones diferentes de la variable dependiente tendrán cantidades distintas de variación a explicar. Por ejemplo, la variación total de *y* y de log(*y*) no son las mismas, y con frecuencia son muy distintas. Comparar las *R*-cuadradas ajustadas de regresiones que tienen estas distintas formas de las variables dependientes no dice nada acerca de qué modelo se ajusta mejor; los modelos se ajustan a dos variables dependientes distintas.

Pregunta 6.4

Explique por qué es lo mismo elegir un modelo maximizando $\bar{R}^2$ que minimizando $\hat{\sigma}$ (el error estándar de la regresión).

Ejemplo 6.4

[Compensación de los directores generales y desempeño de una empresa]

Considere dos modelos estimados que relacionan la compensación de los directores generales (CEO) con el desempeño de la empresa:

$$\widehat{salary} = 830.63 + .0163\, sales + 19.63\, roe \quad \textbf{6.25}$$
$$(223.90)\ \ (.0089)\ \ \ \ (11.08)$$
$$n = 209,\ R^2 = .029,\ \bar{R}^2 = .020$$

y

$$\widehat{lsalary} = 4.36 + .275\, lsales + .0179\, roe \quad \textbf{6.26}$$
$$(0.29)\ \ (.033)\ \ \ \ (.0040)$$
$$n = 209,\ R^2 = .282,\ \bar{R}^2 = .275,$$

donde *roe* es la rentabilidad de la inversión vista en el capítulo 2. Para simplificar, *lsalary* y *lsales* denotan los logaritmos naturales de *salary* y de *sales*. Se sabe, ya, cómo interpretar estas distintas ecuaciones estimadas. Pero, ¿puede decirse que un modelo se ajuste mejor que el otro?

La *R*-cuadrada de la ecuación (6.25) indica que *sales* y *roe* explican sólo cerca de 2.9% de la variación del sueldo de los CEO en la muestra. Tanto *sales* como *roe* tienen una significación estadística marginal.

La ecuación (6.26) indica que el log(*sales*) y *roe* explican aproximadamente 28.2% de la variación de log(*salary*). En términos de bondad de ajuste, esta *R*-cuadrada mayor parecería implicar que el modelo (6.26) es mucho mejor, pero este no es necesariamente el caso. La suma total de cuadrados de *salary* en la muestra es 391,732,982, mientras que la suma total de cuadrados de log(*salary*) es sólo de 66.72. Por tanto, en log(*salary*) hay mucho menos variación que debe ser explicada.

En este punto pueden emplearse otras características distintas de R^2 y $\bar{R}^2$ para decidir entre estos modelos. Por ejemplo, en (6.26) log(*sales*) y *roe* son estadísticamente mucho más significativas que *sales* y *roe* en (6.25), y los coeficientes en (6.26) son quizá de mayor interés. Sin embargo, para estar seguro, se requerirá hacer una comparación válida de la bondad de ajuste.

En la sección 6.4, se presentará una medida de la bondad de ajuste que permite comparar modelos en los que *y* aparece tanto de forma lineal como de forma logarítmica.

Control de demasiados factores en un análisis de regresión

En muchos de los ejemplos vistos y seguramente en el análisis sobre el sesgo de variables omitidas en el capítulo 3, la preocupación ha sido omitir, de un modelo, factores importantes que puedan estar correlacionados con las variables independientes. También puede ocurrir que en un análisis de regresión se controlen *demasiadas* variables.

Si se le da demasiada importancia a la bondad de ajuste, puede ocurrir que en un modelo de regresión se controlen factores que no deberían ser controlados. Para evitar este error, hay que recordar la interpretación *ceteris paribus* de los modelos de regresión múltiple.

Para ilustrar esto, suponga que se hace un estudio para evaluar el impacto de los impuestos estatales a la cerveza sobre los accidentes de tránsito fatales. La idea es que con impuestos más altos a la cerveza se reducirá el consumo de alcohol y, por consiguiente, la posibilidad de conducir bajo los efectos del alcohol, dando como resultado una reducción de los accidentes de tránsito fatales. Para medir el efecto *ceteris paribus* de los impuestos sobre los accidentes fatales, los

accidentes fatales (*fatalities*) pueden modelarse como función de diversos factores, entre los que se encuentra el *impuesto* (*tax*) a la cerveza:

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 miles + \beta_3 percmale + \beta_4 perc16\_21 + \ldots,$$

donde

$miles$ = cantidad total de millas recorridas conduciendo.
$percmale$ = porcentaje de hombres en la población de un estado.
$perc16\_21$ = porcentaje de personas entre 16 y 21 años de edad en la población.

Observe que no se ha incluido una variable que mida el consumo de cerveza *per cápita*. ¿Se está cometiendo un error de omisión de variables? La respuesta es no. Si en esta ecuación se controla el consumo de cerveza, entonces ¿cómo afectarían los impuestos a la cerveza los accidentes de tránsito fatales? En la ecuación

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \ldots,$$

β_1 mide la diferencia en los accidentes fatales debido a un aumento de un punto porcentual en los *impuestos*, manteniendo constante el consumo de cerveza (*beercons*). No es fácil entender por qué esto podría ser de interés. No se deben controlar las diferencias en *beercons* entre los estados, a menos que se desee probar algún tipo de efecto indirecto de los impuestos a la cerveza. Otros factores, tales como distribución del género y de la edad sí deben ser controlados.

Para ver otro ejemplo, suponga que, en un país en desarrollo, se desea estimar el efecto del uso de pesticidas por los agricultores sobre el gasto en salud familiar. Además de las cantidades de pesticidas usadas, ¿debe incluirse, como variable explicativa, la cantidad de visitas al médico? No. Los gastos en salud comprenden las visitas al médico y lo que se desea es captar todos los efectos del uso de pesticidas sobre los gastos en salud. Si se incluye, como variable explicativa, la cantidad de visitas al médico, entonces sólo se estarán midiendo los efectos del uso de los pesticidas sobre gastos en la salud distintos a las visitas al médico. Tiene más sentido usar la cantidad de visitas al médico como variable dependiente en otra regresión contra las cantidades de pesticidas.

Los anteriores son ejemplos de lo que podría llamarse un **sobrecontrol** de factores en la regresión múltiple. Esto suele ocurrir como resultado de la preocupación por los sesgos potenciales que pueden surgir al dejar fuera variables explicativas importantes. Pero es importante recordar la naturaleza *ceteris paribus* de la regresión múltiple. En algunas situaciones no tiene caso mantener ciertos factores fijos debido precisamente a que variarán cuando cambie una variable de política.

Por desgracia, el asunto de si controlar o no ciertos factores no siempre es claro. Por ejemplo, Betts (1995) estudió el efecto de la calidad del bachillerato sobre los ingresos subsecuentes. Él indica que, si una mejor calidad de la escuela da como resultado más educación, entonces controlar en la regresión la educación junto con medidas de la calidad subestima el rendimiento de la calidad. Betts realiza el análisis con y sin años de educación en la ecuación para obtener un intervalo de los efectos estimados de la calidad de la educación.

Para ver de manera explícita cómo enfocarse en R-cuadradas altas puede conducir a problemas, considere el ejemplo del precio de la vivienda visto en la sección 4.5 que ilustra la prueba de hipótesis múltiples. En ese caso se deseaba probar lo razonable de los avalúos de los precios de las viviendas. Se regresó log(*price*) sobre log(*assess*), log(*lotsize*), log(*sqrft*) y *bdrms* y se probó si las últimas tres variables tenían coeficientes poblacionales iguales a cero mientras que log(*assess*) tenía coeficiente igual a uno. Pero, ¿qué pasa si se modifica el objetivo del análisis y se estima un *modelo de precios hedónicos*, que permita obtener los valores marginales de distintos atributos de las viviendas? ¿Se debe incluir log(*assess*) en la ecuación? La R-cuadrada

ajustada de la regresión con log(*assess*) es .762, mientras que la R-cuadrada ajustada sin él es .630. Basándose sólo en la bondad de ajuste debería incluirse log(*assess*). Pero esto es incorrecto si lo que se quiere determinar son los efectos del tamaño del terreno, de la superficie de la vivienda en pies cuadrados y de la cantidad de recámaras sobre el valor de la vivienda. Incluir log(*assess*) en la ecuación equivale a mantener fija una medida del valor y preguntar después cuánto modificará una recámara más otra medida del valor. Esto no tiene sentido para la evaluación de los atributos de una vivienda.

Si se recuerda que modelos distintos tienen propósitos diferentes y se atiende a la interpretación *ceteris paribus* de la regresión, no se incluirán factores inadecuados en un modelo de regresión.

Adición de regresores para reducir la varianza del error

Se acaban de ver algunos ejemplos en los que ciertas variables independientes no deben ser incluidas en un modelo de regresión, aun cuando estas variables estén correlacionadas con la variable independiente. De acuerdo con lo visto en el capítulo 3, se sabe que agregar una variable independiente a la regresión puede exacerbar el problema de multicolinealidad. Por otro lado, como se está sacando algo del término de error, adicionar una variable reduce, por lo general, la varianza del error. Por lo común no se puede saber cuál será el efecto que domine.

Sin embargo, hay un caso que es claro: siempre deben incluirse las variables independientes que afecten a y y que *no estén correlacionadas* con todas las variables independientes de interés. ¿Por qué? Porque agregar esas variables no induce multicolinealidad en la población (y por tanto la multicolinealidad en la muestra será despreciable), pero sí reduce la varianza del error. En muestras grandes, los errores estándar de los estimadores de MCO se reducirán.

Como ejemplo, considérese la estimación de la demanda individual de cerveza como función del precio promedio de la cerveza en un condado o municipio. Será razonable suponer que las características individuales no están correlacionadas con el nivel de precios en el condado y de esta manera una regresión simple del consumo de cerveza sobre el precio en el condado será suficiente para estimar el efecto del precio sobre la demanda individual. Pero se puede obtener un estimador más preciso de la elasticidad precio de la demanda de cerveza mediante la inclusión de características individuales, tales como edad y cantidad de educación. Si estos factores que afectan la demanda no están correlacionados con el precio, entonces el error estándar del coeficiente del precio será menor, por lo menos cuando se tengan muestras grandes.

Para ver otro ejemplo, considere el financiamiento para equipo de cómputo mencionado al inicio de la sección 6.3. Si, además de la variable financiamiento, se controlan otros factores que puedan explicar el promedio general de calificaciones (GPA), tal vez se podrá obtener un estimado más preciso del efecto del financiamiento. Medidas del promedio general de calificaciones en el bachillerato y del ranking, puntuaciones de los exámenes de admisión a la universidad SAT y ACT (por sus siglas en inglés) y variables de antecedentes familiares son buenos candidatos. Dado que los montos del financiamiento son asignados de forma aleatoria, todas las demás variables de control no están correlacionadas con el monto del financiamiento; en la muestra, la multicolinealidad entre el monto del financiamiento y otras variables independientes deberá ser mínima. Pero la adición de los controles extra puede reducir de manera significativa la varianza del error, conduciendo a una estimación más precisa del efecto del financiamiento. Recuerde que aquí el problema no es insesgamiento: se agregue o no el desempeño en el bachillerato y las variables de antecedentes familiares se obtiene un estimador insesgado y consistente.

Por desgracia, en las ciencias sociales los casos en los que se tiene información adicional sobre variables explicativas que no estén correlacionadas con las variables explicativas de interés son raros. Pero vale la pena recordar que cuando existan estas variables, pueden incluirse en el modelo para reducir la varianza del error sin inducir multicolinealidad.

6.4 Predicción y análisis de residuales

En el capítulo 3 se definieron los valores predichos o ajustados de MCO, así como los residuales de MCO. Las **predicciones** son seguramente útiles, pero están sujetas a variaciones de muestreo, debido a que se obtienen empleando los estimadores de MCO. Así, en esta sección se muestra cómo obtener intervalos de confianza para una predicción a partir de la línea de regresión de MCO.

De acuerdo con lo visto en los capítulos 3 y 4, se sabe que los residuales se usan para obtener la suma de los residuales cuadrados y la *R*-cuadrada, de manera que son importantes para la bondad de ajuste y para las pruebas. Algunas veces los economistas estudian los residuales de determinadas observaciones para obtener información acerca de individuos (empresas, viviendas, etc.) de la muestra.

Intervalos de confianza para predicciones

Suponga que se tiene la ecuación estimada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad \textbf{6.27}$$

Cuando se sustituyen valores específicos de las variables independientes, se obtiene una predicción para y, la cual es una estimación del *valor esperado* de y dados los valores específicos de las variables explicativas. Haciendo hincapié, sean c_1, c_2, …, c_k valores específicos de cada una de las k variables independientes, que pueden corresponder o no a datos reales de la muestra. El parámetro que se desea estimar es

$$\begin{aligned} \theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k \\ &= \mathrm{E}(y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k). \end{aligned} \qquad \textbf{6.28}$$

El estimador de θ_0 es

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k. \qquad \textbf{6.29}$$

En la práctica, esto es fácil de calcular. Pero, ¿y si se desea tener alguna medida de la incertidumbre de este valor predicho? Lo natural es construir un intervalo de confianza para θ_0, que esté centrado en $\hat{\theta}_0$.

Para obtener un intervalo de confianza para θ_0, se necesita un error estándar de $\hat{\theta}_0$. Entonces, con un número de *gl* grande, se puede construir un intervalo de confianza de 95% empleando la regla $\hat{\theta}_0 \pm 2 \cdot \mathrm{ee}(\hat{\theta}_0)$. (Como siempre, pueden emplearse los percentiles exactos de una distribución *t*.)

¿Cómo se obtiene el error estándar de $\hat{\theta}_0$? Este es el mismo problema encontrado en la sección 4.4: se necesita obtener un error estándar de una combinación lineal de los estimadores de MCO. Aquí, el problema es incluso más complicado, porque en general todos los estimadores de MCO aparecen en $\hat{\theta}_0$ (a menos que algunos de los c_j sean cero). No obstante, el mismo truco empleado en la sección 4.4 funciona aquí. Se escribe $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$ y esto se sustituye en la ecuación

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

con lo que se obtiene

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \ldots + \beta_k(x_k - c_k) + u. \qquad \textbf{6.30}$$

En otras palabras, a cada observación c_j se le resta x_j, y se vuelve a correr la regresión de

$$y_i \text{ sobre } (x_{i1} - c_1), \ldots, (x_{ik} - c_k), \; i = 1, 2, \ldots, n. \qquad \textbf{6.31}$$

El valor predicho en (6.29) y, lo que es más importante, su error estándar, se obtienen del *intercepto* (o constante) de la regresión (6.31).

Para dar un ejemplo, se obtendrá un intervalo de confianza para la predicción de una regresión del promedio general de calificaciones (GPA) en la universidad, en la que se usa información del bachillerato.

Ejemplo 6.5

[Intervalo de confianza para el promedio general de calificaciones (GPA) en la universidad]

Empleando los datos del archivo GPA2.RAW se obtiene la ecuación siguiente para predecir el promedio general de calificaciones (GPA) en la universidad:

$$\begin{aligned} \widehat{colgpa} &= 1.493 + .00149\, sat - .01386\, hsperc \\ &\quad (0.075)\;\; (.00007)\qquad (.00056) \\ &\quad - .06088\, hsize + .00546\, hsize^2 \\ &\quad (.01650)\qquad\quad (.00227) \\ n &= 4{,}137,\; R^2 = .278,\; \bar{R}^2 = .277,\; \hat{\sigma} = .560, \end{aligned} \qquad \textbf{6.32}$$

donde las estimaciones se han dado con varios dígitos para reducir el error de redondeo. ¿Cuál es el promedio general de calificaciones (GPA) en la universidad, si $sat = 1{,}200$, $hsperc = 30$ y $hsize = 5$ (que significa 500)? Esto se obtiene fácilmente sustituyendo estos valores en la ecuación (6.32): $\widehat{colgpa} = 2.70$ (redondeado a dos cifras decimales). Por desgracia, la ecuación (6.32) no puede emplearse para obtener directamente un intervalo de confianza para *colgpa* en los valores dados de las variables independientes. Una manera sencilla de obtener un intervalo de confianza es definir un nuevo conjunto de variables independientes: $sat0 = sat - 1{,}200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$ y $hsizesq0 = hsize^2 - 25$. Al regresar *colgpa* sobre estas nuevas variables independientes, se obtiene

$$\begin{aligned} \widehat{colgpa} &= 2.700 + .00149\, sat0 - .01386\, hsperc0 \\ &\quad (0.020)\;\; (.00007)\qquad (.00056) \\ &\quad - .06088\, hsize0 + .00546\, hsizesq0 \\ &\quad (.01650)\qquad\quad (.00227) \\ n &= 4{,}137,\; R^2 = .278,\; \bar{R}^2 = .277,\; \hat{\sigma} = .560. \end{aligned}$$

La única diferencia entre esta regresión y la de (6.32) es el intercepto, que es la predicción que se busca, junto con su error estándar, .020. No es una coincidencia que los coeficientes de pendiente, sus errores estándar, la *R*-cuadrada, etc., sean las mismas que antes; esto proporciona una manera de verificar que se hicieron las transformaciones correctas. Construir un intervalo de confianza de 95% para el GPA esperado es fácil: $2.70 \pm 1.96(.020)$ es decir aproximadamente de 2.66 a 2.74. Este intervalo de confianza es bastante estrecho debido al tamaño tan grande de la muestra.

Como la varianza del estimador del intercepto es la menor posible cuando todas las variables explicativas tienen media muestral cero (vea la pregunta 2.5 para el caso de la regresión simple), se sigue, de acuerdo con la regresión (6.31), que la varianza de la predicción es la menor posible en los valores medios de las x_j. (Es decir, $c_j = \bar{x}_j$ para toda j.) Este resultado no debe extrañar, ya que cerca del centro de los datos se tiene la mayor confianza en la línea de regresión. A medida que los valores de las c_j se alejan de las $\bar{x}_j$, Var($\hat{y}$) se vuelve cada vez mayor.

El método anterior permite colocar un intervalo de confianza alrededor de la estimación de MCO de $E(y|x_1, \ldots, x_k)$, para cualesquiera valores de las variables explicativas. En otras palabras, se obtiene un intervalo de confianza para el valor *promedio* de y en la subpoblación de un conjunto dado de variables independientes. Pero un intervalo de confianza para la persona promedio de la subpoblación no es lo mismo que un intervalo de confianza para una unidad específica (individuo, familia, empresa, etc.) de la población. Al formar un intervalo de confianza para un resultado desconocido de y, debe tomarse en cuenta otra fuente importante de variación: la varianza del error no observado, el cual mide la ignorancia de los factores no observados que afectan a y.

Sea y^0 el valor para el cual se desea construir un intervalo de confianza, al cual suele llamársele **intervalo de predicción**. Por ejemplo, y^0 puede representar, una persona o una empresa que no esté en la muestra original. Sean $x_1^0, \ldots, x_k^0$ los nuevos valores de las variables independientes, que se supone se observan, y sea u^0 el error no observado. Por tanto, se tiene

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0 + u^0. \qquad \textbf{6.33}$$

Como antes, la mejor predicción para y^0 es el valor esperado de y^0 dadas las variables explicativas, el cual se estima a partir de la línea de regresión de MCO: $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$. El **error de predicción** al emplear $\hat{y}^0$ para predecir y^0 es

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}^0. \qquad \textbf{6.34}$$

Ahora, $E(\hat{y}^0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_1^0 + E(\hat{\beta}_2)x_2^0 + \ldots + E(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$, debido a que las $\hat{\beta}_j$ son insesgadas. (Como siempre, todos estos valores esperados son condicionales en los valores muestrales de las variables independientes.) Como u^0 tiene media cero, $E(\hat{e}^0) = 0$. Se ha demostrado que el error de predicción esperado es cero.

Al determinar la varianza de $\hat{e}^0$, observe que u^0 no está correlacionada con ninguna de las $\hat{\beta}_j$, porque tampoco lo está con los errores de la muestra empleada para obtener las $\hat{\beta}_j$. De acuerdo con las propiedades básicas de la covarianza (ver apéndice B), u^0 y $\hat{y}^0$ no están correlacionados. Por tanto **varianza del error de predicción** (condicional en todos los valores muestrales de las variables independientes) es la suma de las varianzas:

$$\text{Var}(\hat{e}^0) = \text{Var}(\hat{y}^0) + \text{Var}(u^0) = \text{Var}(\hat{y}^0) + \sigma^2, \qquad \textbf{6.35}$$

donde $\sigma^2 = \text{Var}(u^0)$ es la varianza del error. En $\hat{e}^0$ hay dos fuentes de variación. La primera es el error de muestreo en $\hat{y}^0$, que surge debido a que se han estimado los β_j. Como cada $\hat{\beta}_j$ tiene una varianza proporcional a $1/n$, donde n es el tamaño de la muestra, Var($\hat{y}^0$) es proporcional a $1/n$. Esto significa que, con muestras grandes, Var($\hat{y}^0$) puede ser muy pequeña. En cambio, σ^2 es la varianza del error en la población: ésta no cambia con el tamaño de la muestra. En muchos ejemplos, σ^2 será el término dominante en (6.35).

Bajo los supuestos del modelo lineal clásico, las $\hat{\beta}_j$ y u^0 están distribuidas normalmente y entonces $\hat{e}^0$ tambien está distribuida normalmente (condicional a todos los valores muestrales de las variables explicativas). Antes, se describió cómo obtener un estimador insesgado de $\text{Var}(\hat{y}^0)$, y en el capítulo 3 se obtuvo el estimador insesgado de σ^2. Empleando estos estimadores, se puede definir el error estándar de $\hat{e}^0$ como

$$\text{ee}(\hat{e}^0) = \{[\text{ee}(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}. \qquad \textbf{6.36}$$

Empleando el mismo razonamiento para los estadísticos t de los $\hat{\beta}_j$, $\hat{e}^0/\text{ee}(\hat{e}^0)$ tiene una distribución t con $n - (k + 1)$ grados de libertad. Por tanto,

$$P[-t_{.025} \leq \hat{e}^0/\text{ee}(\hat{e}^0) \leq t_{.025}] = .95,$$

donde $t_{.025}$ es el percentil 97.5 de la distribución t_{n-k-1}. Si $n - k - 1$ es grande, recuerde que $t_{.025} \approx 1.96$. Al sustituir en $\hat{e}^0 = y^0 - \hat{y}^0$ y reordenar se obtiene un intervalo de predicción de 95% para y^0:

$$\hat{y}^0 \pm t_{.025} \cdot \text{ee}(\hat{e}^0); \qquad \textbf{6.37}$$

como de costumbre, excepto cuando el número de *gl* sea pequeño, una buena regla práctica es $\hat{y}^0 \pm 2\text{ee}(\hat{e}^0)$. Este intervalo es más amplio que el de $\hat{y}^0$ debido a la presencia de $\hat{\sigma}^2$ en (6.36); este intervalo suele ser mucho más amplio para reflejar los factores en u^0 que no se han controlado.

Ejemplo 6.6

[Intervalo de confianza para un GPA futuro]

Suponga que se desea un IC de 95% para el GPA futuro de un estudiante de bachillerato para el que $sat = 1{,}200$, $hsperc = 30$ y $hsize = 5$. En el ejemplo 6.5, se obtuvo un intervalo de confianza de 95% para el *promedio* de GPA de todos los estudiantes con las características particulares $sat = 1{,}200$, $hsperc = 30$ y $hsize = 5$. Ahora se quiere un intervalo de confianza de 95% para un estudiante *específico* con estas características. El intervalo de predicción de 95% debe tomar en cuenta la variación en el individuo, características no observadas que afectan el desempeño universitario. Ya se tiene todo lo que se necesita para obtener un IC para *colgpa*. $\text{ee}(\hat{y}^0) = .020$ y $\hat{\sigma} = .560$ y, de esta manera, de acuerdo con (6.36), $\text{ee}(\hat{e}^0) = [(.020)^2 + (.560)^2]^{1/2} \approx .560$. Observe lo pequeño que es $\text{ee}(\hat{y}^0)$ con relación a $\hat{\sigma}$: prácticamente toda la variación en $\hat{e}^0$ proviene de la variación en u^0. El IC de 95% es $2.70 \pm 1.96(.560)$ es decir, aproximadamente de 1.60 a 3.80. Este es un intervalo de confianza amplio y muestra que, con base en los factores incluidos en la regresión, no es posible señalar con exactitud el promedio general de calificaciones futuras de un individuo. [En cierto sentido, esta es una buena noticia, pues significa que el desempeño en el bachillerato y en el examen de admisión (SAT) no predeterminan el desempeño en la universidad.] Evidentemente, las características no observadas varían de manera amplia entre los individuos que tienen una misma puntuación observada en el SAT y un mismo promedio observado en el bachillerato.

Análisis de residuales

Algunas veces es útil examinar las observaciones individuales para ver si el verdadero valor de la variable dependiente es superior o inferior al valor predicho; es decir, examinan los residuales de las observaciones individuales. A este proceso se le llama **análisis de residuales**.

Los economistas examinan los residuales de una regresión con objeto de ayudar en la compra de una vivienda. El ejemplo siguiente sobre precios de la vivienda ilustra el análisis de residuales. El precio de una vivienda está relacionado con diversas características observables de la vivienda. Se puede hacer una lista con todas las características que se consideran importantes, tales como tamaño, cantidad de recámaras y de baños, etc. Se puede emplear una muestra de viviendas para estimar la relación entre el precio y los atributos, con lo que se obtiene un valor predicho y un valor real de cada vivienda. Después, pueden calcularse los residuales, $\hat{u}_i = y_i - \hat{y}_i$. La vivienda con el residual más negativo es, al menos con base en los factores controlados, la más subvalorada con relación a las características *observadas*. Por supuesto, un precio de venta sustancialmente menor a su precio predicho indicará que existen algunas características indeseables de la vivienda que no han sido tomadas en cuenta y que están contenidas en el error no observado. Además de obtener la predicción y el residual, también es útil calcular un intervalo de confianza para el que puede ser el precio de venta de la vivienda, empleando el método descrito en la ecuación (6.37).

Empleando los datos del archivo HPRICE1.RAW, se corre la regresión de *price* sobre *lotsize*, *sqrft* y *bdrms*. En la muestra de 88 viviendas, el residual más negativo es −120.206, que es el de la casa número 81 de la muestra. Por tanto, el precio solicitado por esta vivienda es $120,206 inferior a su precio predicho.

El análisis de residuales tiene otros muchos usos. Una manera de jerarquizar las escuelas de leyes es regresando el sueldo inicial promedio sobre diversas características de los estudiantes [por ejemplo, la puntuación promedio en el examen de admisión a una escuela de leyes (LSAT, por sus siglas en inglés), al GPA promedio, etc.] y obtener un valor predicho y un residual para cada escuela de leyes. La escuela de leyes con el mayor residual tiene el valor agregado predicho más alto. (Por supuesto, todavía es incierta la relación entre el sueldo inicial de un individuo y la media general de una escuela de leyes.) Estos residuales pueden usarse junto con los costos de asistir a cada escuela de leyes para determinar el mejor valor; esto requerirá un descuento adecuado de las ganancias futuras.

El análisis de residuales también es importante en decisiones legales. Un artículo de la revista *New York Times* titulado "Judge Says Pupil's Poverty, Not Segregation, Hurts Scores" ("El juez dice que la pobreza de los alumnos, no la segregación, afecta las calificaciones") (6/28/95) describe un importante caso legal. El problema era si el mal desempeño en exámenes estandarizados del distrito escolar de Hartford, con relación al desempeño de suburbios vecinos se debía a la mala calidad de las escuelas más segregadas. El juez concluyó que "la disparidad en las puntuaciones de los exámenes no indica que Hartford esté trabajando mal o de manera inadecuada con relación a la educación de sus estudiantes o que sus escuelas estén fallando, ya que las puntuaciones predichas de acuerdo con los factores socioeconómicos relevantes se encuentran en los niveles que eran de esperarse". Esta conclusión está basada en un análisis de regresión de las puntuaciones promedio o medias sobre las características socioeconómicas de varios distritos escolares en Connecticut. La conclusión del juez indica que dados los niveles de pobreza de los estudiantes en las escuelas de Hartford, las puntuaciones reales en los exámenes son similares a las predichas con un análisis de regresión: el residual correspondiente a Hartford no era suficientemente negativo para concluir que las escuelas mismas fueron la causa de las bajas puntuaciones en el examen.

Pregunta 6.5

¿Cómo emplearía usted el análisis de residuales para determinar qué atletas profesionales son remunerados de manera excesiva o insuficiente con relación a su desempeño?

Predicción de *y* cuando log(*y*) es la variable dependiente

Como en la economía empírica se emplea con tanta frecuencia la transformación que usa el logaritmo natural, esta subsección se dedica a la predicción de y cuando la variable dependiente

es log(y). Como subproducto se obtiene una medida de la bondad de ajuste para el modelo logarítmico, la cual puede compararse con la R-cuadrada del modelo lineal.

Para obtener una predicción es útil definir $logy = \log(y)$; con lo que se hace hincapié en que lo que predice el modelo es el logaritmo de y

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{6.38}$$

En esta ecuación, las x_j pueden ser transformaciones de otras variables; así, en el ejemplo de los sueldos de los CEO puede tenerse $x_1 = \log(sales)$, $x_2 = \log(mktval)$, $x_3 = ceoten$.

Dados los estimadores de MCO, ya se sabe cómo predecir $logy$ para cualesquiera valores de las variables independientes:

$$\widehat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad \textbf{6.39}$$

Ahora, como la función exponencial elimina el logaritmo, el primer intento para predecir y es exponenciar el valor predicho para log(y): $\hat{y} = \exp(\widehat{logy})$. Esto no funciona; en realidad subestimará de manera *sistemática* el valor esperado de y. En efecto, si el modelo (6.38) sigue los supuestos RLM.1 a RLM.6 del MLC, puede demostrarse que

$$\mathrm{E}(y|\boldsymbol{x}) = \exp(\sigma^2/2)\cdot\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k),$$

donde $\boldsymbol{x}$ denota las variables independientes y σ^2 es la varianza de u. [Si $u \sim \text{Normal}(0,\sigma^2)$, entonces el valor esperado de $\exp(u)$ es $\exp(\sigma^2/2)$.] Esta ecuación muestra que para predecir y se necesita un ajuste sencillo:

$$\hat{y} = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}), \qquad \textbf{6.40}$$

donde $\hat{\sigma}^2$ es simplemente el estimador insesgado de σ^2. Como siempre se reporta $\hat{\sigma}$, el error estándar de la regresión, obtener los valores predichos para y es sencillo. Como $\hat{\sigma}^2 > 0$, $\exp(\hat{\sigma}^2/2) > 1$. Cuando $\hat{\sigma}^2$ es grande, este factor de ajuste puede ser sustancialmente mayor que la unidad.

La predicción en (6.40) no es insesgada, pero es consistente. No hay predicciones insesgadas de y y, en muchos casos, la predicción de (6.40) funciona bien. Sin embargo, esta predicción se apoya en la normalidad del término del error, u. En el capítulo 5, se mostró que los MCO tienen propiedades deseables, aun cuando u no esté distribuido normalmente. Por tanto, es útil tener una predicción que no se apoye en la normalidad. Si simplemente se supone que u es independiente de las variables explicativas, entonces se tiene

$$\mathrm{E}(y|\boldsymbol{x}) = \alpha_0\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k), \qquad \textbf{6.41}$$

donde α_0 es el valor esperado de $\exp(u)$, el cual debe ser mayor que la unidad.

Dada una estimación $\hat{\alpha}_0$, se puede predecir y de la manera siguiente

$$\hat{y} = \hat{\alpha}_0\exp(\widehat{logy}), \qquad \textbf{6.42}$$

que una vez más simplemente debe exponenciar el valor predicho con el modelo logarítmico y multiplicar el resultado por $\hat{\alpha}_0$.

Se sugieren dos métodos para estimar α_0 sin el supuesto de normalidad. El primero se basa en $\alpha_0 = \mathrm{E}[\exp(u)]$. Para estimar α_0 se sustituye la esperanza poblacional por un promedio muestral

y después se sustituyen los errores no observados, u_i, por los residuales de MCO, $\hat{u}_i = \log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$. Esto conduce al estimador del método de momentos (vea el apéndice C)

$$\hat{\alpha}_0 = n^{-1}\sum_{i=1}^{n}\exp(\hat{u}_i). \qquad \textbf{6.43}$$

No debe extrañar que $\hat{\alpha}_0$ sea un estimador consistente de α_0, pero no es insesgado debido a que dentro de una función no lineal se ha sustituido u_i por $\hat{u}_i$. Esta versión de $\hat{\alpha}_0$ es un caso especial de lo que Duan (1983) llamó un estimador no paramétrico de retransformación **estimador smearing**. Como los residuales de MCO tienen promedio muestral cero, puede demostrarse que, para cualquier conjunto de datos, $\hat{\alpha}_0 > 1$. (Técnicamente, $\hat{\alpha}_0$ será igual a uno si todos los residuales de MCO son cero, pero esto no ocurre nunca en una aplicación interesante.) El que $\hat{\alpha}_0$ sea necesariamente mayor que uno es conveniente porque debe ser cierto que $\alpha_0 > 1$.

Otra estimación de α_0 se basa en una regresión simple a través del origen. Para ver cómo funciona, se define $m_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$, de manera que, de acuerdo con la ecuación (6.41), $\mathrm{E}(y_i|m_i) = \alpha_0 m_i$. Si las m_i pudieran observarse, mediante la regresión de y_i sobre m_i sin intercepto se obtendría un estimador insesgado de α_0. En lugar de esto, las β_j se sustituyen por sus estimaciones de MCO y se obtiene $\hat{m}_i = \exp(\widehat{logy}_i)$, donde, por supuesto, los $\widehat{logy}_i$ son los valores ajustados obtenidos de la regresión de $logy_i$ sobre $x_{i1}, \ldots, x_{ik}$ (con intercepto). Entonces $\check{\alpha}_0$ [para distinguirla de la $\hat{\alpha}_0$ de la ecuación (6.43)] es la pendiente estimada de MCO obtenida de la regresión simple de y_i sobre $\hat{m}_i$ (sin intercepto):

$$\check{\alpha}_0 = \left(\sum_{i=1}^{n}\hat{m}_i^2\right)^{-1}\left(\sum_{i=1}^{n}\hat{m}_i y_i\right). \qquad \textbf{6.44}$$

A $\check{\alpha}_0$ se le llamará la estimación de la regresión de α_0. Al igual que $\hat{\alpha}_0$, $\check{\alpha}_0$ es consistente pero no insesgada. Curiosamente, no se garantiza que, $\check{\alpha}_0$ sea mayor que uno, aunque en la mayoría de las aplicaciones lo es. Si $\check{\alpha}_0$ es menor que uno, y en especial si es mucho menor que uno, es muy probable que se viole el supuesto de independencia entre u y x_j. Si $\check{\alpha}_0 < 1$, una posibilidad es emplear la estimación de (6.43), aunque esto puede sólo estar enmascarando algún problema con el modelo lineal de $\log(y)$. A continuación se resumen los pasos:

PREDICCIÓN DE *y* CUANDO LA VARIABLE DEPENDIENTE ES log(*y*):

1. Obtener los valores ajustados, $\widehat{logy}_i$, y los residuales, $\hat{u}_i$, mediante la regresión de $logy$ sobre $x_1, \ldots, x_k$.
2. Obtener $\hat{\alpha}_0$ de la ecuación (6.43) o $\check{\alpha}_0$ de la ecuación (6.44).
3. Para valores dados de $x_1, \ldots, x_k$, obtener $\widehat{logy}$ mediante (6.42).
4. Obtener la predicción $\hat{y}$ mediante (6.42) (con $\hat{\alpha}_0$ o $\check{\alpha}_0$).

A continuación se muestra cómo predecir el sueldo de los CEO empleando este procedimiento.

Ejemplo 6.7

[Predicción del sueldo de los directores generales]

El modelo de interés es

$$\log(salary) = \beta_0 + \beta_1\log(sales) + \beta_2\log(mktval) + \beta_3 ceoten + u,$$

de manera que β_1 y β_2 son elasticidades y $100\cdot\beta_3$ es una semielasticidad. La ecuación estimada empleando el archivo CEOSAL2.RAW es

$$\widehat{lsalary} = 4.504 + .163\ lsales + .109\ lmktval + .0117\ ceoten \qquad \textbf{6.45}$$
$$(.257)\ \ (.039)\qquad (.050)\qquad (.0053)$$
$$n = 177,\ R^2 = .318,$$

donde, para mayor claridad *lsalary* denota el logaritmo de *salary*, y de manera similar *lsales* y *lmktval*. A continuación se obtiene $\hat{m}_i = \exp(\widehat{lsalary}_i)$ para cada una de las observaciones de la muestra.

La estimación no paramétrica de retransformación de Duan de (6.43) es aproximadamente $\hat{\alpha}_0 = 1.136$ y la estimación de la regresión de acuerdo con (6.44) es $\tilde{\alpha}_0 = 1.117$. Se puede usar cualquier estimación para predecir *salary* para cualesquiera valores de *sales*, *mktval* y *ceoten*. Se va a determinar la predicción para *sales* = 5,000 (lo que significa \$5 mil millones porque *sales* está en millones), *mktval* = 10,000 (o \$10 mil millones) y *ceoten* = 10. De acuerdo con (6.45), la predicción para *lsalary* es $4.504 + .163\cdot\log(5{,}000) + .109\cdot\log(10{,}000) + .0117(10) \approx 7.013$ y $\exp(7.013) \approx 1{,}110.983$. Empleando la estimación de α_0 obtenido con (6.43), el sueldo predicho es aproximadamente 1,262.077, es decir, \$1,262.077. Usando el estimado que se obtiene con (6.44), el sueldo predicho es aproximadamente \$1,240,968. Éstos difieren uno de otro en mucho menos de lo que cada uno difiere de la ingenua predicción de \$1,110,983.

Los métodos anteriores pueden usarse en la obtención de predicciones para determinar qué tan bien puede explicar y el modelo que usa $\log(y)$ como variable dependiente. Ya se tienen mediciones para los modelos en los que y es la variable dependiente: R-cuadrada y R-cuadrada ajustada. El objetivo es hallar una medida de la bondad de ajuste para el modelo $\log(y)$ que pueda compararse con una R-cuadrada de un modelo en el que y sea la variable dependiente.

Existen varias maneras de definir una medida de la bondad de ajuste después de transformar un modelo con $\log(y)$ para predecir y. Aquí se presenta un método que es fácil de realizar y con el que se obtiene el mismo valor ya sea que α_0 se estime como en (6.40), (6.43) o (6.44). Para motivar esta medida, recuerde que en la ecuación de regresión lineal estimada mediante MCO,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \qquad \textbf{6.46}$$

la R-cuadrada usual es simplemente el cuadrado de la correlación entre y_i y $\hat{y}_i$ (vea la sección 3.2). Ahora, si en lugar de esto se calculan valores ajustados de acuerdo con (6.42) —es decir, $\hat{y}_i = \hat{\alpha}_0 m_i$ para todas las observaciones i— entonces tiene sentido emplear, como una R-cuadrada, el cuadrado de la correlación entre y_i y estos valores ajustados. Como a la correlación no le afecta que se multiplique por una constante, no importa qué estimación de α_0 se emplee. En realidad, esta medida R-cuadrada para y [no para $\log(y)$] es precisamente el cuadrado de la correlación entre y_i y $\hat{m}_i$. Esto se puede comparar de manera directa con la R-cuadrada de la ecuación (6.46). [Como el cálculo de la R-cuadrada no depende de la estimación de α_0, no permite elegir entre (6.40), (6.43) y (6.44). Pero se sabe que (6.44) minimiza la suma de los residuales cuadrados entre y_i y $\hat{m}_i$ sin una constante. En otras palabras, dados los $\hat{m}_i$, $\tilde{\alpha}_0$ se elige de manera que produzca el mejor ajuste con base en la suma de los residuales cuadrados. En lo que se está interesado aquí es en elegir entre el modelo lineal para y y $\log(y)$, y de esta manera es adecuada una medida de R-cuadrada que no dependa de la manera en que se estima α_0.]

Ejemplo 6.8

[Predicción de los sueldos de los directores generales (CEO)]

Después de obtener los $\hat{m}_i$ simplemente se obtiene la correlación entre $salary_i$ y $\hat{m}_i$; ésta es .493. Su cuadrado es aproximadamente .243 y esta es una medida de qué tan bien explica el modelo logarítmico la variación de *salary* no de log(*salary*). [La R^2 obtenida con (6.45), .318, indica que el modelo logarítmico explica aproximadamente 31.8% de la variación en log(*salary*).]

Como modelo lineal alternativo, suponga que se estima un modelo con todas las variables en forma lineal:

$$salary = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u. \qquad \textbf{6.47}$$

La clave es que la variable dependiente es *salary*. En el lado derecho se podrían usar los logaritmos de *sales* o *mktval*, pero, si (*salary*) aparece de forma lineal, es más razonable tener todos los valores en dólares en forma lineal. La *R*-cuadrada de la estimación de esta ecuación empleando las mismas 177 observaciones es .201. De manera que el modelo logarítmico explica más de la variación en *salary*, y por tanto se prefiere a (6.47) por razones de la bondad de ajuste. El modelo logarítmico se prefiere también debido a que parece más realista y sus parámetros son más fáciles de interpretar.

Si en el modelo (6.38) se conserva la base completa de supuestos del modelo lineal clásico, con facilidad se pueden obtener intervalos de predicción para $y^0 = \exp(\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0 + u^0)$ una vez que se haya estimado el modelo lineal para log(*y*). Recuerde que $x_1^0, x_2^0, \ldots, x_k^0$ son valores conocidos y que u^0 es el error no observado que determina de manera parcial y^0. De acuerdo con la ecuación (6.37), un intervalo de predicción de 95% para $logy^0 = \log(y^0)$ es simplemente $\widehat{logy}^0 \pm t_{.025} \cdot \text{ee}(\hat{e}^0)$, donde $\text{ee}(\hat{e}^0)$ se obtiene mediante la regresión de log(*y*) sobre $x_1, \ldots, x_k$ empleando las *n* observaciones originales. Sean $c_l = -t_{.025} \cdot \text{ee}(\hat{e}^0)$ y $c_u = t_{.025} \cdot \text{ee}(\hat{e}^0)$ los límites inferior y superior del intervalo de predicción para $logy^0$. Es decir, $P(c_l \leq logy^0 \leq c_u) = .95$. Dado que la función exponencial es estrictamente creciente, también se tiene que $P[\exp(c_l) \leq \exp(logy^0) \leq \exp(c_u)] = .95$, es decir, $P[\exp(c_l) \leq y^0 \leq \exp(c_u)] = .95$. Por tanto, se pueden tomar $\exp(c_l)$ y $\exp(c_u)$ como los límites inferior y superior, respectivamente, de un intervalo de predicción de 95% para y^0. Para *n* grande, $t_{.025} = 1.96$, y de esta manera un intervalo de predicción de 95% para y^0 es $\exp[-1.96 \cdot \text{ee}(\hat{e}^0)] \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ a $\exp[-1.96 \cdot \text{ee}(\hat{e}^0)] \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$, donde $\mathbf{x}^0\hat{\boldsymbol{\beta}}$ es una abreviación de $\hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_k x_k^0$. Recuerde, los $\hat{\beta}_j$ y $\text{ee}(\hat{e}^0)$ se obtienen mediante la regresión con log(*y*) como variable dependiente. Como en (6.38) se asume la normalidad de *u*, para obtener una predicción puntual de y^0. probablemente se emplee (6.40). A diferencia de la ecuación (6.37), esta predicción puntual no se encontrará a la mitad entre los límites inferior y superior $\exp(c_l)$ y $\exp(c_u)$. Se pueden obtener distintos intervalos de predicción de 95% eligiendo distintas cantidades en la distribución t_{n-k-1}. Si $q_{\alpha 1}$ y $q_{\alpha 2}$ son cuantiles con $\alpha_2 - \alpha_1 = .95$, entonces se pueden elegir $c_l = q_{\alpha 1}\text{ee}(\hat{e}^0)$ y $c_u = q_{\alpha 2}\text{ee}(\hat{e}^0)$.

Como ejemplo, considérese la regresión del sueldo de CEO, donde la predicción sobre los mismos valores de *sales*, *mktval* y *ceoten* que en el ejemplo 6.7. El error estándar de la regresión para (6.43) es aproximadamente .505 y el error estándar de $\widehat{logy}^0$ es .075. Por tanto, empleando la ecuación (6.36), $\text{ee}(\hat{e}^0) \approx .511$; como en el ejemplo del promedio general de calificaciones (GPA), la varianza del error domina el error de estimación en los parámetros, aun cuando aquí el tamaño de la muestra sea de sólo 177. Un intervalo de predicción de 95% para $salary^0$ es $\exp[-1.96 \cdot (.511)] \exp(7.013)$ a $\exp[1.96 \cdot (.511)] \exp(7.013)$, es decir, aproximadamente 408.071 a 3,024.678, es decir $408.071 a $3,024.678. Este intervalo de predicción de 95%, tan amplio, para el sueldo de CEO, a los valores dados de ventas, valor de mercado y de antigüedad, indica que existen muchos más factores que no fueron incluidos en la regresión que determina el sueldo. Dicho sea de paso, la predicción puntual para el sueldo (*salary*), empleando (6.40), es aproximadamente $1,262, 075 —superior a las predicciones empleando las otras estimaciones de α_0 y más cercana al límite inferior que al límite superior del intervalo de predicción de 95 por ciento.

RESUMEN

En este capítulo se vieron algunos temas importantes del análisis de regresión múltiple.

En la sección 6.1 se mostró que cualquier cambio en las unidades de medición de una variable independiente, modifica los coeficientes de MCO como era de esperarse: si x_j se multiplica por c, su coeficiente se dividirá entre c. Si la variable dependiente se multiplica por c, *todos* los coeficientes de MCO se multiplicarán por c. Modificar las unidades de medición de una variable no afecta ni a los estadísticos t ni a los estadísticos F.

Se analizaron los coeficientes beta, los cuales miden los efectos de las variables independientes en unidades de desviaciones estándar. Los coeficientes beta se obtienen de la regresión usual de MCO después de transformar la variable dependiente y las variables independientes en valores-z.

Como ya se ha visto en varios ejemplos, la forma funcional logarítmica proporciona coeficientes que se interpretan como efectos porcentuales. En la sección 6.2 se analizaron sus ventajas adicionales. Se vio también cómo calcular el efecto porcentual exacto en caso de que en un modelo logarítmico lineal un coeficiente sea grande. Modelos con términos cuadráticos permiten efectos marginales, ya sea crecientes o decrecientes. Modelos con interacciones permiten que el efecto marginal de una variable explicativa dependa del nivel de otra variable explicativa.

Se introdujo la R-cuadrada ajustada, $\bar{R}^2$, como una alternativa a la R-cuadrada usual en la medición de la bondad de ajuste. Mientras que R^2 nunca puede disminuir cuando se agrega una variable más a la regresión, $\bar{R}^2$ sí penaliza la cantidad de regresores y puede disminuir cuando se agrega una variable independiente. Esto hace que $\bar{R}^2$ se prefiera cuando se trata de elegir entre modelos no anidados con cantidades diferentes de variables explicativas. En la comparación de modelos con variables dependientes diferentes no se puede usar ni R^2 ni $\bar{R}^2$. Sin embargo, como se mostró en la sección 6.4, es bastante fácil obtener medidas de la bondad de ajuste para elegir entre y y $\log(y)$ como variable dependiente.

En la sección 6.3 se analizó el problema algo sutil de confiar demasiado en R^2 o en $\bar{R}^2$ para llegar a un modelo final: es posible que en un modelo de regresión se controlen demasiados factores. Debido a esto es importante pensar de antemano en las especificaciones del modelo, en especial en la naturaleza *ceteris paribus* de la ecuación de regresión. Las variables explicativas que afectan a y y que no están correlacionadas con todas las demás variables explicativas pueden emplearse para reducir la varianza del error sin inducir multicolinealidad.

En la sección 6.4 se demostró cómo obtener un intervalo de confianza para una predicción hecha a partir de la línea de regresión de MCO. También se mostró cómo construir un intervalo de confianza para un valor futuro, desconocido de y.

Ocasionalmente se desea predecir y cuando en un modelo de regresión se ha usado $\log(y)$ como variable dependiente. En la sección 6.4 se explica un sencillo método para esto. Por último, algunas veces se tiene interés en conocer el signo y la magnitud de los residuales de determinadas observaciones. Para determinar si los valores predichos para ciertos miembros de una muestra son mucho mayores o mucho menores que los valores reales puede emplearse el análisis de residuales.

TÉRMINOS CLAVE

Análisis de residuales
Bootstrap
Coeficientes beta
Coeficientes estandarizados
Efecto de interacción
Error de predicción
Error estándar bootstrap
Estimador no paramétrico de retransformación (estimador smearing)
Funciones cuadráticas
Intervalo de predicción
Métodos de remuestreo
Modelos no anidados
Predicciones
R-cuadrada ajustada
R-cuadrada poblacional
Sobrecontrol
Varianza del error de predicción

PROBLEMAS

6.1 La ecuación siguiente se estimó empleando los datos del archivo CEOSAL1.RAW:

$$\widehat{\log(salary)} = 4.322 + .276\log(sales) + .0215\,roe - .00008\,roe^2$$
$$(.324)\ \ (.033)\qquad\qquad (.0129)\qquad (.00026)$$
$$n = 209,\ R^2 = .282.$$

Esta ecuación permite que *roe* tenga un efecto decreciente sobre log(*salary*). ¿Es esto, en general, necesario? Explique por qué.

6.2 Sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ las estimaciones de MCO de la regresión de y_i sobre $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. Dadas constantes distintas de cero $c_1, \ldots, c_k$, argumente que el intercepto y las pendientes de MCO de la regresión de $c_0 y_i$ sobre $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$, están dadas por $\tilde{\beta}_0 = c_0\hat{\beta}_0$, $\tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Sugerencia*: use el hecho de que las $\hat{\beta}_j$ resuelven las condiciones de primer orden en (3.13), y que las $\tilde{\beta}_j$ deben resolver las condiciones de primer orden relacionadas con la reescalación de las variables dependiente e independiente.]

6.3 Empleando los datos del archivo RDCHEM.RAW, mediante MCO se obtuvo la ecuación siguiente:

$$\widehat{rdintens} = 2.613 + .00030\,sales - .0000000070\,sales^2$$
$$(.429)\ \ (.00014)\qquad\qquad (.0000000037)$$
$$n = 32,\ R^2 = .1484.$$

i) ¿En qué punto se vuelve negativo el efecto de *sales* sobre *rdintens*?

ii) ¿Conservaría usted el término cuadrático del modelo? Explique.

iii) Defina *salesbil* como las ventas medidas en miles de millones de dólares: *salesbil* = *sales*/1,000. Escriba de nuevo la ecuación estimada con *salesbil* y $salesbil^2$ como variables independientes. No olvide dar los errores estándar y la *R*-cuadrada. [*Sugerencia*: observe que $salesbil^2 = sales^2/(1{,}000)^2$.]

iv) ¿Qué ecuación prefiere con objeto de dar los resultados?

6.4 El modelo siguiente permite que el rendimiento de la educación sobre el salario dependa de la cantidad total de educación de los dos padres, denominada *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ\cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

i) Muestre que, de forma decimal, el rendimiento de un año más de educación en este modelo es

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

¿Qué signo espera que tenga β_2? ¿Por qué?

ii) Empleando los datos del archivo WAGE2.RAW, la ecuación estimada es

$$\widehat{\log(wage)} = 5.65 + .047\ educ + .00078\ educ{\cdot}pareduc + .019\ exper + .010\ tenure$$
$$(.13)\ \ (.010)\ \ \ \ (.00021)\ \ \ \ (.004)\ \ \ \ (.003)$$
$$n = 722,\ R^2 = .169.$$

(Sólo 722 observaciones contienen información completa sobre la educación de los padres.) Interprete el coeficiente del término de interacción. Puede ayudar elegir dos valores específicos para *pareduc* —por ejemplo, *pareduc* = 32 si ambos padres tienen educación universitaria o *pareduc* = 24 si los dos padres tienen bachillerato— y de esta manera estimar el rendimiento de *educ*.

iii) Si a esta ecuación se le agrega *pareduc* como una variable aparte, se obtienen:

$$\widehat{\log(wage)} = 4.94 + .097\ educ + .033\ pareduc - .0016\ educ{\cdot}pareduc + .020\ exper + .010\ tenure$$
$$(.38)\ \ (.027)\ \ \ \ (.017)\ \ \ \ (.0012)\ \ \ \ (.004)\ \ \ \ (.003)$$
$$n = 722,\ R^2 = .174.$$

¿Depende ahora el rendimiento estimado de la educación positivamente de la educación de los padres? Pruebe la hipótesis nula de que el rendimiento de la educación no depende de la educación de los padres.

6.5 En el ejemplo 4.2, en donde el porcentaje de estudiantes que obtienen una puntuación aprobatoria en el examen del décimo grado (*math*10) es la variable dependiente, ¿tiene sentido incluir *sci11* —porcentaje de estudiantes del grado undécimo que aprueban el examen de ciencias— como una variable explicativa más?

6.6 Cuando a la ecuación estimada en (6.19) se le agregan $atndrte^2$ y $ACT{\cdot}atndrte$ la R-cuadrada se vuelve .232. Al nivel de significancia de 10%, ¿son estos términos adicionales conjuntamente significativos? ¿Los incluiría usted en el modelo?

6.7 Las siguientes tres ecuaciones se estimaron empleando 1,534 observaciones del archivo 401K.RAW:

$$\widehat{prate} = 80.29 + 5.44\ mrate + .269\ age - .00013\ totemp$$
$$(.78)\ \ (.52)\ \ \ \ (.045)\ \ \ \ (.00004)$$
$$R^2 = .100,\ \bar{R}^2 = .098.$$

$$\widehat{prate} = 97.32 + 5.02\ mrate + .314\ age - 2.66\ \log(totemp)$$
$$(1.95)\ \ (0.51)\ \ \ \ (.044)\ \ \ \ (.28)$$
$$R^2 = .144,\ \bar{R}^2 = .142.$$

$$\widehat{prate} = 80.62 + 5.34\ mrate + .290\ age - .00043\ totemp + .0000000039\ totemp^2$$
$$(.78)\ \ (.52)\ \ \ \ (.045)\ \ \ \ (.00009)\ \ \ \ (.0000000010)$$
$$R^2 = .108,\ \bar{R}^2 = .106.$$

¿Cuál de estos tres modelos prefiere usted? ¿Por qué?

6.8 Suponga que se desea estimar el efecto del alcohol (*alcohol*) sobre el promedio general de calificaciones en la universidad (*colGPA*). Además de la información acerca del consumo del alcohol y el promedio general de calificaciones, también se obtiene información sobre la asistencia a clases (porcentaje de asistencia a clases, que se denomina *attend*). Se cuenta también con la calificación en una prueba estandarizada *SAT*) y con el promedio general de calificaciones en el bachillerato (*hsGPA*).

i) En un modelo de regresión múltiple, ¿debe incluirse *attend* además de *alcohol* como variables explicativas en un modelo de regresión múltiple? (Reflexione sobre cómo se interpretaría β_{alcohol}.)

ii) ¿Hay que incluir *SAT* y *hsGPA* como variables explicativas? Explique.

6.9 Si se empieza en (6.38) bajo los supuestos del MLC, suponiendo que n es grande e ignorando el error de estimación en las $\hat{\beta}_j$, un intervalo de predicción de 95% para y^0 es $[\exp(-1.96\hat{\sigma})\exp(\widehat{logy}^0), \exp(1.96\hat{\sigma})\exp(\widehat{logy}^0)]$. La predicción puntual para y^0 es $\hat{y}^0 = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}^0)$.

i) ¿Para qué valores de $\hat{\sigma}$ estará la predicción puntual en el intervalo de predicción de 95%? ¿Es posible que se satisfaga esta condición en la mayor parte de las aplicaciones?

ii) Compruebe que la condición del inciso i) se verifica en el ejemplo del sueldo de CEO.

EJERCICIOS EN COMPUTADORA

C6.1 Del archivo KIELMC.RAW emplee los datos de 1981 para responder las preguntas siguientes. Estos datos son de viviendas vendidas durante 1981 en North Andover, Massachusetts; 1981 fue el año en que se inició la construcción de un incinerador local de basura.

i) Para estudiar el efecto de la ubicación del incinerador sobre los precios de la vivienda, considere el modelo de regresión simple

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

donde *price* es el precio de la vivienda en dólares y *dist* es la distancia de la vivienda al incinerador, medida en pies. Interpretando esta ecuación de manera causal, ¿que signo espera usted que tenga β_1 si la presencia del incinerador hace decrecer el precio de la vivienda? Estime esta ecuación e interprete los resultados.

ii) Al modelo de regresión simple del inciso i) agregue las variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths* y *age*, donde *intst* es la distancia de la vivienda a la carretera interestatal, *area* es el área de la vivienda en pies cuadrados, *land* es el tamaño del terreno en pies cuadrados, *rooms* es la cantidad de habitaciones, *baths* es la cantidad de baños y *age* la antigüedad de la vivienda en años. Ahora, ¿qué concluye usted acerca del efecto del incinerador? Explique por qué los incisos i) y ii) dan resultados contradictorios.

iii) Al modelo del inciso ii) agregue la variable $[\log(intst)]^2$. ¿Qué pasa ahora? ¿Qué concluye usted acerca de la importancia de la forma funcional?

iv) ¿Es significativo el cuadrado de log(*dist*) cuando se agrega al modelo del inciso iii)?

C6.2 Para este ejercicio emplee los datos del archivo WAGE1.RAW.

i) Use MCO para estimar la ecuación

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

y dé los resultados empleando el formato acostumbrado.

ii) ¿Es $exper^2$ estadísticamente significativa al nivel de 1%?

iii) Empleando la aproximación

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

determine el rendimiento aproximado que tiene el quinto año de experiencia. ¿Cuál es el rendimiento aproximado del vigésimo año de experiencia?

iv) ¿Cuál es el valor de *exper* al que más experiencia disminuye el log(*wage*) predicho? ¿En esta muestra cuántas personas tienen una experiencia mayor que ese nivel?

C6.3 Considere un modelo en el que el rendimiento de la educación depende de la cantidad de experiencia de trabajo (y viceversa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$

i) Muestre que el rendimiento de un año más de educación (en forma decimal, manteniendo *exper* constante, es $\beta_1 + \beta_3 exper$.

ii) Establezca la hipótesis nula de que el rendimiento de la educación no depende del nivel de *exper*. ¿Cuál piensa que sea la alternativa adecuada?

iii) Emplee los datos del archivo WAGE2.RAWpara probar la hipótesis nula del inciso ii) contra la alternativa dada por usted.

iv) Sea θ_1 el rendimiento de la educación (en forma decimal) cuando $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Obtenga $\hat{\theta}_1$ y un intervalo de confianza de 95% para θ_1. (*Sugerencia*: escriba $\beta_1 = \theta_1 - 10\beta_3$ y sustitúyalo en la ecuación; después ordénelo. Esto da la regresión para obtener el intervalo de confianza para θ_1.)

C6.4 Para hacer este ejercicio, emplee los datos del archivo GPA2.RAW.

i) Estime el modelo

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

donde *sat* es el puntaje del examen de admisión a la universidad y *hsize* es el tamaño del grupo que termina sus estudios de bachillerato (en miles); dé los resultados en la forma habitual. ¿Es el término cuadrático estadísticamente significativo?

ii) Empleando la ecuación estimada en el inciso i), ¿cuál es el tamaño "óptimo" para un grupo de bachillerato? Justifique su respuesta.

iii) ¿Es representativo este análisis del desempeño de *todos* los estudiantes que terminan bachillerato? Explique.

iv) Encuentre el tamaño óptimo de una escuela de bachillerato, empleando log(*sat*) como variable dependiente. ¿Es esto muy distinto a lo que obtuvo en el inciso ii)?

C6.5 Para hacer este ejercicio, emplee los datos del archivo HPRICE1.RAW.

i) Estime el modelo

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

y dé los resultados en el formato habitual de MCO.

ii) Determine el valor que se predice para log(*price*), cuando *lotsize* = 20,000, *sqrft* = 2,500 y *bdrms* = 4. Empleando los métodos de la sección 6.4, halle el valor que se predice para *price* a estos mismos valores de la variable explicativa.

iii) Si se trata de explicar la variación de *price*, diga si prefiere el modelo del inciso i) o el modelo

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u.$$

C6.6 Para hacer este ejercicio, emplee los datos del archivo VOTE1.RAW.

i) Considere un modelo con una interacción entre los gastos:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

¿Cuál es el efecto parcial de *expendB* sobre *voteA*, cuando se mantienen constantes *prtystrA* y *expendA*? ¿Cuál es el efecto parcial de *expendA* sobre *voteA*? ¿Es obvio el signo que se espera para β_4?

ii) Estime la ecuación del inciso i) y dé los resultados de la manera habitual. ¿Es estadísticamente significativo el término de interacción?

iii) Determine el promedio de *expendA* en la muestra. Fije *expendA* en 300 (que significa \$300,000). ¿Cuál es el efecto estimado de \$100,000 más gastados por el candidato B sobre *voteA*? ¿Es grande este efecto?

iv) Ahora fije *expendB* en 100. ¿Cuál es el efecto estimado de $\Delta expendA = 100$ sobre *voteA*? ¿Es esto razonable?

v) Ahora, estime un modelo en el que la interacción se remplace por *shareA*, la participación porcentual del candidato A en los gastos de campaña. ¿Tiene sentido mantener constantes *expendA* y *expendB* y variar *shareA*?

vi) (Se requiere cálculo.) En el modelo del inciso v), determine el efecto parcial de *expendB* sobre *voteA*, manteniendo constantes *prtystrA* y *expendA*. Evalúe este modelo cuando *expendA* = 300 y *expendB* = 0 y comente los resultados.

C6.7 Para hacer este ejercicio, emplee los datos del archivo ATTEND.RAW.

i) En el modelo del ejemplo 6.3 argumente que

$$\Delta stndfnl/\Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Emplee la ecuación 6.19 para estimar el efecto parcial cuando *priGPA* = 2.59 y *atndrte* = 82. Interprete su estimación.

ii) Muestre que la ecuación puede escribirse como

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u,$$

donde $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. (Observe que el intercepto ha cambiado, pero esto no tiene importancia.) Use esto para obtener el error estándar de $\hat{\theta}_2$ de acuerdo con el inciso i).

iii) Suponga que en lugar de *priGPA*(*atndrte* − 82), emplea (*priGPA* − 2.59)· (*atndrte* − 82). ¿Cómo interpreta ahora los coeficientes de *atndrte* y de *priGPA*?

C6.8 Para hacer este ejercicio, emplee los datos del archivo HPRICE.RAW.

i) Estime el modelo

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

y dé los resultados de la manera habitual, dando también el error estándar de regresión. Obtenga el precio que se predice si se sustituye $lotsize = 10{,}000$, $sqrft = 2{,}300$ y $bdrms = 4$; redondee el precio obtenido al dólar más cercano.

ii) Corra una regresión que le permita colocar un intervalo de confianza de 95% en torno del valor predicho en el inciso i). Observe que su predicción diferirá un poco debido al error de redondeo.

iii) Sea $price^0$ el precio futuro no conocido al que se vende la vivienda con las características empleadas en los incisos i) y ii). Encuentre un IC de 95% para $price^0$ y analice la amplitud de este intervalo de confianza.

C6.9 La base de datos NBASAL.RAW contiene información de los sueldos y estadísticas sobre las carreras de 269 jugadores de la National Basketball Association (NBA).

i) Estime un modelo que relacione puntos por juego (*points*) con años en la liga (*exper*), edad (*age*), y años que se ha jugado en la universidad (*coll*). Incluya un término cuadrático en *exper*; las demás variables deben aparecer en forma lineal. Dé los resultados en la forma habitual.

ii) Manteniendo constantes los años en la universidad (*coll*) y la edad (*age*), ¿a qué valor de la experiencia un año adicional de ésta reduce los puntos por juego? ¿Es esto razonable?

iii) ¿Por qué piensa usted que *coll* tenga un coeficiente negativo y estadísticamente significativo? (*Sugerencia:* los jugadores de la NBA pueden ser reclutados antes de terminar sus carreras universitarias e incluso directamente al salir del bachillerato.)

iv) Agregue a la ecuación un término cuadrático en *age*. ¿Es necesario ese término? ¿Qué parece implicar este término acerca de los efectos de *age*, una vez controlado experiencia y educación?

v) Ahora regrese log(*wage*) sobre *points*, *exper*, $exper^2$, *age* y *coll*. Dé los resultados en el formato habitual.

vi) Pruebe si en la regresión del inciso v) son *age* y *coll* conjuntamente significativas. ¿Qué implica esto acerca de si la edad (*age*) y la educación tienen efectos separados sobre el salario (*wage*) una vez tomadas en cuenta la productividad y la antigüedad?

C6.10 Para hacer este ejercicio, emplee los datos del archivo BWGHT2.RAW.

i) Mediante MCO estime la ecuación

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

donde *bwght* es el peso de los niños al nacer y *npvis* es el número de visitas prenatales al médico; dé los resultados de la manera habitual. ¿Es significativo el término cuadrático?

ii) Muestre que, de acuerdo con la ecuación del inciso i), el número de visitas prenatales que maximiza log(*bwght*) se estima que es alrededor de 22. En la muestra, ¿cuántas mujeres tienen por lo menos 22 visitas prenatales?

iii) ¿Es razonable que el peso al nacer disminuya con más de 22 visitas prenatales? Explique.

iv) Agregue a esta ecuación la edad de la madre, empleando una forma funcional cuadrática. Manteniendo constante *npvis* ¿cuál es la edad de la madre en la que el peso del niño al nacer alcanza su máximo? ¿Qué proporción de las mujeres de la muestra tienen una edad mayor a la "óptima"?

v) ¿Considera usted que la edad de la madre y el número de visitas prenatales explican mucho de la variación en log(*bwght*)?

vi) Empleando términos cuadráticos tanto para *npvis* como para *age* (edad), diga si para predecir *bwght* es mejor emplear el logaritmo natural de *bwght* o *bwght* en nivel original.

C6.11 Para verificar algunas de las afirmaciones hechas en la sección 6.3 emplee el archivo APPLE.RAW.

i) Corra la regresión de *ecolbs* sobre *ecoprc*, *regprc* y dé los resultados de la manera habitual, dando también la *R*-cuadrada y la *R*-cuadrada ajustada. Interprete los coeficientes de las variables del precio y haga un comentario sobre sus signos y magnitudes.

ii) ¿Son estadísticamente significativas las variables del precio? Dé los valores-*p* de las pruebas *t* individuales.

iii) ¿Cuál es el rango de los valores ajustados para *ecolbs*? ¿En qué proporción de la muestra se tiene $ecolbs = 0$? Analice.

iv) ¿Considera usted que las variables del precio, juntas, explican suficiente la variación en *ecolbs*? Explique.

v) A la regresión del inciso i) agregue las variables *faminc*, *hhsize* (tamaño de la familia), *educ* y *age* (edad). Encuentre el valor-*p* para su significancia conjunta. ¿Qué concluye usted?

C6.12 Emplee el subconjunto del archivo 401KSUBS.RAW con $fsize = 1$; esto restringe el análisis a los hogares de una sola persona; vea también el ejercicio para computadora C4.8.

i) ¿Cuál es la edad de las personas más jóvenes en esta muestra? ¿Cuántas personas tienen esa edad?

ii) En el modelo

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

¿cuál es la interpretación literal de β_2? ¿Tiene mucho interés en sí misma?

iii) Estime el modelo del inciso ii) y dé los resultados de la manera habitual. ¿Le preocupa que el coeficiente de *age* (edad) sea negativo? Explique.

iv) Dado que las personas más jóvenes de la muestra tienen 25 años, es razonable pensar que, dado un determinado nivel de ingreso, la menor cantidad promedio de activo financiero neto es a la edad de 25 años. Recuerde que el efecto parcial de *age* sobre *nettfa* es $\beta_2 + 2\beta_3 age$, de manera que este efecto parcial a la edad de 25 años es $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; llámese a esto θ_2. Determine $\hat{\theta}_2$ y obtenga el valor-*p* de dos colas para probar H_0: $\theta_2 = 0$. Debe concluir que $\hat{\theta}_2$ es pequeño y estadísticamente muy poco significativo. [*Sugerencia:* una manera de hacer esto es estimar el modelo $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, donde el intercepto, α_0, es diferente de β_0. Existen también otras maneras de hacerlo.]

v) Dado que la evidencia contra H_0: $\theta_2 = 0$ es muy débil, iguálela a cero y estime el modelo

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

En términos de la bondad de ajuste, ¿es mejor este modelo que el del inciso ii)?

vi) Dada la ecuación estimada en el inciso v), haga $inc = 30$ (aproximadamente, el valor promedio) y grafique la relación entre *nettfa* y *age*, pero sólo para $age \geq 25$. Describa lo que ve.

vii) Verifique si es necesario incluir un término cuadrático para *inc*.

C6.13 Para este ejercicio emplee los datos del archivo MEAP00_01.

i) Estime mediante MCO el modelo

$$math4 = \beta_0 + \beta_2 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

donde *math*4 es el porcentaje de aprobación en matemáticas en 4o. grado, *lexppp* es el logaritmo del gasto por alumno, *lenroll* es el logaritmo del número de alumnos en la escuela, *lunch* es el porcentaje de estudiantes con desayuno gratuito o subsidiado; dé los resultados en la forma habitual. ¿Es cada una de las variables explicativas estadísticamente significativa al nivel de 5%?

ii) Obtenga los valores ajustados a partir de la regresión del inciso i). ¿Cuál es el rango de los valores ajustados? ¿Cómo es este rango en comparación con el rango de los datos reales en *math*4?

iii) Obtenga los residuales correspondientes a la regresión del inciso i). ¿Cuál es el código de la escuela (*bcode*) que tiene el residual (*positivo*) mayor? Interprete este residual.

iv) Agregue a la ecuación términos cuadráticos de todas las variables explicativas y pruebe su significancia conjunta. ¿Dejaría usted estos términos en el modelo?

v) Volviendo al modelo del inciso i), divida la variable independiente y cada una de las variables explicativas entre su desviación estándar muestral y vuelva a correr la regresión. (Incluya un intercepto a menos que primero también reste a cada variable su media.) En términos de unidades de desviaciones estándar, ¿cuál de las variables explicativas tiene el mayor efecto sobre la tasa de aprobación en matemáticas?

Apéndice 6A

6A. Breve introducción al bootstrapping

En muchos casos en los que se dificulta obtener matemáticamente las fórmulas para el error estándar, o cuando se cree que éstas no son aproximaciones muy buenas a la verdadera variación de muestreo de un estimador, puede uno apoyarse en un **método de remuestreo**. La idea general es tratar los datos observados como una población de donde se pueden sacar muestras. El método de remuestreo más común es el **bootstrap**. (En realidad existen varias versiones de bootstrap, pero el más general y de más fácil aplicación, al que se le llama *bootstrap no paramétrico* es el que se describe aquí.)

Suponga que se tiene una estimación $\hat{\theta}$ de un parámetro poblacional, θ. Esta estimación, que puede ser función de estimaciones de MCO (o de estimaciones que se verán en capítulos posteriores), fue obtenido de una muestra aleatoria de tamaño n. Se desea obtener un error estándar de $\hat{\theta}$ que pueda emplearse para calcular estadísticos t o intervalos de confianza. Naturalmente, un error estándar válido puede obtenerse calculando la estimación a partir de varias muestras aleatorias obtenidas de los datos originales.

La implementación es sencilla. Si se numeran las observaciones de la 1 a la n, y se obtienen de forma aleatoria n de estos números, con reposición. Esto produce un nuevo conjunto de datos (de tamaño n) que consta de los datos originales, pero en el que muchas de las observaciones aparecen repetidas (salvo en el raro caso que se obtenga la base original). Cada vez que se toma una muestra aleatoria de los datos originales, puede estimarse θ empleando el mismo procedimiento que se empleó con los datos originales. Sea $\hat{\theta}^{(b)}$ la estimación obtenida por bootstrap de la muestra b. Ahora, repitiendo el remuestreo y la estimación m veces,