CAPÍTULO 7

Análisis de regresión múltiple con información cualitativa: variables binarias (o *dummy*)

En los capítulos anteriores, la variable dependiente y las variables independientes en los modelos de regresión múltiple han tenido un significado *cuantitativo*. Algunos ejemplos son salario por hora, años de educación, promedio general de calificaciones en la universidad, cantidad de contaminación atmosférica, nivel de ventas de una empresa y número de arrestos. En cada caso, la magnitud de la variable proporciona información útil. En el trabajo empírico es necesario incluir también factores *cualitativos* en los modelos de regresión. El género o la raza de una persona, la industria de una empresa (manufactura, minorista, etc.) y la región de Estados Unidos (norte, sur, este, etc.) en la que se encuentra una ciudad son considerados factores cualitativos.

La mayor parte de este capítulo se dedica a las variables *independientes* cualitativas. Después de analizar, en la sección 7.1, la manera adecuada de describir variables cualitativas, en las secciones 7.2, 7.3 y 7.4 se describe cómo incorporar variables explicativas cualitativas a los modelos de regresión múltiple. En estas secciones se ven también casi todas las maneras más usuales de manejar variables independientes cualitativas en el análisis de regresión de corte transversal.

En la sección 7.5 se analiza una variable dependiente binaria, que es un tipo especial de variable dependiente cualitativa. En este caso, el modelo de regresión múltiple tiene una interpretación interesante y se le llama modelo de probabilidad lineal. Aunque ha sido muy difamado por muchos econometristas, la sencillez del modelo de probabilidad lineal lo hace útil en muchos contextos empíricos. Sus desventajas se describen en la sección 7.5, pero éstas suelen ser secundarias en el trabajo empírico.

7.1 Descripción de la información cualitativa

Los factores cualitativos surgen casi siempre en forma de información bivariada: una persona es mujer u hombre; una persona tiene o no computadora; una empresa ofrece o no un determinado tipo de plan de pensión a sus empleados; en un estado existe o no la pena de muerte. En todos estos ejemplos, la información que interesa puede ser captada empleando una **variable binaria** o una variable cero-uno. En econometría a las variables binarias se les suele llamar **variables binarias o *dummy***, aunque este nombre no es especialmente descriptivo.

Al definir una variable binaria hay que decidir a qué evento se le asigna el valor uno y a cuál el valor cero. Por ejemplo, en un estudio para determinar el salario de los individuos, puede definirse *female* como una variable binaria que tome el valor uno para mujer y el valor cero para hombre. En este caso, el nombre de la variable indica el evento que tiene valor uno. Esta misma información se capta definiendo *male* (hombre) igual a uno si la persona es hombre y cero si la persona es mujer. Cualquiera de éstas es mejor que emplear *gender* (género) porque este nombre

TABLA 7.1

Enumeración parcial de los datos del archivo WAGE1.RAW

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

no indica cuándo la variable binaria es uno: ¿*gender* = 1 corresponde a hombre o a mujer? Cómo se le llame a las variables no tiene importancia en la obtención de los resultados de la regresión, pero siempre ayuda elegir nombres que hagan más clara la ecuación y la exposición.

Pregunta 7.1

Suponga que, en un estudio para hacer una comparación entre los resultados de la elección para los demócratas y los republicanos, se desea indicar el partido de cada candidato. ¿Es un nombre como *party* (partido) una buena elección, en este caso, para una variable binaria? ¿Cuál sería un mejor nombre?

Suponga que en el ejemplo del salario se ha empleado *female* para indicar el género. Se define además una variable binaria *married* (casado) igual a uno si la persona está casada y cero si no es así. En la tabla 7.1 se muestra una enumeración parcial de los datos sobre salario que pueden obtenerse. Se ve que la persona 1 es mujer y que no está casada. La persona 2 es mujer y está casada. La persona 3 es hombre y no está casado y así sucesivamente.

¿Por qué se usan los valores cero y uno para describir información cualitativa? En cierto sentido, estos valores son arbitrarios: otros dos valores cualesquiera podrían servir igual. La verdadera ventaja de capturar la información cualitativa empleando variables cero-uno es que esto conduce a modelos de regresión en los que los parámetros tienen interpretaciones muy naturales, como se verá ahora.

7.2 Una sola variable binaria independiente

¿Cómo se incorpora la información binaria a los modelos de regresión? En el caso más sencillo en el que sólo hay una variable binaria explicativa, ésta simplemente se agrega a la ecuación como una variable independiente. Por ejemplo, considere el sencillo modelo siguiente para determinar el salario por hora:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \qquad \textbf{7.1}$$

Se emplea δ_0 como parámetro de *female* para resaltar la interpretación de los parámetros que multiplican a las variables binarias; más adelante se usará la notación que resulte más conveniente.

En el modelo 7.1, sólo hay dos factores que afectan al salario: el género y la educación. Como *female* $= 1$ si la persona es mujer y *female* $= 0$ si la persona es hombre, el parámetro δ_0 tiene la interpretación siguiente: δ_0 es la diferencia del salario por hora entre hombres y mujeres, *dada* una misma cantidad de educación (y un mismo término del error, u). De esta manera, el coeficiente δ_0 determina si hay discriminación en contra de las mujeres: si, para un mismo nivel de los demás factores, $\delta_0 < 0$, las mujeres ganan, en promedio, menos que los hombres.

En términos de expectativas, considerando el supuesto de media condicional cero $\mathrm{E}(u|female,educ) = 0$, entonces

$$\delta_0 = \mathrm{E}(wage|female = 1,educ) - \mathrm{E}(wage|female = 0,educ).$$

Como *female* $= 1$ corresponde a mujer y *female* $= 0$ corresponde a hombre, esto puede escribirse de manera más sencilla como

$$\delta_0 = \mathrm{E}(wage|female,educ) - \mathrm{E}(wage|male,educ). \qquad \textbf{7.2}$$

Lo importante aquí es que el nivel de educación es el mismo para las dos expectativas; la diferencia, δ_0, se debe sólo al género.

Esta situación puede representarse gráficamente como un **desplazamiento del intercepto** entre hombres y mujeres. En la figura 7.1 se muestra el caso $\delta_0 < 0$, de manera que, por hora, los hombres ganan más, en una cantidad fija, que las mujeres. Esta diferencia no depende de la cantidad de educación, y esto explica por qué las líneas de salario-educación de hombres y mujeres son paralelas.

En este punto el lector se preguntará por qué no se incluye también en (7.1) una variable binaria, por ejemplo, *male* (hombres), que sea uno para hombres y cero para mujeres. Esto sería redundante. En (7.1) el intercepto para hombres es β_0 y el intercepto para mujeres es $\beta_0 + \delta_0$. Dado que hay dos grupos, sólo se necesitan dos interceptos. Esto significa que además de β_0 sólo se necesita usar *una* variable binaria; se eligió incluir una variable binaria para mujeres. Usar dos variables binarias introduciría colinealidad perfecta, ya que *female* + *male* $= 1$, lo que significa que *male* es una función lineal perfecta de *female*. Incluir variables binarias para los dos géneros es el ejemplo más sencillo de lo que se conoce como **trampa de las variables binarias**, que surge cuando demasiadas variables binarias describen una determinada cantidad de grupos. Este problema se analizará más adelante.

En (7.1) se eligió hombres como **grupo base** o **grupo de referencia** (*benchmark*), es decir, el grupo contra el que se hacen las comparaciones. A esto se debe que β_0 sea el intercepto para hombres y δ_0 sea la *diferencia* entre los interceptos para hombres y para mujeres. También podría haberse elegido mujeres como grupo base, expresando el modelo como

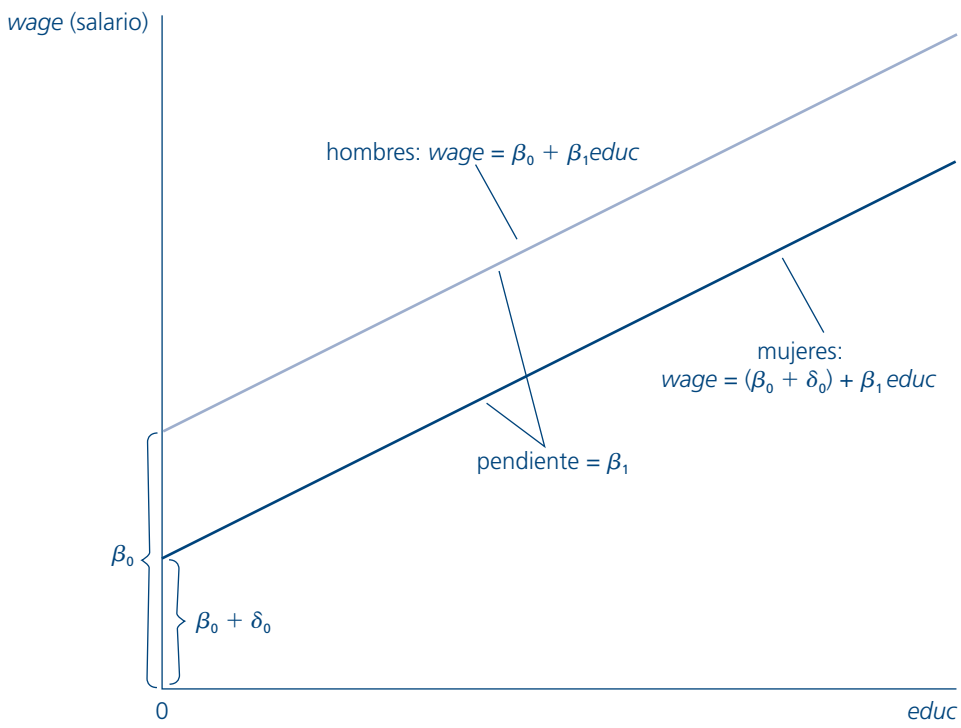
$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

donde el intercepto para mujeres es α_0 y el intercepto para hombres es $\alpha_0 + \gamma_0$; esto implica que $\alpha_0 = \beta_0 + \delta_0$ y $\alpha_0 + \gamma_0 = \beta_0$. En cualquier aplicación no tiene importancia qué grupo se elija como grupo base, pero no se debe olvidar qué grupo es el grupo base.

Algunos investigadores prefieren eliminar el intercepto general y emplear variables binarias para cada grupo. La ecuación será entonces $wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$, donde el intercepto para los hombres es β_0 y el intercepto para las mujeres es α_0. En este caso no hay trampa

FIGURA 7.1

Gráfica de $wage = \beta_0 + \delta_0 female + \beta_1 educ$ en la que $\delta_0 < 0$.



de las variables binarias, porque no se tiene un intercepto general. Sin embargo, esta formulación tiene poco que ofrecer, ya que probar la diferencia entre los interceptos es más complicado y, además, para regresiones sin intercepto, no existe un acuerdo general sobre cómo calcular la *R*-cuadrada. Por tanto, aquí siempre se incluirá un intercepto general para el grupo base.

Nada cambia mucho cuando intervienen más variables explicativas. Tomando hombres como grupo base, un modelo en el que, además de la educación se controle la experiencia y la antigüedad es

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u. \qquad \textbf{7.3}$$

Si tanto *educ* como *exper* y *tenure* son características importantes para la productividad, la hipótesis nula de que *no* hay diferencia entre hombres y mujeres es H_0: $\delta_0 = 0$. La alternativa de que existe discriminación contra las mujeres es H_1: $\delta_0 < 0$.

¿Cómo se puede probar que existe discriminación en los salarios? La respuesta es sencilla: simplemente se estima el modelo mediante MCO, *exactamente* como antes, y se usa el estadístico *t* habitual. Cuando algunas de las variables independientes se definen como variables binarias no cambia nada de la mecánica de MCO ni de la teoría estadística. La única diferencia encontrada hasta ahora es la interpretación del coeficiente de la variable binaria.

Ejemplo 7.1

[Ecuación para el salario por hora]

Empleando los datos del archivo WAGE1.RAW, se estimará el modelo (7.3). Por ahora, como variable dependiente se usará *wage* y no log(*wage*):

$$\widehat{wage} = -1.57 - 1.81\,female + .572\,educ \\ (.72) \quad (.26) \quad (.049) \\ + .025\,exper + .141\,tenure \\ (.012) \quad (.021) \\ n = 526,\ R^2 = .364. \tag{7.4}$$

El intercepto negativo —el intercepto para hombres, en este caso— no tiene mucho significado, porque en la muestra ninguna de las variables *educ*, *exper* o *tenure* (antigüedad) tiene valor cero. El coeficiente de *female* (mujer) es interesante porque mide la diferencia promedio entre el salario por hora de una mujer y de un hombre, dados los *mismos* niveles de *educ*, *exper* y *tenure*. Si se toman un hombre y una mujer con los mismos niveles de educación, experiencia y antigüedad, la mujer gana, en promedio $1.81 menos por hora que el hombre. (Recuerde que estos son salarios de 1976.)

Es importante recordar que, como se ha realizado una regresión múltiple en la que se han controlado *educ*, *exper* y *tenure*, la diferencia de $1.81 en el salario no puede ser explicada por diferencias en los niveles de educación, experiencia o antigüedad entre hombres y mujeres. Se puede concluir que tal diferencia se debe al género o a factores relacionados con el género que no han sido controlados en la regresión. [En dólares de 2003, la diferencia en el salario es aproximadamente $3.23(1.81) \approx 5.85$.]

Es interesante comparar el coeficiente de *female* en la ecuación (7.4) con la estimación que se obtiene cuando todas las demás variables explicativas se eliminan de la ecuación:

$$\widehat{wage} = 7.10 - 2.51\,female \\ (.21) \quad (.30) \\ n = 526,\ R^2 = .116. \tag{7.5}$$

Los coeficientes en (7.5) tienen una interpretación sencilla. El intercepto es el salario promedio de los hombres en la muestra (si $female = 0$), de manera que los hombres ganan, en promedio, $7.10 dólares por hora. El coeficiente de *female* es la diferencia entre el salario promedio de los hombres y el de las mujeres. Por tanto, en la muestra, el salario promedio de las mujeres es $7.10 - 2.51 = 4.59$, es decir, $4.59 por hora. (Dicho sea de paso, en la muestra hay 274 hombres y 252 mujeres.)

La ecuación (7.5) proporciona una manera sencilla de realizar una *prueba de comparación de las medias* entre dos grupos, que en este caso son hombres y mujeres. La diferencia estimada, -2.51, tiene un estadístico t de -8.37, que es estadísticamente muy significativo (y, por supuesto, que $2.51 también es económicamente una diferencia grande). En general, una regresión simple sobre una constante y una variable binaria es una manera sencilla de comparar las medias de dos grupos. Para que la prueba t habitual sea válida es necesario asumir que el supuesto de homocedasticidad se satisface, lo que significa que las varianzas poblacionales de los salarios de los hombres y de las mujeres son iguales.

La diferencia estimada entre los salarios de los hombres y de las mujeres es más grande en (7.5) que en (7.4) debido a que en (7.5) no se controlan las diferencias en educación, experiencia y antigüedad, y en promedio éstas tienen valores menores para las mujeres que para los hombres en la muestra. La ecuación (7.4) da una estimación más confiable de la brecha *ceteris paribus* entre los salarios según el género, e indica una diferencia aún muy grande.

En muchos casos, las variables binarias independientes reflejan elecciones de los individuos u otras unidades económicas (a diferencia de algo predeterminado como el género). En tales situaciones el asunto de la causalidad es de nuevo central. En el ejemplo siguiente se desea saber si tener una computadora personal es *causa* de un promedio de calificaciones superior en la universidad.

Ejemplo 7.2

[Efecto que tiene poseer una computadora sobre el promedio general de calificaciones (GPA) en la universidad]

Para determinar los efectos que poseer una computadora tiene sobre el promedio general de calificaciones se estima en el modelo

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

donde la variable binaria *PC* es uno si el estudiante posee una computadora personal y cero si no es así. Hay varias razones por las que poseer una computadora puede tener algún efecto sobre *colGPA*. Los trabajos de un estudiante pueden ser de mayor calidad si se realizan en una computadora y se puede ahorrar tiempo al no tener que esperar en un laboratorio de cómputo. Por supuesto, también un estudiante que posee una computadora puede que tienda más a jugar con los juegos de la computadora o a navegar por Internet, de manera que no es obvio que δ_0 sea positivo. Las variables *hsGPA* (promedio general de calificaciones en el bachillerato) y *ACT* (resultados en el examen de admisión) se emplean como controles: puede que sea más probable que los mejores estudiantes, de acuerdo con las puntuaciones en el *GPA* del bachillerato y en el *ACT*, tengan computadora. Estos factores se controlan porque se quiere conocer el efecto promedio sobre *colGPA* si se toma un estudiante al azar y se le da una computadora.

Empleando los datos en el archivo GPA1.RAW, se obtiene

$$\widehat{colGPA} = 1.26 + .157\,PC + .447\,hsGPA + .0087\,ACT \quad \textbf{7.6}$$
$$(.33)\ \ (.057)\qquad (.094)\qquad\ (.0105)$$
$$n = 141,\ R^2 = .219.$$

Esta ecuación implica que el GPA que se pronostica para un estudiante que posee una PC es aproximadamente .16 puntos más alto que el de un estudiante comparable que no tiene PC (recuerde que tanto *colGPA* como *hsGPA* se dan en una escala de cuatro puntos). Este efecto también es estadísticamente muy significativo, siendo $t_{PC} = .157/.057 \approx 2.75$.

¿Qué ocurre si de esta ecuación se eliminan *hsGPA* y *ACT*? Eliminar la última variable tendrá un efecto muy pequeño, ya que su coeficiente y su estadístico *t* son muy pequeños. Pero *hsGPA* es muy significativa y eliminarla puede afectar la estimación de β_{PC}. Regresando *colGPA* sobre *PC* se obtiene una estimación para *PC* igual aproximadamente a .170, con un error estándar de .063; en este caso, $\hat{\beta}_{PC}$ y su estadístico *t* no cambian mucho.

En los ejercicios al final del capítulo, se le pedirá que en la ecuación controle otros factores para ver si el efecto de la posesión de una computadora desaparece o si por lo menos se vuelve notablemente menor.

Cada uno de los ejemplos anteriores puede considerarse relevante en el **análisis de política**. En el primer ejemplo interesaba la discriminación de género en la fuerza de trabajo. En el segundo, el efecto de la posesión de una computadora sobre el desempeño en la universidad. Un caso especial del análisis de política es la **evaluación de programas**, en donde interesa conocer el efecto de programas económicos o sociales sobre las personas, las empresas, los vecindarios, las ciudades, etcétera.

En el caso más sencillo existen dos grupos de personas. El **grupo control** no participa en el programa; el **grupo experimental** o **grupo de tratamiento** sí. Estos nombres provienen de la literatura de las ciencias experimentales y no deben tomarse literalmente. Salvo en casos raros, tanto el grupo de control como el de tratamiento no son aleatorios. Sin embargo, en algunos

casos, el análisis de regresión múltiple puede emplearse para controlar otros factores con objeto de estimar el efecto causal del programa.

Ejemplo 7.3

[Efecto de apoyos de capacitación sobre las horas de ésta]

Empleando los datos del archivo JTRAIN.RAW sobre empresas manufactureras de Michigan en 1988, se obtuvo la ecuación estimada siguiente:

$$\widehat{hrsemp} = 46.67 + 26.25\ grant - .98\ \log(sales) \\ (43.41)\quad (5.59)\qquad (3.54) \\ - 6.07\ \log(employ) \qquad (7.7) \\ (3.88) \\ n = 105,\ R^2 = .237.$$

La variable dependiente es horas de capacitación por empleado, al nivel de la empresa. La variable *grant* es una variable binaria igual a uno si en 1988 la empresa recibió una subvención para capacitación e igual a cero si no fue así. Las variables *sales* y *employ* corresponden a ventas anuales y a cantidad de empleados, respectivamente. No es posible emplear *hrsemp* en forma logarítmica, debido a que para 29 de las 105 empresas usadas en la regresión *hrsemp* es cero.

La variable *grant* es estadísticamente muy significativa, siendo $t_{grant} = 4.70$. Controlando ventas (*sales*) y empleados (*employ*), las empresas que recibieron una subvención dieron a cada trabajador, una capacitación de 26.25 horas más, en promedio. Dado que la cantidad promedio de horas de capacitación por empleado en la muestra es aproximadamente 17, siendo el valor máximo 164, *grant* (subvención) tiene un efecto importante sobre la capacitación, como era de esperarse.

El coeficiente de log(*sales*) es pequeño y muy poco significativo. El coeficiente de log(*employ*) significa que, si una empresa es 10% mayor, capacitará a sus empleados .61 hora menos. Su estadístico *t* es −1.56, que es sólo marginal en términos estadísticos significativos.

Como ocurre con cualquier otra variable independiente, es necesario preguntarse si el efecto de una variable cualitativa es causal. En la ecuación (7.7), ¿la diferencia en capacitación entre las empresas que recibieron subvención y las que no se debe a la subvención, o la recepción de la subvención es simplemente un indicador de algo más? Puede ser que las empresas que recibieron la subvención, de cualquier manera, en promedio, hubieran capacitado más a sus empleados aun sin subvención. No hay nada en este análisis que indique que se ha estimado un efecto causal; es necesario saber cómo se eligieron las empresas para que recibieran la subvención. Sólo se puede esperar que se hayan controlado tantos factores como sea posible relacionados con que la empresa haya recibido una subvención y con su nivel de capacitación.

En la sección 7.6, así como en capítulos posteriores, se volverá a ver el análisis de políticas con variables binarias.

Interpretación de los coeficientes de variables explicativas binarias cuando la variable dependiente es log(*y*)

En una especificación usual en el trabajo práctico, la variable dependiente aparece en forma logarítmica y una o más variables binarias aparecen como variables independientes. ¿Cómo se interpretan, en este caso, los coeficientes de las variables binarias? No sorprenderá que los coeficientes tengan una interpretación *porcentual*.

Ejemplo 7.4

[Regresión para el precio de la vivienda]

Empleando los datos del archivo HPRICE1.RAW, se obtiene la ecuación

$$\widehat{\log(price)} = -1.35 + .168\log(lotsize) + .707\log(sqrft) + .027\,bdrms + .054\,colonial \qquad (7.8)$$
$$(.65)\quad (.038)\qquad\qquad (.093)\qquad\qquad (.029)\qquad\qquad (.045)$$
$$n = 88,\ R^2 = .649.$$

Todas las variables se explican por sí mismas, excepto *colonial*, que es una variable binaria igual a uno si la casa es de estilo colonial. ¿Qué significa el coeficiente de *colonial*? Para valores dados de *lotsize*, *sqrft* y *bdrms*, la diferencia en $\widehat{\log(price)}$ entre una casa de estilo colonial y una de otro estilo es .054. Esto significa que se predice que una casa de estilo colonial se venderá en aproximadamente 5.4% más, manteniendo constantes todos los demás factores.

Este ejemplo muestra que cuando en un modelo la variable dependiente es $\log(y)$ el coeficiente de una variable binaria, después de multiplicarlo por 100, se interpreta como la diferencia porcentual en y, manteniendo todos los demás factores constantes. Cuando el coeficiente de una variable binaria indica un cambio proporcional grande en y, la diferencia porcentual exacta puede obtenerse exactamente, como en el caso del cálculo de la semielasticidad en la sección 6.2.

Ejemplo 7.5

[Ecuación del logaritmo del salario por hora]

Se reestimará la ecuación para el salario, del ejemplo 7.1, empleando $\log(wage)$ como variable dependiente y agregando términos cuadráticos en *exper* y *tenure* (antigüedad):

$$\widehat{\log(wage)} = .417 - .297\,female + .080\,educ + .029\,exper - .00058\,exper^2 + .032\,tenure - .00059\,tenure^2 \qquad (7.9)$$
$$(.099)\quad (.036)\qquad (.007)\qquad (.005)\qquad\qquad (.00010)\qquad (.007)\qquad (.00023)$$
$$n = 526,\ R^2 = .441.$$

Empleando el mismo método que en el ejemplo 7.4, el coeficiente de *female* (mujer) implica que dados los mismos valores de *educ*, *exper* y *tenure*, las mujeres ganan aproximadamente 100(.297) = 29.7% menos que los hombres. Este resultado se puede mejorar calculando la diferencia porcentual exacta entre los salarios predichos. Lo que se quiere es la diferencia proporcional entre los salarios de las mujeres y de los hombres, manteniendo todos los demás factores constantes: $(\widehat{wage}_F - \widehat{wage}_M)/\widehat{wage}_M$. Lo que se tiene, de acuerdo con (7.9), es

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -.297.$$

Exponenciando y restando uno se obtiene

$$(\widehat{wage}_F - \widehat{wage}_M)/\widehat{wage}_M = \exp(-.297) - 1 \approx -.257.$$

Esta estimación más exacta implica que el salario de una mujer es, en promedio, 25.7% inferior al salario comparable de un hombre.

Si en el ejemplo 7.4 se hace la misma corrección se obtiene $\exp(.054) - 1 \approx .0555$, es decir, aproximadamente 5.6%. Esta corrección tiene un efecto menor en el ejemplo 7.4 que en el ejemplo del salario, debido a que la magnitud del coeficiente de la variable binaria es mucho menor en (7.8) que en (7.9).

En general, si $\hat{\beta}_1$ es el coeficiente de una variable binaria, por ejemplo x_1, siendo $\log(y)$ la variable dependiente, la diferencia porcentual exacta en la y predicha para $x_1 = 1$ versus $x_1 = 0$ es

$$100 \cdot [\exp(\hat{\beta}_1) - 1]. \qquad \textbf{7.10}$$

La estimación $\hat{\beta}_1$ puede ser positiva o negativa, y es importante preservar su signo al calcular (7.10).

El método logarítmico de aproximación tiene la ventaja de proporcionar una estimación entre las magnitudes obtenidas empleando cada grupo como grupo base. En particular, aunque la ecuación (7.10) da una estimación mejor que $100 \cdot \hat{\beta}_1$ del porcentaje en el que y para $x_1 = 1$ es mayor que y para $x_1 = 0$, (7.10) no es una buena estimación si se cambia el grupo base. En el ejemplo 7.5 se puede estimar el porcentaje en el que el salario de un hombre es superior al salario comparable de una mujer y esta estimación es $100 \cdot [\exp(-\hat{\beta}_1) - 1] = 100 \cdot [\exp(.297) - 1] \approx 34.6$. La aproximación, basada en $100 \cdot \hat{\beta}_1$, 29.7, se encuentra entre 25.7 y 34.6 (y cercano a la mitad). Por tanto, es razonable decir que "la diferencia que se predice entre los salarios de hombres y mujeres es aproximadamente 29.7%", sin tener que decir cuál es el grupo base.

7.3 Uso de variables binarias en categorías múltiples

En una misma ecuación pueden emplearse varias variables independientes binarias. Por ejemplo, a la ecuación (7.9) se le puede agregar la variable binaria *married* (casado). El coeficiente de *married* da la diferencia proporcional (aproximada) entre los salarios de los casados y de los solteros, manteniendo constantes género, *educ*, *exper* y *tenure* (antigüedad). Cuando se estima este modelo, el coeficiente de *married* (dando el error estándar entre paréntesis) es .053 (.041), y el coeficiente de *female* se convierte en −.290 (.036). Por tanto, se estima que la "prima de casado" es de aproximadamente 5.3%, pero no es estadísticamente distinta de cero ($t = 1.29$). Una limitación importante de este modelo es que se supone que la prima de casado es la misma para hombres que para mujeres; esto se soluciona en el ejemplo siguiente.

Ejemplo 7.6

[Ecuación para el logaritmo del salario por hora]

Ahora se estimará un modelo que toma en cuenta las diferencias entre cuatro grupos: hombres casados, mujeres casadas, hombres solteros y mujeres solteras. Para esto, se debe elegir un grupo base; se elige hombres solteros. Después se define una variable binaria para cada uno de los grupos restantes. Llámesele

a estas variables *marrmale*, *marrfem* y *singfem*. Introduciendo estas tres variables en (7.9) (y, por supuesto, eliminando *female*, ya que ahora es redundante) se obtiene

$$\begin{aligned}\widehat{\log(wage)} &= .321 + .213\,marrmale - .198\,marrfem \\ &\qquad (.100)\ \ (.055)\qquad\quad (.058) \\ &\quad - .110\,singfem + .079\,educ + .027\,exper - .00054\,exper^2 \\ &\qquad (.056)\qquad\ (.007)\qquad (.005)\qquad (.00011) \\ &\quad + .029\,tenure - .00053\,tenure^2 \\ &\qquad (.007)\qquad\ (.00023) \\ n &= 526,\ R^2 = .461.\end{aligned} \tag{7.11}$$

Todos los coeficientes, excepto *singfem*, tienen un estadístico *t* bastante mayor a dos, en valor absoluto. El estadístico *t* para *singfem* es aproximadamente −1.96, que apenas es significativo al nivel de 5% contra la alternativa de dos colas.

Para interpretar los coeficientes de las variables binarias, hay que recordar que el grupo base es hombres solteros. De esta manera, las estimaciones para las tres variables binarias miden la diferencia proporcional en el salario *con relación* a los hombres solteros. Por ejemplo, se estima que, manteniendo constantes los niveles de educación, experiencia y antigüedad, los hombres casados ganan aproximadamente 21.3% más que los solteros. [La estimación más precisa que se obtiene con (7.10) es aproximadamente 23.7%.] Una mujer casada, por otro lado, se predice que gana 19.8% menos que un hombre soltero siendo los niveles de otras variables los mismos.

Como en (7.11) el grupo base está representado por el intercepto, sólo se han incluido variables binarias para tres de los cuatro grupos. Si en (7.11) se introdujera una variable binaria para hombres casados se caería en la trampa de la variable binaria porque se introduciría colinealidad perfecta. Algunos paquetes para regresión corrigen, de manera automática, este error, mientras que otros simplemente indican que hay colinealidad perfecta. Lo mejor es especificar con cuidado las variables binarias, porque entonces se está forzando a interpretar de forma adecuada el modelo final.

Aunque en (7.11) el grupo base es hombres solteros, esta ecuación puede usarse para obtener la diferencia estimada entre cualesquiera dos de los grupos. Como el intercepto general es común a todos los grupos, ésta puede ignorarse al hallar las diferencias. Así, la diferencia proporcional estimada entre las mujeres solteras y casadas es $-.110 - (-.198) = .088$, lo que significa que las mujeres solteras ganan aproximadamente 8.8% más que las casadas. Por desgracia, la ecuación (7.11) no puede emplearse para probar si la diferencia estimada entre mujeres solteras y casadas es estadísticamente significativa. Conocer los errores estándar para *marrfem* y *singfem* no es suficiente para llevar a cabo esta prueba (vea la sección 4.4). Lo más fácil es elegir uno de estos grupos como grupo base y volver a estimar la ecuación. Con esto no cambia nada importante, pero la estimación buscada y su error estándar se obtienen de manera directa

$$\begin{aligned}\widehat{\log(wage)} &= .123 + .411\,marrmale + .198\,singmale + .088\,singfem + \ldots, \\ &\qquad (.106)\ \ (.056)\qquad\quad (.058)\qquad\quad (.052)\end{aligned}$$

donde, por supuesto, ninguno de los coeficientes o errores estándar que no se reportan han cambiado. La estimación de *singfem* es, como se esperaba, .088. Ahora, se tiene un error estándar que corresponde a esta estimación. El estadístico *t* para la hipótesis nula de que en la población no hay diferencia entre las mujeres casadas y solteras es $t_{singfem} = .088/.052 \approx 1.69$. Esta es una evidencia marginal contra la hipótesis nula. Se observa también que la diferencia estimada entre hombres casados y mujeres casadas es estadísticamente muy significativa ($t_{marrmale} = 7.34$).

El ejemplo anterior ilustra un principio general para la inclusión de variables binarias para indicar grupos diferentes: si el modelo de regresión ha de tener interceptos diferentes para, por ejemplo, g grupos o categorías, en el modelo se deberán incluir $g - 1$ variables binarias y un intercepto. El intercepto correspondiente al grupo base es el intercepto general del modelo, y el coeficiente de la variable binaria de un determinado grupo representa la diferencia estimada entre el intercepto de ese grupo y el grupo base. Incluir g variables binarias y un intercepto dará como resultado la trampa de la variable binaria. Una alternativa es incluir g variables binarias y eliminar el intercepto general. Algunas veces es útil incluir g variables binarias sin un intercepto general, pero tiene dos desventajas prácticas. Primero, hace más complicado probar diferencias en relación con un grupo base. Segundo, cuando no se incluye un intercepto general, los paquetes para regresión suelen modificar la manera en que calculan R-cuadrada. En particular, en la fórmula $R^2 = 1 - \text{SRC/STC}$, la suma total de cuadrados, STC, es sustituida por una suma total de cuadrados que no centra las y_i en torno a su media, por ejemplo, $\text{STC}_0 = \sum_{i=1}^{n} y_i^2$. A la R-cuadrada que se obtiene, por ejemplo $R_0^2 = 1 - \text{SRC/STC}_0$, se le suele llamar ***R*-cuadrada descentrada**. Por desgracia, R_0^2 pocas veces es adecuada como una medida de la bondad de ajuste. Siempre se tiene $\text{STC}_0 \geq \text{STC}$, presentándose la igualdad sólo cuando $\bar{y} = 0$. Con frecuencia STC_0 es mucho más grande que STC, lo que hace que R_0^2 sea mucho más grande que R^2. Por ejemplo, si en el ejemplo anterior se regresa log(*wage*) sobre *marrmale*, *singmale*, *marrfem*, *singfem* y las demás variables explicativas —sin intercepto— la R-cuadrada que se obtiene con Stata, que es R_0^2, es .948. Esta R-cuadrada elevada es un error por no centrar la suma total de cuadrados en los cálculos. En la ecuación (7.11) se da la R-cuadrada correcta, que es .461. Algunos paquetes para regresión, como Stata, tienen una opción para hacer que se calcule la R-cuadrada centrada aun cuando no se haya incluido un intercepto general, y casi siempre es aconsejable usar esta opción. En la inmensa mayoría de los casos, en cualquier R-cuadrada que se base en la comparación de una SRC y una STC, la STC deberá haber sido calculada centrando las y_i en torno a $\bar{y}$. Esta STC puede entenderse como la suma de los cuadrados residuales que se obtiene si se usa la media muestral, $\bar{y}$, para predecir todas las y_i. Desde luego, se puede esperar muy poco de un modelo en el que lo único que se mide es su ajuste en relación con el uso de una constante como predictor. En un modelo sin intercepto que tenga un mal ajuste, es posible que $\text{SRC} > \text{STC}$, lo que significa que R^2 será negativa. La R-cuadrada descentrada estará siempre entre cero y uno, lo que posiblemente explica por qué suele ser la que se emplea si no se indica otra cosa cuando en los modelos de regresión no se estima un intercepto.

Pregunta 7.2

En la base de datos del archivo MLB1.RAW, sobre los salarios en el béisbol, a los jugadores se les asigna una de seis posiciones: *frstbase* (primera base), *scndbase* (segunda base), *thrdbase* (tercera base), *shrtstop* (parador en corto), *outfield* (jardinero) o *catcher* (receptor). ¿Cuáles son las variables binarias que deben incluirse como variables independientes en el modelo para considerar las diferencias de salario entre estas posiciones?

Incorporación de información ordinal mediante el uso de variables binarias

Suponga que se desea estimar el efecto de la calificación crediticia de la ciudad sobre las tasas de interés de los bonos municipales (*MBR*). Varias compañías financieras, por ejemplo Moody's Investors Service and Standard Poor's, califican la calidad de la deuda de los gobiernos locales, dependiendo de la calificación de cosas tales como la probabilidad de incumplimiento. (Los gobiernos locales prefieren tasas de interés bajas con objeto de reducir los costos de sus préstamos.) Para simplificar, suponga que las calificaciones van de cero a cuatro, siendo cero la peor calificación crediticia y cuatro la mejor. Este es un ejemplo de una **variable ordinal**. Llámesele a esta variable *CR*. La pregunta que hay que responder es: ¿Cómo se incorpora *CR* en un modelo que explique *MBR*?

Una posibilidad es incluir *CR* como se incorporaría cualquier otra variable explicativa:

$$MBR = \beta_0 + \beta_1 CR + otros\ factores,$$

donde no se muestra explícitamente qué otros factores están en el modelo. Entonces β_1 es la variación de *MBR*, en puntos porcentuales, cuando *CR* aumenta una unidad, permaneciendo todos los demás factores constantes. Por desgracia, es bastante difícil interpretar un aumento de *CR* de una unidad. Se conoce el significado cuantitativo de un año más de educación o de un dólar más que se gasta por estudiante, pero cosas como las calificaciones crediticias suelen tener sólo un significado ordinal. Se sabe que una *CR* de cuatro es mejor que una *CR* de tres, pero ¿es la diferencia entre cuatro y tres igual a la diferencia entre uno y cero? Si no es así, entonces no tiene sentido suponer que un aumento de una unidad en *CR* tiene un efecto constante sobre *MBR*.

Una mejor idea, que se puede emplear dado que *CR* toma relativamente pocos valores, es definir variables binarias para cada valor de *CR*. Así, sea $CR_1 = 1$ si $CR = 1$ y $CR_1 = 0$ si no es así; $CR_2 = 1$ si $CR = 2$ y $CR_2 = 0$ si no es así; etc. En efecto, se toma la calificación crediticia y se convierte en cinco categorías. Así se puede estimar el modelo

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otros\ factores. \qquad \textbf{7.12}$$

Siguiendo la regla para la inclusión de variables binarias en el modelo, se incluyen cuatro de estas variables, porque se tienen cinco categorías. La categoría omitida aquí es la calificación crediticia de cero, que de esta manera es el grupo base. (Esta es la razón por la que no se tiene que definir una variable binaria para esta categoría.) Los coeficientes tienen una interpretación sencilla: δ_1 es la diferencia en *MBR* (permaneciendo los demás factores constantes) entre una municipalidad con una calificación crediticia de uno y una con una calificación crediticia de cero; δ_2 es la diferencia en *MBR* entre una municipalidad con una calificación crediticia de dos y una con una calificación crediticia de cero, etc. Se ha permitido que el movimiento entre cada dos calificaciones crediticias tenga efectos diferentes, con lo que usar (7.12) es mucho más flexible que incluir *CR* como una sola variable. Una vez que se han definido estas variables binarias, es sencillo estimar (7.12).

Pregunta 7.3

En el modelo (7.12), ¿cómo probaría la hipótesis nula de que la calificación crediticia no tiene efecto en *MBR*?

La ecuación (7.12) contiene, como caso especial, el modelo en el que el efecto parcial es constante. Una manera de expresar las tres restricciones para que impliquen un efecto parcial constante es $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$ y $\delta_4 = 4\delta_1$. Sustituyendo en la ecuación (7.12) y reordenando, se obtiene $MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + otros\ factores$. Ahora, el término que multiplica a δ_1 es simplemente la variable original para la calificación del crédito, *CR*. Para obtener el estadístico *F* para probar las restricciones de efecto parcial constante, se obtiene la *R*-cuadrada no restringida de (7.12) y la *R*-cuadrada restringida de la regresión de *MBR* sobre *CR* y los demás factores que se han controlado. El estadístico *F* se obtiene como en la ecuación (4.41) con $q = 3$.

Ejemplo 7.7

[Efectos del atractivo físico sobre el salario]

Hamermesh y Biddle (1944) emplearon mediciones del atractivo físico en una ecuación de salario. (El archivo BEAUTY.RAW contiene menos variables, pero más observaciones de las usadas por Hamermesh

y Biddle.) Un entrevistador dio a cada una de las personas de la muestra una calificación de acuerdo con su atractivo físico, empleando cinco categorías (feo, poco atractivo, regular, bien parecido, hermoso o guapo). Dado que en los dos extremos hay muy pocas personas, para el análisis de regresión los autores colocaron a las personas en tres grupos: promedio, superior al promedio (*abvavg*) e inferior al promedio (*belavg*), donde el grupo base es promedio (*average*). Empleando los datos de la encuesta Quality of Employment Survey de 1977, después de controlar las características de productividad habituales, Hamermesh y Biddle estimaron una ecuación para hombres:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.046)}{.164}\, belavg + \underset{(.033)}{.016}\, abvavg + otros\ factores$$
$$n = 700, \bar{R}^2 = .403$$

y una para mujeres:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.066)}{.124}\, belavg + \underset{(.049)}{.035}\, abvavg + otros\ factores$$
$$n = 409, \bar{R}^2 = .330.$$

Los otros factores controlados en la regresión son educación, experiencia, antigüedad, estado civil y raza; para una lista más completa vea la tabla 3 en el artículo de Hamermesh y Biddle. Para ahorrar espacio, en ese artículo no se dan los coeficientes de las otras variables ni tampoco el intercepto.

En el caso de los hombres, se estima que aquellos cuya apariencia es inferior al promedio ganan aproximadamente 16.4% menos que un hombre con una apariencia promedio e igual en los demás aspectos (educación, experiencia, antigüedad, estado civil y raza). El efecto es estadísticamente distinto de cero con $t = -3.57$. De manera similar, los hombres con un aspecto superior al promedio ganan un estimado de 1.6% más, aunque este efecto no es estadísticamente significativo ($t < .5$).

Una mujer con una apariencia inferior al promedio gana aproximadamente 12.4% menos que una mujer con una apariencia promedio y por lo demás comparable a ella, con $t = -1.88$. Como ocurrió con los hombres, la estimación para *abvavg* no es estadísticamente distinta de cero.

Hay casos en los que una variable ordinal toma demasiados valores, por lo que no puede incluirse una variable binaria para cada valor. Por ejemplo, el archivo LAWSCH85.RAW contiene datos sobre los salarios iniciales medianos de egresados de escuelas de leyes. Una de las variables explicativas clave es la calificación dada a la escuela de leyes. Como cada escuela tiene una calificación distinta, es claro que no puede incluirse una variable binaria para cada calificación. Si no se quiere incluir directamente la calificación en la ecuación, las calificaciones pueden dividirse en categorías. En el ejemplo siguiente se muestra cómo se hace esto.

Ejemplo 7.8

[Efectos del ranking de una escuela de leyes sobre el salario inicial]

Se definen las variables binarias *top10*, *r11_25*, *r26_40*, *r41_60*, *r61_100* que toman el valor uno cuando la variable *rank* (ranking) cae dentro del rango correspondiente. Como grupo base se consideran las escuelas con un ranking que no esté entre las 100 mejores. La ecuación estimada es

$$\widehat{\log(salary)} = 9.17 + .700\, top10 + .594\, r11\_25 + .375\, r26\_40 + .263\, r41\_60 + .132\, r61\_100 + .0057\, LSAT + .014\, GPA + .036 \log(libvol) + .0008 \log(cost)$$
$$(.41)\quad (.053)\qquad (.039)\qquad (.034)\qquad (.028)\qquad (.021)\qquad (.0031)\qquad (.074)\qquad (.026)\qquad (.0251)$$
$$n = 136,\ R^2 = .911,\ \bar{R}^2 = .905. \qquad \textbf{7.13}$$

Se ve inmediatamente que todas las variables binarias que definen los diferentes rangos son estadísticamente muy significativas. La estimación para *r61_100* significa que, manteniendo constantes *LSAT*, *GPA*, *libvol* y *cost*, el salario medio en una escuela de leyes clasificada entre 61 y 100 es aproximadamente 13.2% superior que en una escuela de leyes que no esté clasificada entre las 100 mejores. La diferencia entre una escuela de las 10 mejores y una escuela no clasificada entre las 100 mejores es bastante grande. Empleando el cálculo exacto dado en la ecuación (7.10) se obtiene exp(.700) − 1 ≈ 1.014, con lo que el salario medio que se predice es más de 100% superior en las 100 mejores escuelas que en una de las escuelas no clasificada entre las 100 mejores.

Como indicación de si el dividir el ranking en categorías representa una mejora, se puede comparar la *R*-cuadrada ajustada de (7.13) con la *R*-cuadrada ajustada cuando se incluye *rank* (ranking) como una sola variable: la primera es .905 y la última es .836, de manera que está justificada la mayor flexibilidad de (7.13).

Es interesante observar que una vez que las posiciones del ranking se colocan en las (algo arbitrarias) categorías dadas, todas las demás variables se vuelven poco significativas. Una prueba de significancia conjunta para *LSAT*, *GPA*, log(*libvol*) y log(*cost*) da un valor-*p* de .055, el cual es apenas significativo. Cuando se incluye *rank* en su forma original, el valor-*p* para la significancia conjunta es cero a cuatro posiciones decimales.

Un último comentario acerca de este ejemplo. Al obtener las propiedades de mínimos cuadrados ordinarios, se asumió que se tenía una muestra aleatoria. En esta aplicación se viola ese supuesto debido a la forma en que se define *rank*: la posición en el ranking de una escuela depende necesariamente de las posiciones de las demás escuelas de la muestra y siendo así, los datos no pueden representar muestreos independientes de la población de todas las escuelas de leyes. Esto no ocasiona ningún problema serio, siempre y cuando el término de error no esté correlacionado con las variables explicativas.

7.4 Interacciones en las que intervienen variables binarias

Interacciones entre variables binarias

Así como variables con un significado cuantitativo pueden estar relacionadas en los modelos de regresión, también pueden estarlo las variables binarias. En realidad ya vimos esto en el ejemplo 7.6, en donde se definieron cuatro categorías con base en el estado civil y en el género. En efecto, este modelo puede retomarse agregando un **término de interacción** entre *female* (mujer) y *married* (casado) y donde *female* y *married* aparezcan por separado. Esto permite que la prima de casado dependa del género, como lo hace en la ecuación (7.11). Para propósitos de comparación, el modelo estimado con el término de interacción *female-married* es

$$\widehat{\log(wage)} = .321 - .110\, female + .213\, married - .301\, female{\cdot}married + \ldots, \qquad \textbf{7.14}$$
$$(.100)\quad (.056)\qquad (.055)\qquad (.072)$$