sueldo que es explicada por el rendimiento sobre el capital no depende de si el sueldo se mide en dólares o en miles de dólares o de si el rendimiento sobre el capital es un porcentaje o un decimal. Esta idea intuitiva puede verificarse matemáticamente: empleando la definición de R^2 se puede mostrar que, en efecto R^2 no varía ante los cambios de unidades de y o de x.

Incorporación de no linealidades en la regresión simple

Hasta ahora, se ha fijado la atención en relaciones *lineales* entre las variables dependiente e independiente. Como se dijo en el capítulo 1, las relaciones lineales no son suficientemente generales para todas las aplicaciones económicas. Por fortuna, es bastante fácil incorporar muchas no linealidades en el análisis de regresión simple mediante una definición apropiada de las variables dependiente e independiente. Aquí se verán dos posibilidades que surgen con frecuencia en la práctica.

En la literatura de las ciencias sociales, con frecuencia se encuentran ecuaciones de regresión en las que la variable dependiente aparece en forma logarítmica. ¿A qué se debe esto? Recuerde el ejemplo sueldo-educación, en el que se hizo la regresión del salario por hora sobre años de educación. La pendiente estimada fue de 0.54 [vea la ecuación (2.27)], lo que significa que se predice que por cada año más de educación el salario por hora aumentará 54 centavos. Debido a que la ecuación (2.27) es lineal, 54 centavos es el aumento, ya sea por el primer o por el vigésimo año de educación; cosa que no parece razonable.

Una mejor caracterización para el cambio del salario de acuerdo con la educación puede que sea que por cada año más de educación el salario aumente un *porcentaje* constante. Por ejemplo, que un aumento en la educación de cinco a seis años haga que el salario aumente, por ejemplo, 8% (*ceteris paribus*), y que un aumento en la educación de 11 a 12 años, haga que el salario aumente también 8%. Un modelo con el que (aproximadamente) se obtiene un efecto porcentual constante es

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{2.42}$$

donde $\log(\cdot)$ denota el logaritmo *natural*. (Vea en el apéndice A un repaso de los logaritmos.) En particular, si $\Delta u = 0$, entonces

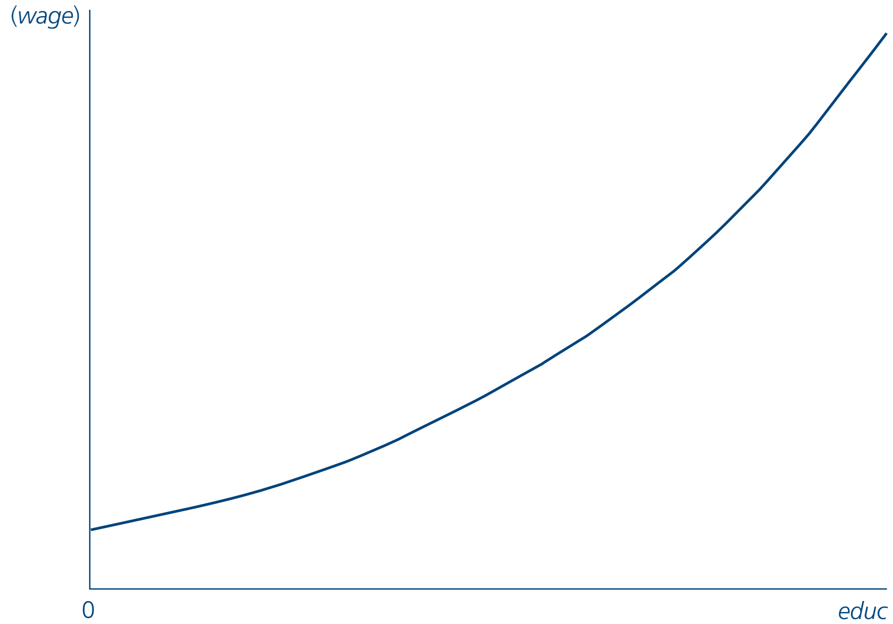
$$\%\Delta wage \approx (100 \cdot \beta_1)\Delta educ. \tag{2.43}$$

Observe que β_1 se multiplica por 100 para obtener el cambio porcentual de *wage* por un año más de educación. Como el cambio porcentual es el mismo por cada año adicional de educación, el cambio de *wage* por un año más de educación *aumenta* a medida que la educación lo hace; en otras palabras, (2.42) implica un rendimiento *creciente* de la educación. Exponenciando (2.42), se obtiene $wage = \exp(\beta_0 + \beta_1 educ + u)$. En la figura 2.6 se grafica esta ecuación, con $u = 0$.

La estimación de un modelo como el de la ecuación (2.42) es sencilla cuando se usa regresión simple. Sólo se define la variable dependiente, y, como $y = \log(wage)$. La variable independiente se representa por $x = educ$. La mecánica de MCO sigue siendo la misma que antes: las estimaciones del intercepto y de la pendiente están dadas por las fórmulas (2.17) y (2.19). En otras palabras, $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen mediante una regresión por MCO de $\log(wage)$ sobre *educ*.

FIGURA 2.6

***wage* = exp(β_0 + β_1*educ*), con β_1 > 0.**



Ejemplo 2.10

[Ecuación logarítmica del salario]

Empleando los mismos datos que en el ejemplo 2.4, pero usando log(*wage*) como variable dependiente, se obtiene la siguiente relación:

$$\widehat{\log(wage)} = 0.584 + 0.083\ educ \qquad \textbf{2.44}$$

$$n = 526, R^2 = 0.186.$$

El coeficiente de *educ* tiene una interpretación porcentual multiplicándolo por 100: $\widehat{wage}$ aumenta 8.3% por cada año más de educación. Esto es a lo que los economistas se refieren cuando hablan de "rendimiento de un año más de educación".

Es importante recordar que la principal razón para emplear el logaritmo de *wage* en la ecuación (2.42) es imponer a la educación un efecto porcentual constante sobre *wage*. Una vez obtenida la ecuación (2.42), el logaritmo natural de *wage* apenas se menciona. En particular, *no* es correcto decir que un año más de educación incrementa log(*wage*) 8.3 por ciento.

En la ecuación (2.42) el intercepto no tiene mucho significado, ya que da el log(*wage*) predicho, cuando *educ* = 0. La *R*-cuadrada muestra que *educ* explica cerca de 18.6% de la variación en log(*wage*) (*no* en *wage*). Por último, la ecuación (2.44) puede que no capte por completo la no linealidad de la relación entre salarios y escolaridad. Si hay "efectos de diploma", entonces el duodécimo año de educación —terminación del bachillerato— deberá ser mucho más valioso que el undécimo año. En el capítulo 7 se verá cómo considerar esta clase de no linealidades.

Otro uso importante del logaritmo natural es la obtención de un **modelo de elasticidad constante**.

Ejemplo 2.11

[Sueldo de los CEO y ventas de la empresa]

Se va a estimar un modelo de elasticidad constante en el que relacione el sueldo de los CEO con las ventas de la empresa. El conjunto de datos es el mismo que se usó en el ejemplo 2.3, salvo que ahora se relaciona *sueldo* con *ventas*. Sea *sales* las ventas anuales de la empresa medidas en millones de dólares. Un modelo de elasticidad constante es

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \tag{2.45}$$

donde β_1 es la elasticidad de *salary* respecto a *sales*. Este modelo cae dentro de los modelos de regresión simple mediante la definición de la variable dependiente como $y = \log(salary)$ y de la variable independiente como $x = \log(sales)$. Estimando esta ecuación mediante MCO se obtiene

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \tag{2.46}$$

$$n = 209, R^2 = 0.211.$$

El coeficiente de $\log(sales)$ es la elasticidad estimada de *salary* (sueldo) respecto a *sales* (ventas). Esto implica que por cada aumento de 1% en las ventas de la empresa hay un aumento de aproximadamente 0.257% en el sueldo de los CEO —la interpretación usual de una elasticidad.

Las dos formas funcionales vistas en esta sección aparecerán con frecuencia en el resto de este libro. Estos modelos con logaritmos naturales se han visto aquí debido a que se encuentran con frecuencia en la práctica. En el caso de la regresión múltiple la interpretación de estos modelos no será muy diferente.

También es útil observar lo que ocurre con las estimaciones del intercepto y la pendiente cuando las unidades de medición de la variable dependiente cambian y esta variable aparece en forma logarítmica. Dado que el cambio a dicha forma aproxima un cambio proporcional, es razonable que no suceda *nada* con la pendiente. Esto se puede ver escribiendo cada observación i de la variable reescalada como $c_1 y_i$. La ecuación original es $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$. Si se agrega $\log(c_1)$ a ambos lados, se obtiene $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$, es decir $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. ((Recuerde que la suma de logaritmos es igual al logaritmo de sus productos, como se muestra en el apéndice A.) Por tanto, la pendiente sigue siendo β_1, pero el intercepto es ahora $\log(c_1) + \beta_0$. De manera similar, si la variable independiente es $\log(x)$ y se modifican las unidades de x antes de obtener el logaritmo, la pendiente sigue siendo la misma, pero el intercepto cambia. En el problema 2.9 se pide verificar esto.

Esta subsección se termina resumiendo las cuatro formas funcionales que se obtienen empleando ya sea la variable original o su logaritmo natural. En la tabla 2.3, x y y representan las variables en su forma original. Al modelo en el que y es la variable dependiente y x es la variable independiente se le llama modelo *nivel-nivel* debido a que las variables aparecen en sus unidades de medición original. Al modelo en el que $\log(y)$ es la variable dependiente y x la variable independiente se le llama modelo *log-nivel*. El modelo *log-nivel* no será analizado aquí explícitamente debido a que se encuentra con menos frecuencia en la práctica. No obstante, en capítulos posteriores se verán ejemplos de este modelo.

TABLA 2.3

Resumen de las formas funcionales en las que se emplean logaritmos

Modelo	Variable dependiente	Variable independiente	Interpretación de β_1
Nivel-nivel	y	x	$\Delta y = \beta_1 \Delta x$
Nivel-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-nivel	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

En la última columna de la tabla 2.3 se da la interpretación de β_1. En el modelo log-nivel, a $100 \cdot \beta_1$ se le suele conocer como la **semielasticidad** de y respecto a x. Como se dijo en el ejemplo 2.11, en el modelo log-log, β_1 es la **elasticidad** de y respecto a x. La tabla 2.3 merece cuidadoso estudio, ya que con frecuencia se hará referencia a ella en el resto del libro.

Significado de regresión "lineal"

Al modelo de regresión simple que se ha estudiado en este capítulo también se le llama modelo de regresión *lineal* simple. Sin embargo, como se acaba de ver, el modelo general también permite ciertas relaciones *no lineales*. Entonces, ¿qué significa aquí "lineal"? Observando la ecuación (2.1) se puede ver que $y = \beta_0 + \beta_1 x + u$. La clave es que esta ecuación es lineal en los *parámetros* β_0 y β_1. No hay restricción alguna en la manera en que y y x estén relacionadas con las variables originales, explicada y explicativa, de interés. Como se vio en los ejemplos 2.10 y 2.11, y y x pueden ser los logaritmos naturales de estas variables, lo que es muy común en las aplicaciones. Pero esto no es todo. Por ejemplo, nada impide que se use la regresión simple para estimar un modelo como el siguiente: $cons = \beta_0 + \beta_1\sqrt{inc} + u$, donde $cons$ es el consumo anual e inc es el ingreso anual.

Mientras que la mecánica de la regresión simple no depende de la manera en que estén definidas y y x, la interpretación de los coeficientes sí depende de sus definiciones. Para realizar un trabajo empírico exitoso, es mucho más importante tener la capacidad de interpretar los coeficientes que tener destreza para calcular fórmulas como la (2.19). Cuando se estudie la regresión múltiple se logrará mucho más práctica en la interpretación de las estimaciones de la línea de regresión de MCO.

Hay cantidad de modelos que *no pueden* ser enmarcados en el modelo de regresión lineal ya que no son lineales en sus parámetros; un ejemplo es: $cons = 1/(\beta_0 + \beta_1 inc) + u$. La estimación de estos modelos nos lleva al campo de los *modelos de regresión no lineal*, que queda fuera del alcance de este libro. Para la mayoría de las aplicaciones, es suficiente elegir un modelo que se pueda situar dentro del marco de la regresión lineal.

2.5 Valores esperados y varianzas de los estimadores de MCO

En la sección 2.1 se definió el modelo poblacional $y = \beta_0 + \beta_1 x + u$, y se dijo que el principal supuesto para que el análisis de regresión simple sea útil es que el valor esperado de u dado