Wooldridge - Análisis de regresión con datos de corte transversal (Selección)[1]

A. El modelo de regresión simple

El modelo de regresión simple puede utilizarse para estudiar la relación entre dos variables. Por razones que se verán más adelante, este modelo presenta limitaciones como herramienta general para el análisis empírico. No obstante, algunas veces es adecuado como herramienta empírica. Aprender a interpretar el modelo de regresión simple es una buena práctica para estudiar la regresión múltiple, lo cual se hará en capítulos subsiguientes.

2.1 Definición del modelo de regresión simple

Gran parte de los análisis en econometría aplicada parten de la premisa siguiente: y y x son dos variables que representan alguna población y se desea "explicar y en términos de x" o "estudiar cómo varía y cuando varía x". En el capítulo 1 se vieron algunos ejemplos: y es rendimiento de cultivos de frijol de soya y x la cantidad de fertilizante; y es salario por hora y x los años de educación escolar e y es la tasa de delincuencia en una comunidad y x la cantidad de policías.

Para establecer un modelo que "explique y en términos de x" hay que tomar en consideración tres aspectos. Primero, dado que entre las variables nunca existe una relación exacta, ¿cómo pueden tenerse en cuenta otros factores que afecten a y? Segundo, ¿cuál es la relación funcional entre y y x? Y, tercero, ¿cómo se puede estar seguro de que la relación entre y y x sea una relación *cæteris paribus* entre y y x (si es ése el objetivo buscado)?

Estas ambigüedades pueden resolverse estableciendo una ecuación que relacione y con x. Una ecuación sencilla es

$$[2.1] \qquad y = \beta_0 + \beta_1 x + u$$

La ecuación (2.1), que se supone válida en la población de interés, define el modelo de regresión lineal simple. A esta ecuación también se le llama modelo de regresión lineal de dos variables o modelo de regresión lineal bivariada debido a que en este modelo se relacionan las dos variables x y y. A continuación se analizará el significado de cada una de las cantidades que aparecen en la ecuación (2.1). [Dicho sea de paso, el origen del término "regresión" no tiene una importancia especial para la mayoría de las aplicaciones econométricas modernas, por lo que no se explicará aquí. Ver la historia del análisis de regresión en Stigler[2].]

---

[1] Tomado en forma parcial de Jeffrey M. Wooldridge - *Introducción a la econometría. Un enfoque moderno*. 4a. edición, 2009, Capítulo 2. Se ha simplificado la exposición, y se han dejado de lado los tecnicismos, pero se mantuvo la numeración de secciones, tablas y ecuaciones y gráficos. También hemos salteado otras cuestiones que ya están incluidas en otras lecturas.

[2] Stephen M. Stigler, *The History of Statistics - The Measurement of Uncertainty before 1900*, 1990. Pueden leer en internet el capítulo 1 (*Least Squares and the combination of observations*).

Cuando las variables y y x se relacionan mediante la ecuación (2.1) se les da diversos nombres que se usan indistintamente: a y se le conoce como la variable dependiente, la variable explicada, la variable de respuesta, la variable predicha o el regresando; a x se le conoce como la variable independiente, la variable explicativa, la variable de control, la variable predictora o el regresor. (Para x también se usa el término covariada.) En econometría con frecuencia se usan los términos "variable dependiente" y "variable independiente". Pero hay que hacer notar que aquí el término "independiente" no se refiere al concepto estadístico de independencia entre variables aleatorias (vean el apéndice B).

Los términos variables "explicada" y "explicativa" son probablemente los más descriptivos. "Respuesta" y "control" se usan más en las ciencias experimentales, en donde la variable x está bajo el control del experimentador. Aquí no se usarán los términos "variable predicha" y "predictor", aunque éstos a veces se encuentran en aplicaciones relacionadas sólo con la predicción y no con la causalidad. La terminología que se empleará aquí para la regresión simple se resume en la tabla 2.1.

Terminología en la regresión simple

y	x
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictora
Regresando	Regresor

Tabla 2.1

La variable u, llamada *término de error*, o *perturbación* en la relación, representa factores distintos a x que afectan a y. Un análisis de regresión simple en realidad trata a todos los factores que afectan a y, y que son distintos a x como factores no observados. Es útil considerar a u como abreviación de *unobserved* (no observado, en inglés).

La ecuación (2.1) también resuelve el problema de la relación funcional entre y y x. Si los demás factores en u permanecen constantes, de manera que el cambio en u sea cero, $\Delta u = 0$, entonces x tiene un efecto *lineal* sobre y:

[2.2] $$\Delta y = \beta_1 \Delta x \text{ si } \Delta u = 0.$$

Por tanto, el cambio en y es simplemente β_1 multiplicado por el cambio en x. Esto significa que β_1 es el parámetro de la *pendiente* en la relación entre y y x, cuando todos los demás factores en u permanecen constantes; este parámetro es de interés primordial en la economía aplicada. El parámetro del *intercepto* β_0, algunas veces llamado término constante, tiene también su utilidad, aunque es raro que tenga una importancia central en el análisis.

Ejemplo 2.1 Rendimiento del frijol de soya y el fertilizante

Supongan que el rendimiento del frijol de soya está determinado por el modelo

[2.3] $rendimiento = \beta_0 + \beta_1 fertilizante + u,$

de manera que $y = rendimiento$ y $x = fertilizante$. Al investigador agrícola le interesa el efecto del fertilizante sobre el rendimiento, cuando todos los demás factores permanecen constantes. Este efecto está dado por β_1. El término del error u comprende factores como calidad de la tierra, precipitación pluvial, etc. El coeficiente β_1 mide el efecto del fertilizante sobre el rendimiento, cuando todos los demás factores permanecen constantes: $\Delta rendimiento = \beta_1 \Delta fertilizante$.

Ejemplo 2.2 Una ecuación sencilla para el salario

Un modelo en el que se relaciona el salario de una persona con la educación observada y con otros factores no observados es

[2.4] $salario = \beta_0 + \beta_1 educ + u.$

Si *salario* se mide en dólares por hora y *educ* se mide en años de educación, entonces β_1 mide la variación en el salario por hora por cada año adicional de educación, cuando todos los demás factores permanecen constantes. Entre estos factores se encuentran experiencia laboral, capacidades innatas, antigüedad en el empleo actual, ética laboral y otra gran cantidad de cosas.

La linealidad de la ecuación (2.1) implica que todo cambio de x en una unidad tiene siempre el mismo efecto sobre y, sin importar el valor inicial de x. En muchas aplicaciones de la economía esto no es muy realista. *Así, en el ejemplo del salario y la educación, es deseable permitir que haya rendimientos crecientes*: un año más en educación escolar debe tener un efecto mayor que el que tuvo el año anterior. En la sección 2.4 se verá cómo tener estas posibilidades.

El problema más difícil de abordar es si el modelo dado por la ecuación (2.1) en realidad permite formular conclusiones *cæteris paribus* acerca de cómo afecta x a y. Se acaba de ver que en la ecuación (2.2) β_1 mide el efecto de x sobre y, cuando todos los demás factores (en u) permanecen constantes. ¿Resuelve esto el problema de la causalidad? Por desgracia, no. ¿Cómo esperar conocer el efecto *cæteris paribus* de x sobre y, cuando todos los demás factores permanecen constantes, si se ignoran esos otros factores?

En la sección 2.5 se muestra que la única manera de obtener estimadores confiables de β_0 y β_1 a partir de los datos de una muestra aleatoria, es haciendo una suposición que restrinja la manera en que la variable no observable u está relacionada con la variable explicativa x. Sin esta restricción, no es posible estimar el efecto *cæteris paribus*, β_1. Como u y x son variables aleatorias, se necesita un concepto basado en la probabilidad.

Antes de establecer esta suposición clave acerca de la relación entre x y u se puede hacer una suposición acerca de u. En tanto el intercepto β_0 aparezca en la ecuación,

nada se altera al suponer que el valor promedio de *u* en la población, es cero. Matemáticamente,

[2.5] $E(u) = 0$.

El supuesto (2.5) no dice nada acerca de la relación entre *u* y *x*, sólo afirma algo acerca de la distribución de los efectos no observables en la población. Al usar como ilustración los ejemplos anteriores, puede verse que el supuesto (2.5) no es muy restrictivo. *En el ejemplo 2.1, no se modifica nada al normalizar los factores no observados que afectan el rendimiento del frijol de soya, por ejemplo la calidad de la tierra, para hacer que en la población de todas las parcelas cultivadas su promedio sea cero*. Lo mismo ocurre con los factores no observados del ejemplo 2.2. Sin pérdida de generalidad, se puede suponer que en la población de todas las personas trabajadoras, cosas como la capacidad promedio sea cero.

Ahora, volvamos al supuesto crucial sobre la manera en que están relacionadas *u* y *x*. Una medida natural de la relación entre dos variables aleatorias es el *coeficiente de correlación*. (Vean definición y propiedades en el apéndice B.) Si *u* y *x* no están correlacionadas, entonces, como variables aleatorias, no están relacionadas linealmente. Suponer que *u* y *x* no están correlacionadas es un avance para definir el sentido en el que *u* y *x* estarán relacionadas en la ecuación (2.1). Sin embargo, el avance no es suficiente, ya que la correlación sólo mide dependencia lineal entre *u* y *x*. La correlación tiene una propiedad un poco contra intuitiva: es posible que *u* no esté correlacionada con *x* y que, sin embargo, esté correlacionada con funciones de *x* como, por ejemplo, x^2. (Vean una mayor explicación en la sección B.4.) Esta posibilidad no es aceptable para la mayoría de los propósitos de la regresión, ya que causa problemas para interpretar el modelo y obtener propiedades estadísticas. Un supuesto mejor involucra el *valor esperado de u dado x*.

Como *u* y *x* son variables aleatorias, se puede definir la distribución condicional de *u* dado cualquier valor de *x*. En particular, para cada *x*, se puede obtener el valor esperado (o promedio) de *u* en la porción de la población descrita por el valor de *x*. El supuesto crucial es que el valor promedio de *u* no depende del valor de *x*. Este supuesto se expresa como

[2.6] $E(u \mid x) = E(u)$.

La ecuación (2.6) indica que el valor promedio de los factores no observables es el mismo en todas las fracciones de la población determinados por los valores de *x* y que este promedio común es necesariamente igual al promedio de *u* en toda la población. Cuando se satisface el supuesto (2.6) se dice que *u es media independiente de x*. (Por supuesto, la independencia de la media es una consecuencia de la independencia entre *u* y *x*, un supuesto usado frecuentemente en probabilidad y estadística básicas.) Combinando la independencia de media con el supuesto (2.5), se obtiene el *supuesto de media condicional cero*, $E(u \mid x) = 0$. Es vital recordar que la ecuación (2.6) es el supuesto importante; el supuesto (2.5) sólo define el intercepto, β_0.

¿Qué conlleva la ecuación (2.6) en el ejemplo del salario? Para simplificar el análisis, supongan que u es capacidad innata. Entonces la ecuación (2.6) requiere que el promedio de la capacidad sea el mismo sin importar los años de educación escolar. Por ejem-

plo, si E (*capaci* | *8*) denota las capacidades promedio en el grupo de personas con ocho años de educación escolar y E (*capcaci* | *16*) denota las capacidades promedio entre todas las personas con 16 años de educación escolar, entonces la ecuación (2.6) implica que estos valores deben ser iguales. En efecto, el promedio de capacidad debe ser el mismo en todos los niveles de educación. *Si, por ejemplo, se piensa que la capacidad promedio aumenta con los años de educación, entonces la ecuación (2.6) es falsa*. (*Esto ocurriría si, en promedio, las personas con mayor capacidad eligieran tener más educación*.) Como las capacidades innatas no pueden ser observadas, no hay manera de saber si la capacidad promedio es la misma en todos los niveles de educación. Pero esta es una cuestión que debe ser tomada en consideración antes de confiar en el análisis de regresión simple.

En el ejemplo del fertilizante, si las cantidades de fertilizante se eligen independientemente de otras características de las parcelas, entonces la ecuación (2.6) es válida: la calidad promedio de la tierra no dependerá de la cantidad de fertilizante. Pero, si a las parcelas de mejor calidad se les aplica más fertilizante, entonces el valor esperado de u variará de acuerdo con el nivel del fertilizante y la ecuación (2.6) no es válida.

El supuesto de media condicional cero proporciona otra interpretación de β_1 que suele ser útil. Tomando el valor esperado de (2.1) condicionado a x usando $E(u \mid x) = 0$ se tiene

[2.8] $$E(y \mid x) = \beta_0 + \beta_1 x.$$

La ecuación (2.8) muestra que la función de regresión poblacional (FRP), $E(y \mid x)$, es una función lineal de x. La linealidad significa que por cada aumento de una unidad en x el valor esperado de y se modifica en la cantidad β_1. Dado cualquier valor de x, la distribución de y está centrada en $E(y \mid x)$, como se ilustra en la figura 2.1.

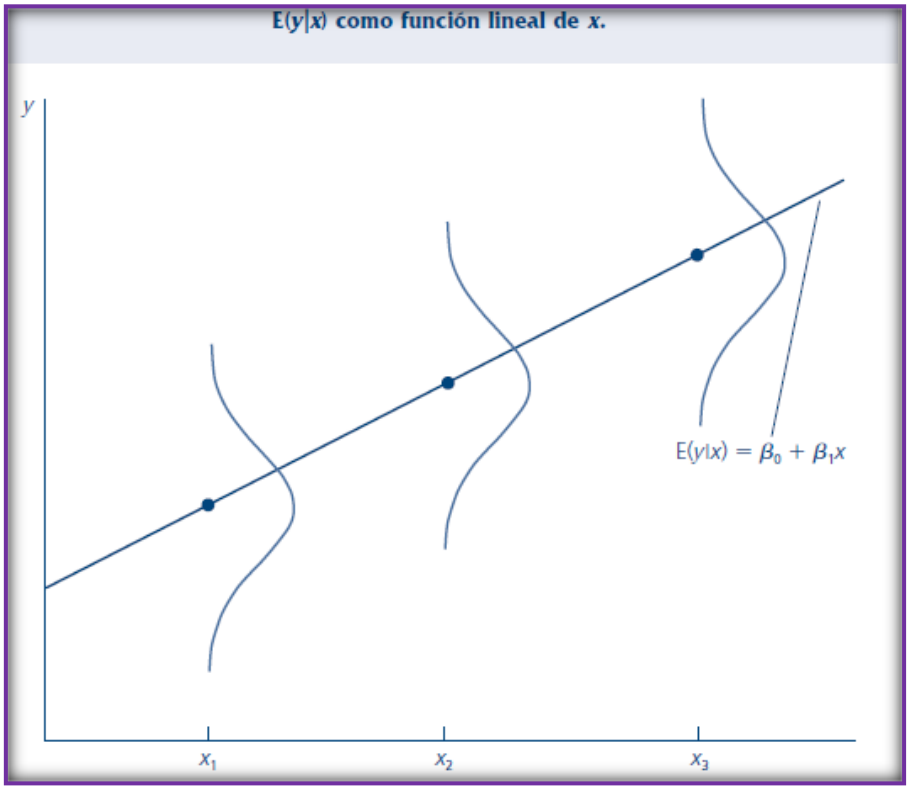

Figura 2.1

Es importante entender que la ecuación (2.8) dice cómo varía el valor promedio de y de acuerdo con la variación de x; esta ecuación no dice que y sea igual a $\beta_0 + \beta_1 x$ para cada una de las unidades de la población. Supongan, por ejemplo, que x sea el promedio obtenido en el bachillerato, e y sea el promedio obtenido en la universidad y que, además, se sepa que $E(promUniv \mid promBach) = 1.5 + 0.5\, promBach$. [Claro que, en la práctica, nunca se conocen ni el intercepto ni la pendiente poblacional, pero para entender la ecuación (2.8) resulta útil suponer, por un momento, que se conocen.] Esta ecuación sobre calificaciones proporciona el promedio de las calificaciones de universidad de entre todos los estudiantes que tienen una determinada calificación de bachillerato. De esta manera, suponga que *Prombach* = 3.6. Entonces, el promedio de *Promuniv* de todos los que terminan el bachillerato y asisten a la universidad y que en el bachillerato tuvieron *Prombach* = 3.6 es $1.5 + 0.5\,(3.6) = 3.3$. *No se está diciendo que todos los estudiantes que tengan Prombach = 3.6 tendrán 3.3 como promedio en la universidad; es claro que esto es falso*. La FRP da una relación entre el promedio de y y diferentes valores de x. Algunos de los estudiantes que tengan *Prombach* = 3.6 obtendrán en la universidad un promedio de calificaciones mayor a 3.3 y otros obtendrán promedios más bajos. Que el verdadero *Promuniv* sea mayor o menor a 3.3 depende de los factores no observables en u, y éstos varían entre los estudiantes aun entre los que pertenecen a la porción de la población con *Prombach* = 3.6.

Dado el supuesto de media condicional cero $E(u \mid x) = 0$, es útil ver la ecuación (2.1) como una que divide a y en dos componentes. A la parte $\beta_0 + \beta_1 x$, que representa $E(y \mid x)$, se le llama *parte sistemática* de y, es decir, es la parte de y explicada por x y a u se le llama la parte *no sistemática*, o *la parte de y que no es explicada por x*. En el capítulo 3, en donde se introducirá más de una variable explicativa, se analizará cómo determinar qué tan grande es la parte sistemática con relación a la parte no sistemática.

En la sección siguiente, se usarán los supuestos (2.5) y (2.6) para obtener estimadores de β_0 y β_1 a partir de una muestra aleatoria de datos dada. El supuesto de media condicional cero también tiene un papel crucial en el análisis estadístico de la sección 2.6.

2.2 *Obtención de las estimaciones de mínimos cuadrados ordinarios*

[No incluida aquí.]

2.3 *Propiedades de MCO en cualquier muestra de datos [MCO: Mínimos Cuadrados Ordinarios]*

En la sección anterior, se dedujeron las fórmulas para las estimaciones, por MCO, del intercepto y de la pendiente. En esta sección, se verán algunas otras propiedades algebraicas de la línea de regresión ajustada de MCO. Hay que recordar que estas propiedades, por construcción, son válidas para cualquier muestra de datos. La tarea más difícil —considerar las propiedades de MCO en todas las posibles muestras aleatorias de datos— se posponen hasta la sección 2.5.

Varias de las propiedades algebraicas que se van a deducir pueden parecer muy simples. Sin embargo, entenderlas ayudará a comprender lo que pasa con las estimaciones de MCO y con los estadísticos con ellos relacionados al manipular los datos de ciertas

maneras, por ejemplo, cuando se modifican las unidades de medición de las variables dependiente o independiente.

Valores ajustados y residuales

Se supone que las estimaciones del intercepto y de la pendiente, β°_0 y β°_1, han sido obtenidas para los datos muestrales dados.[3] Una vez que se tienen β°_0 y β°_1, se puede obtener el valor ajustado y°_i correspondiente a cada observación. [Esto se indica en la ecuación (2.20).][4] Por definición, todos los valores ajustados y°_i se encuentran sobre la línea de regresión de MCO. El residual de MCO correspondiente a la observación i, u°_i, es la diferencia entre y_i y su valor ajustado, como se indica en la ecuación (2.21). Si u°_i es positivo, la línea predice un valor inferior al de y_i; si u°_i es negativo, la línea predice un valor superior al de y_i. Lo ideal para la observación i es cuando $u^{\circ}_i = 0$, pero en la mayoría de los casos, todos los residuales son distintos de cero. *En otras palabras, no es necesario que ninguno de los puntos de los datos se encuentre exactamente sobre la línea de MCO*.

Ejemplo 2.6 Sueldo de los CEO y rendimiento sobre el capital

La tabla 2.2 contiene una lista de las primeras 15 observaciones de la base de datos de los CEO, así como los valores ajustados, a los que se les llama *salarygorro* (*sueldogorro*), y los residuales, a los que se les llama *ugorro*.

Valores ajustados y residuales de los primeros 15 CEO

obsno	*roe*	*salary (sueldo)*	*salarygorro*	*ugorro*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

Tabla 2.2

[3] A lo largo de estas clases, a los estimadores de los parámetros de población β_i se los indicará como β°_i agregando el supraíndice $^{\circ}$ al correspondiente parámetro.
[4] La ecuación [2.20] es la siguiente: $y^{\circ}_i = \beta^{\circ}_0 + \beta^{\circ}_1 x_i$.

Los primeros cuatro CEO tienen un sueldo menor que el que se predice empleando la línea de regresión de MCO (2.26);[5] en otras palabras, dado únicamente el roe de las empresas, estos CEO ganan menos de lo que se predice. Como se puede ver, por el *ugorro* positivo, el quinto CEO gana más de lo que se predice de acuerdo con la línea de regresión de MCO.

Propiedades algebraicas de los estadísticos de MCO

Las estimaciones de MCO y sus correspondientes estadísticos tienen varias propiedades útiles. A continuación se verán las tres más importantes.

(1) La suma, y por tanto el promedio muestral de los residuales de MCO, es cero. Matemáticamente,

[2.30] $\sum_{i=1}^{n} \hat{u}_i = 0.$

Esta propiedad no necesita ser probada; es consecuencia inmediata de la condición de primer orden (2.14) de MCO,[6] si se recuerda que los residuales están definidos por $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. En otras palabras, las estimaciones de MCO $\hat{\beta}_0$ y $\hat{\beta}_1$ se eligen de manera que la suma de los residuales sea cero (para cualquier base de datos). Esto no dice nada acerca del residual de una determinada observación i.

(2) La covarianza muestral entre los regresores y los residuales de MCO es cero. Esto es consecuencia de la condición de primer orden (2.15), que en términos de los residuales puede expresarse como

[2.31] $\sum_{i=1}^{n} x_i \hat{u}_i = 0.$

El promedio muestral de los residuales de MCO es cero, por lo que el lado izquierdo de la ecuación (2.31) es proporcional a la covarianza entre las x_i y los $\hat{u}_i$.

(3) [Denotando como $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ e $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$], el punto $(\bar{x}, \bar{y})$ se encuentra siempre sobre la línea de regresión de MCO. En otras palabras, si en la ecuación (2.20) se sustituye x por $\bar{x}$, el valor predicho es $\bar{y}$. Esto es exactamente lo que dice la ecuación (2.20).[7]

Escribiendo cada y_i como su valor ajustado, más su residual, se obtiene otra manera de interpretar la regresión de MCO. Para cada i se tiene

---

[5] Esta línea de regresión fue estimada de la siguiente manera:

[2.26] $salarygorro = 963{,}191 + 18{,}501\ roe$

[6] Dada una muestra de datos, se eligen estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ que resuelvan las siguientes ecuaciones:

[2.14] $n^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

[2.15] $n^{-1} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$

De estas ecuaciones se pueden obtener soluciones para $\hat{\beta}_0$ y $\hat{\beta}_1$

[7] Esta ecuación es la siguiente:

[2.20] $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.$

[2.32] $y_i = \hat{y}_i + \hat{u}_i$.

De acuerdo con la propiedad (1), el promedio de los residuales es cero; lo que es equivalente a que el promedio muestral de los valores ajustados, $\hat{y}_i$, es igual al promedio muestral de las y_i, es decir $\hat{y}^m = y^m$. Además, con base en las propiedades (1) y (2) se puede mostrar que la covarianza muestral entre $\hat{y}_i$ y $\hat{u}_i$ es cero. Por tanto, se puede considerar que el método de MCO descompone cada y_i en dos partes, un valor ajustado y un residual. Los valores ajustados y los residuales no están correlacionados en la muestra.

Se definen la *suma total de cuadrados (STC)*, la *suma explicada de cuadrados (SEC)* y la *suma residual de cuadrados (SRC)* (conocida también como suma de residuales cuadrados), como sigue:

[2.33] $STC \equiv \sum_{i=1}^{n} (y_i - y^m)^2$.

[2.34] $SEC \equiv \sum_{i=1}^{n} (\hat{y}_i - y^m)^2$.

[2.35] $SRC \equiv \sum_{i=1}^{n} (\hat{u}_i)^2$.

La STC es una medida de la variación muestral total en las y_i; es decir, mide qué tan dispersas están las y_i en la muestra. Si se divide la STC por n - 1, se obtiene la varianza muestral de y, que se analiza en el apéndice C. De manera similar, la SEC mide la variación muestral de las $\hat{y}_i$ (donde se usa el hecho de que $\hat{y}^m = y^m$) y la SRC mide la variación muestral de los $\hat{u}_i$. La variación total de y puede expresarse como la suma de la variación explicada más la variación no explicada *SRC*. Por tanto,

[2.36] $STC = SEC + SRC$.

Sólo una advertencia acerca de STC, SEC y SRC. No hay un acuerdo general para los nombres y las siglas que se emplean para las tres cantidades definidas en las ecuaciones (2.33), (2.34) y (2.35). Para la suma total de cuadrados se usa STC o SCT, de manera que hay un poco de confusión. Desafortunadamente, a la suma explicada de cuadrados suele llamársele también “suma de cuadrados de la regresión”. Si se emplea su abreviación natural para este término, con facilidad puede confundirse con el término “suma residual de cuadrados”. En algunos paquetes para regresión a la suma explicada de cuadrados se le llama “suma de cuadrados del modelo”.

Para complicar las cosas, a la suma residual de cuadrados se le suele llamar “suma de cuadrados de los errores”. Esto es en especial desafortunado ya que, como se verá en la sección 2.5, los errores y los residuales son cantidades diferentes. Por tanto, aquí a (2.35) se le llamará la suma residual de cuadrados o la suma de residuales cuadrados. Es preferible emplear la abreviación SRC para denotar la suma de residuales cuadrados, debido a que ésta es más común en los paquetes para econometría (SSR en los paquetes en inglés).

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? Recuerden 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% (*cæteris 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

[2.42] $\log(wage) = \beta_0 + \beta_1 educ + u,$

donde log (.) denota el logaritmo natural. (Vean en el apéndice A un repaso de los logaritmos.)

En particular, si $\Delta u = 0$, entonces

[2.43] $\%\Delta wage \approx (100.\beta_1)\,\Delta educ.$

Observen 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 (absoluto) 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 $u = 0$.

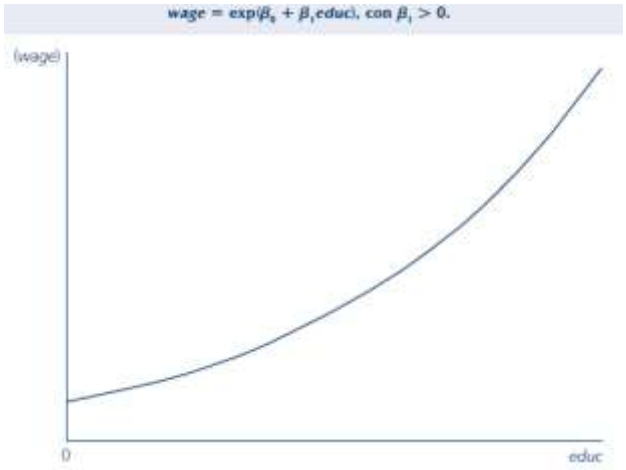

Figura 2.6

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. En otras palabras, $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen mediante una regresión por MCO de $log(wage)$ sobre $educ$.

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 se 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

[2.45] $$log(salary) = \beta_0 + \beta_1 log(sales) + u,$$

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

[2.46] $$\widehat{log\,(salary)} = 4.822 + 0.257\, log(sales) \qquad n = 209,\ R^2 = 0.211.$$

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 $log(c1)$ a ambos lados, se obtiene $log(c_1) + log(yi) = [log(c_1) + \beta_0] + \beta_1 x_i + u_i$, es decir $log(c_1 y_i) = [log(c_1) + \beta_0] + \beta_1 x_i + u_i$. (Recuerden 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 e 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.

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	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Tabla 2.3

En la última columna de la tabla 2.3 se explica la interpretación de β_1. En el modelo log-nivel, a $100.\beta_1$ se lo 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 un 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 lo 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 $y = \beta_o + \beta_1 x + u$. *La clave es que esta ecuación es lineal en los parámetros β_o 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: $cons = \square\beta_o + \beta_1 \sqrt{ing} + u$, donde *cons* es el consumo anual e *ing* 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: $cons=1/(\beta_o+\beta_1 \sqrt{ing})+u$. 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 $y = \beta_o + \beta_1 x + u$, y se dijo que el principal supuesto para que el análisis de regresión simple sea útil es que el valor esperado de u dado cualquier valor de x sea cero. En las secciones 2.2, 2.3 y 2.4, se discutieron las propiedades algebraicas de las estimaciones de MCO. Ahora se vuelve al modelo poblacional para estudiar las propiedades estadísticas de MCO. En otras palabras, ahora β°_o y β°_1 se considerarán como *estimadores* de los parámetros β_0 y β_1 que aparecen en el modelo poblacional. Esto significa que se estudiarán las propiedades de las distribuciones de los β°_0 y β°_1 que resultan de las diversas muestras aleatorias que es posible obtener de la población. (En el apéndice C se encuentra la definición de estimador, así como un repaso de sus propiedades.)

Insesgamiento de los estimadores MCO

Se empezará por demostrar el insesgamiento de los estimadores de MCO bajo un conjunto sencillo de supuestos. Para referencias futuras, estos supuestos se enumeran empleando el prefijo "RLS" como siglas de regresión lineal simple. El primer supuesto define el modelo poblacional.

Supuesto RLS.1 Linealidad de los parámetros

En el modelo poblacional, la variable dependiente, y, está relacionada con la variable independiente, x, y con el error (o perturbación), u, de la manera siguiente

[2.47] $$y = \beta_o + \beta_1 x + u$$

donde β_o y β_1 representan los parámetros poblacionales, del intercepto y pendiente, respectivamente.

Para ser realistas, al plantear el modelo poblacional, y, x, y u son consideradas como variables aleatorias. En la sección 2.1 se analizó, con cierta profundidad, la interpretación de este modelo y se dieron algunos ejemplos. En la sección anterior, se vio que la ecuación (2.47) no es tan restrictiva como a primera vista pareciera; eligiendo y y x de manera adecuada, se pueden obtener interesantes relaciones no lineales (por ejemplo, modelos de elasticidad constante).

La idea es usar los datos de y y de x para estimar los parámetros β_0 y, en especial, β_1. Se supone que los datos se obtienen de una muestra aleatoria. (Vean en el apéndice C un repaso sobre muestreo aleatorio.)

Supuesto RLS.2 Muestreo aleatorio

Se cuenta con una muestra aleatoria de tamaño n, $\{(x_i, y_i): i = 1, 2,..., n\}$, que sigue el modelo poblacional de la ecuación.

En capítulos posteriores sobre el análisis de series de tiempo y problemas de selección de la muestra habrá que ocuparse de la falla del supuesto de muestreo aleatorio. No todas las muestras de corte transversal pueden considerarse como resultado de un muestreo aleatorio, aunque muchas pueden serlo.

En términos de la muestra aleatoria, la ecuación (2.47) se expresa como

[2.48] $$y_i = \beta_o + \beta_1 x_i + u_i, i \square 1, 2,..., n,$$

donde u_i es el error o la perturbación para la observación i (por ejemplo, en la persona i, la empresa i, la ciudad i, etc.). Así que u_i contiene los efectos no observables de la observación i que afectan a y_i. Esta u_i no debe confundirse con los residuales, u°_i, definidos en la sección 2.3. Más adelante se analizará la relación entre los errores y los residuales. Para interpretar β_0 y β_1 en una determinada aplicación, la ecuación (2.47) es la más informativa, pero la ecuación (2.48) es necesaria para algunas deducciones estadísticas.

Dado un conjunto de datos, la relación (2.48) puede graficarse como se muestra en la figura 2.7.

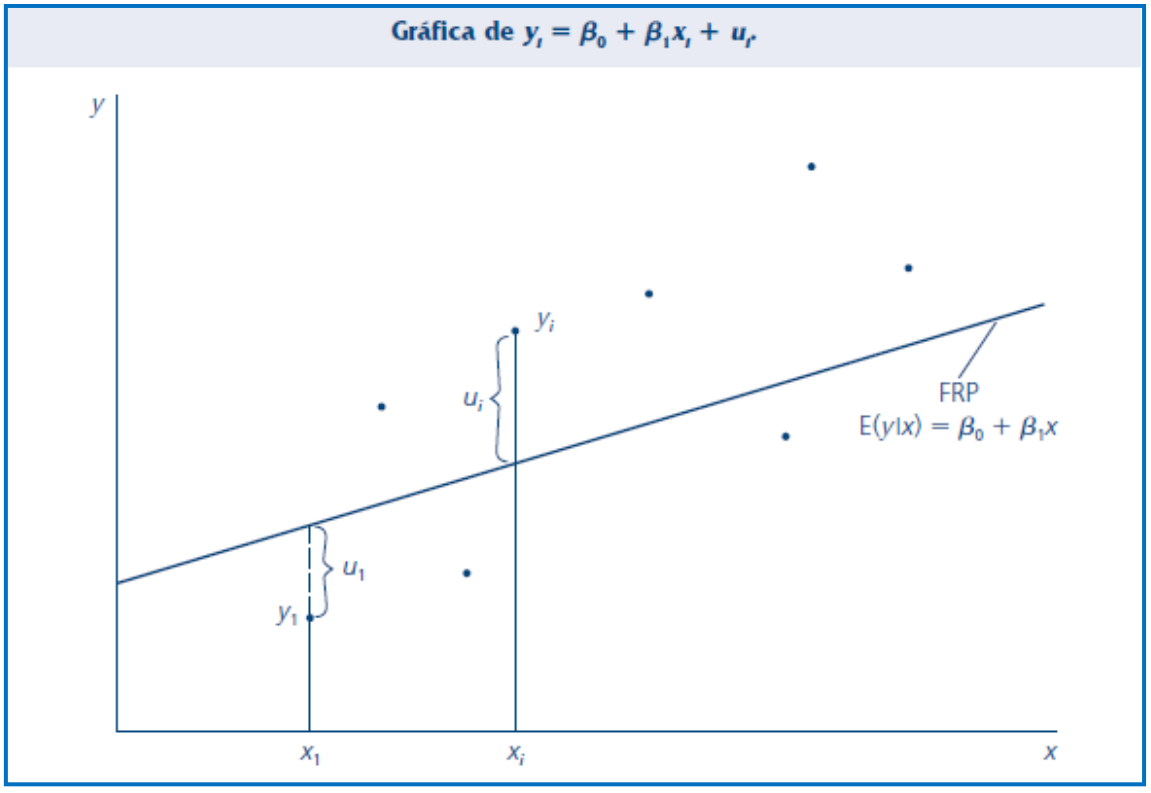

Figura 2.7

Con base en lo visto en la sección 2.2, las estimaciones de MCO del intercepto y de la pendiente sólo están definidas si en la muestra hay variación de la variable explicativa. A continuación se agrega a la lista de supuestos la variación de las x_i.

Supuesto RLS.3 Variación muestral de la variable explicativa

No todos los valores muestrales de x, a saber $\{x_i, i = 1,\ldots, n\}$, son iguales, es decir, no todos tienen el mismo valor.

Este es un supuesto muy débil —ciertamente poco útil de destacar, pero necesario de cualquier manera—. Si x varía en la población, entonces las muestras aleatorias típicamente mostrarán variación en x, a menos que la variación poblacional sea mínima o el tamaño de la muestra sea muy pequeño. Una sencilla inspección de los estadísticos de las x_i indicará si el supuesto RLS.3 se satisface o no: si la desviación estándar muestral de las x_i es cero, entonces el supuesto RLS.3 no se satisface; si no es así, este supuesto se satisface.

Por último, con objeto de obtener estimadores insesgados de β_0 y β_1, es necesario imponer el supuesto de media condicional cero, visto con cierto detalle en la sección 2.1. A continuación se agrega éste a la lista de supuestos.

Supuesto RLS.4 Media condicional cero

Para todo valor de la variable explicativa, el valor esperado del error u es cero. Es decir,

$$E(u \mid x) = 0.$$

En una muestra aleatoria, este supuesto implica que $E(u_i \mid x_i) = 0$, para toda $i=1, 2,\ldots, n$.

Además de restringir la relación que hay en la población entre u y x, el supuesto de media condicional cero —junto con el supuesto del muestreo aleatorio— permite una conveniente simplificación técnica. En particular, es posible obtener las propiedades estadísticas de los estimadores de MCO como condicionales sobre los valores de x_i en la muestra. *Técnicamente, al hacer cálculos estadísticos, condicionar sobre los valores muestrales de la variable independiente es lo mismo que tratar a las x_i como fijas en muestras repetidas, lo cual se entiende como sigue*. Primero se escogen n valores muestrales para $x_1, x_2,\ldots, x_n$. (Éstos pueden repetirse.) Dados estos valores, se obtiene después una muestra de y (efectivamente al obtener una muestra aleatoria de las u_i). A continuación, se obtiene otra muestra de y, usando los mismos valores para $x_1, x_2,\ldots, x_n$. Después se obtiene otra muestra de y, usando los mismos $x_1, x_2,\ldots, x_n$. Y así sucesivamente.

La situación de que las x_i sean fijas en muestras repetidas no es muy realista en los contextos no experimentales. Por ejemplo, al muestrear individuos para el caso del salario y la educación, no tiene mucho sentido pensar en elegir de antemano los valores de *educ* y después muestrear individuos que tengan esos determinados niveles de educación. El muestreo aleatorio, en el que los individuos se eligen de manera aleatoria para conocer su salario y su educación, es representativo de la manera en que se obtienen las bases de datos para el análisis empírico en las ciencias sociales. Una vez que se supone que $E(u \mid x) = 0$, y que se tiene un muestreo aleatorio, no cambia nada en los cálculos al tratar a las x_i como no aleatorias. El riesgo es que el supuesto de que las x_i sean fijas en muestras repetidas siempre implica que u_i y x_i sean independientes. Para decidir si con

un análisis de regresión simple se van a obtener estimadores insesgados, es crucial pensar en los términos del supuesto RLS.4.

Teorema 2.1 Insesgamiento de los estimadores de MCO

Empleando los supuestos RLS.1 a RLS.4,

[2.53] $E(\hat{\beta}_0) = \beta_0$ y $E(\hat{\beta}_1) = \beta_1$,

para cualquier valor de β_0 y β_1. Es decir, $\hat{\beta}_0$ es un estimador insesgado de β_0 y $\hat{\beta}_1$ es un estimador insesgado de β_1. *Demostración en pág. 50 del libro.*

Hay que recordar que el insesgamiento es una propiedad de las distribuciones muestrales de $\hat{\beta}_0$ y $\hat{\beta}_1$, que no dice nada acerca de las estimaciones que se obtienen a partir de una determinada muestra. Se espera que, si la muestra es de alguna manera *representativa*, la estimación deberá estar "cerca" del valor poblacional. *Desafortunadamente, siempre es posible tener una muestra con la que la estimación puntual que se obtenga esté lejos de β_1, y nunca se podrá saber con seguridad si éste es el caso*. En el apéndice C se presenta un repaso sobre estimadores insesgados y un ejercicio de simulación en la tabla C.1 que ilustra el concepto de insesgamiento.

En general, el insesgamiento no se cumple cuando no se satisface alguno de los cuatro supuestos. Esto significa que es muy importante reflexionar sobre la veracidad de cada supuesto en la aplicación particular de que se trate. El supuesto RLS.1 requiere que y y x estén relacionadas linealmente y tengan una perturbación aditiva. Es claro que esto puede no darse. Pero también se sabe que pueden escogerse y y x de manera que den interesantes relaciones no lineales. Cuando la ecuación (2.47) no se satisface, se requieren otros métodos más avanzados que quedan fuera del alcance de este libro.

Más adelante habrá que relajar el supuesto RLS.2, el del muestreo aleatorio, al tratar el análisis de las series de tiempo. ¿Y respecto a su uso en análisis de cortes transversales? El muestreo aleatorio puede no satisfacerse en un corte transversal si la muestra no es representativa de la población subyacente; en realidad, hay bases de datos que se obtienen muestreando exageradamente, de manera intencionada, diferentes partes de la población. En los capítulos 9 y 17 se analizarán los problemas del muestreo no aleatorio.

Como ya se ha visto, el supuesto RLS.3 es casi siempre satisfecho en las aplicaciones interesantes de la regresión. Sin este supuesto no es posible ni siquiera obtener los estimadores de MCO.

El supuesto a considerar ahora es RLS.4 si éste se satisface, los estimadores de MCO son insesgados. Y de manera similar, si RLS.4 no se satisface, los estimadores de MCO son, por lo general, *sesgados*. Hay maneras de determinar cuál puede ser la dirección y el tamaño de este sesgo; esto se estudiará en el capítulo 3.

La posibilidad de que x esté correlacionada con u es una preocupación común en el análisis de regresión simple con datos no experimentales, como se indicó mediante varios ejemplos en la sección 2.1. Cuando se usa regresión simple, si u contiene factores que afectan a y y que están correlacionados con x, puede obtenerse una correlación

espuria: esto significa que se encuentra una relación entre y y x que en realidad se debe a factores no observados que afectan a y y que resultan estar correlacionados con x.

Ejemplo 2.12 Desempeño de los estudiantes en matemáticas y el programa de desayunos escolares

Sea *math10* el porcentaje de estudiantes que aprueban el examen estandarizado de matemáticas en el primer año de bachillerato de una escuela. Suponga que se desea estimar el efecto del programa federal de desayunos escolares sobre el desempeño de los estudiantes. Por supuesto, se espera que este programa tenga, un efecto *cæteris paribus*, positivo sobre el desempeño: si todos los demás factores permanecen constantes, si a un estudiante que es tan pobre como para no tener una buena alimentación se le beneficia con el programa de desayunos escolares, su desempeño deberá mejorar. Sea *lnchprg* el porcentaje de estudiantes beneficiados con el programa de desayunos escolares. Entonces, un modelo de regresión simple es

[2.54] $$math10 = \square\ \beta_0 + \beta_1\, lnchprg + u,$$

donde u contiene características de la escuela y del estudiante que afectan el desempeño general de la escuela. Empleando los datos de una base, correspondientes a 408 escuelas de Michigan durante el ciclo escolar 1992-1993, se obtiene

[2.55] $$\widehat{math10} = 32.14 - 0.319\, lnchprg, \qquad n=408\ R^2= 0.171$$

Esta ecuación predice que si el porcentaje de alumnos que reciben el desayuno escolar aumenta 10 puntos porcentuales, el porcentaje de alumnos que aprueban el examen de matemáticas *decrecerá* aproximadamente 3.2 puntos porcentuales. ¿Se puede creer que un aumento en el porcentaje de estudiantes que reciben el desayuno escolar cause un peor desempeño? Casi seguro que no. Una explicación es que el término del error u de la ecuación (2.54) esté correlacionado con *lnchprg*. En realidad, u contiene factores como la tasa de pobreza de los niños que asisten a la escuela, lo cual afecta el desempeño del estudiante y está fuertemente correlacionado con que se otorgue el programa del desayuno. Variables tales como la calidad de la escuela y los recursos de que dispone están contenidas en u, y pueden estar correlacionadas con *lnchprg*. *Es importante recordar que la estimación $\beta_1 = -0.319$ sólo es válida para esta muestra en particular, pero su signo y magnitud hacen sospechar que u y x estén correlacionadas, haciendo que la regresión simple esté sesgada*.

Además de las variables omitidas, hay otras razones que hacen que x esté correlacionada con u en el modelo de regresión simple. Dado que este mismo tema surge en el análisis de regresión múltiple, el tratamiento sistemático de este problema se pospondrá hasta entonces.

Varianza de los estimadores de mínimos cuadrados ordinarios

Además de saber que la distribución muestral de $\hat{\beta}_1$ está centrada en β_1 ($\hat{\beta}_1$ es insesgado), también es importante saber qué tanto puede esperarse que $\hat{\beta}_1$ se aleje, en promedio, de β_1. Entre otras cosas, esto permite elegir el mejor estimador de todos, o por lo menos, de una amplia clase de estimadores insesgados. La medida de la dispersión de la distribución de $\hat{\beta}_1$ (y de $\hat{\beta}_0$) con la que es más fácil trabajar es con la varianza o su

raíz cuadrada, la desviación estándar. (Vean en el apéndice C un análisis más detallado.)

La varianza de los estimadores de MCO puede ser calculada con los supuestos RLS.1 a RLS.4. Sin embargo, estas expresiones son un poco complicadas. En lugar de esto, se agregará un supuesto tradicional en el análisis de corte transversal y que establece que la varianza de los factores inobservables, u, condicionales en x, es constante. Esto se conoce como el supuesto de *homocedasticidad* o de "varianza constante".

Supuesto RLS.5 Homocedasticidad

El error u tiene la misma varianza para cualquier valor de la variable explicativa. En otras palabras,

$$Var\,(u \mid x) = \sigma^2.$$

Hay que señalar que el supuesto de homocedasticidad es totalmente distinto al de media condicional cero, $E\,(u \mid x) = 0$. El supuesto RLS.4 se refiere al valor esperado de u, mientras que el RLS.5 está relacionado con la varianza de u (ambos condicionales sobre x). Recuerde que el insesgamiento de los estimadores de MCO se demostró sin usar el supuesto RLS.5: el supuesto de homocedasticidad no se emplea en la demostración de que β°_0 y β°_1 son insesgados. El supuesto RLS.5 se agrega debido a que simplifica los cálculos de las varianzas de β°_0 y β°_1 y a que implica que los mínimos cuadrados ordinarios tienen ciertas propiedades de *eficiencia*, que se verán en el capítulo 3. Si se supone que u y x son independientes, entonces la distribución de u dada x no depende de x, y entonces $E\,(u \mid x) = E(u) = 0$ y $Var(u \mid x) = \sigma^2$. Pero la independencia es algunas veces muy fuerte de suponer.

Como $Var\,(u \mid x) = E(u^2 \mid x) - [E(u \mid x)]^2$ y $E(u \mid x) = 0$, $\sigma^2 = E(u^2 \mid x)$, lo cual significa que σ^2 es también la esperanza *incondicional* de u^2. Por tanto, $\sigma^2 = E\,(u^2) = Var\,(u)$, debido a que $E\,(u) = 0$. En otras palabras, σ^2 es la varianza *incondicional* de u y por esto a σ^2 también se le suele llamar la varianza del error o varianza de la perturbación. La raíz cuadrada de σ^2, *σ, es la desviación estándar del error*. Una σ mayor, indica que la distribución de los factores inobservables que afectan a y tiene una mayor dispersión.

Con frecuencia es útil escribir los supuestos RLS.4 y RLS.5 en términos de la media condicional y de la varianza condicional de y:

[2.55] $$E\,(y \mid x) = \beta_0 + \beta_1 x.$$

[2.56] $$Var\,(y \mid x) = \sigma^2.$$

En otras palabras, la esperanza condicional de y dada x es lineal en x, pero la varianza de y dada x es constante. Esta situación se grafica en la figura 2.8 donde $\beta_0 > 0$ y $\beta_1 > 0$.

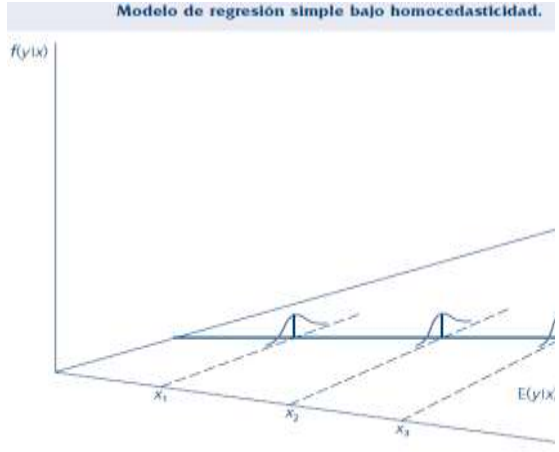

Figura 2.8

Cuando *Var (u | x)* depende de *x*, se dice que el término del error muestra ***heterocedasticidad*** (o varianza no constante). Como *Var (u | x) = Var (y | x),* la heterocedasticidad está presente siempre que *Var (y | x)* sea función de *x*.

Ejemplo 2.13 Heterocedasticidad en una ecuación del salario

Con objeto de obtener un estimador insesgado del efecto *cæteris paribus* de *educ* sobre *wage*, es necesario suponer que $E(u \mid educ) = 0$, y esto implica que $E(wage \mid educ) = \beta_0+\beta_1 educ$. Si se agrega el supuesto de homocedasticidad, luego $Var(u \mid educ)=\sigma^2$ no depende del nivel de educación, lo que es lo mismo que suponer que $Var(wage \mid educ)=\sigma^2$. Por tanto, mientras se permite que el salario promedio aumente con el nivel de educación —esta es la tasa de incremento que se quiere estimar— se supone que la variabilidad del salario en torno a su media es constante en todos los niveles de educación. Esto no parece estar de acuerdo con la realidad. Parece ser más posible que las personas con más educación tengan más intereses y más oportunidades de trabajo, lo que puede hacer que a niveles de educación más altos haya mayor variabilidad en el salario. Las personas con niveles de educación muy bajos tienen menos oportunidades y suelen tener que trabajar por el salario mínimo; esto contribuye a reducir la variabilidad del salario en los niveles de educación bajos. Esta situación se muestra en la figura 2.9. Por último, si el supuesto RLS.5 se satisface o no es una cuestión empírica y en el capítulo 8 se muestra cómo probar el supuesto RLS.5.

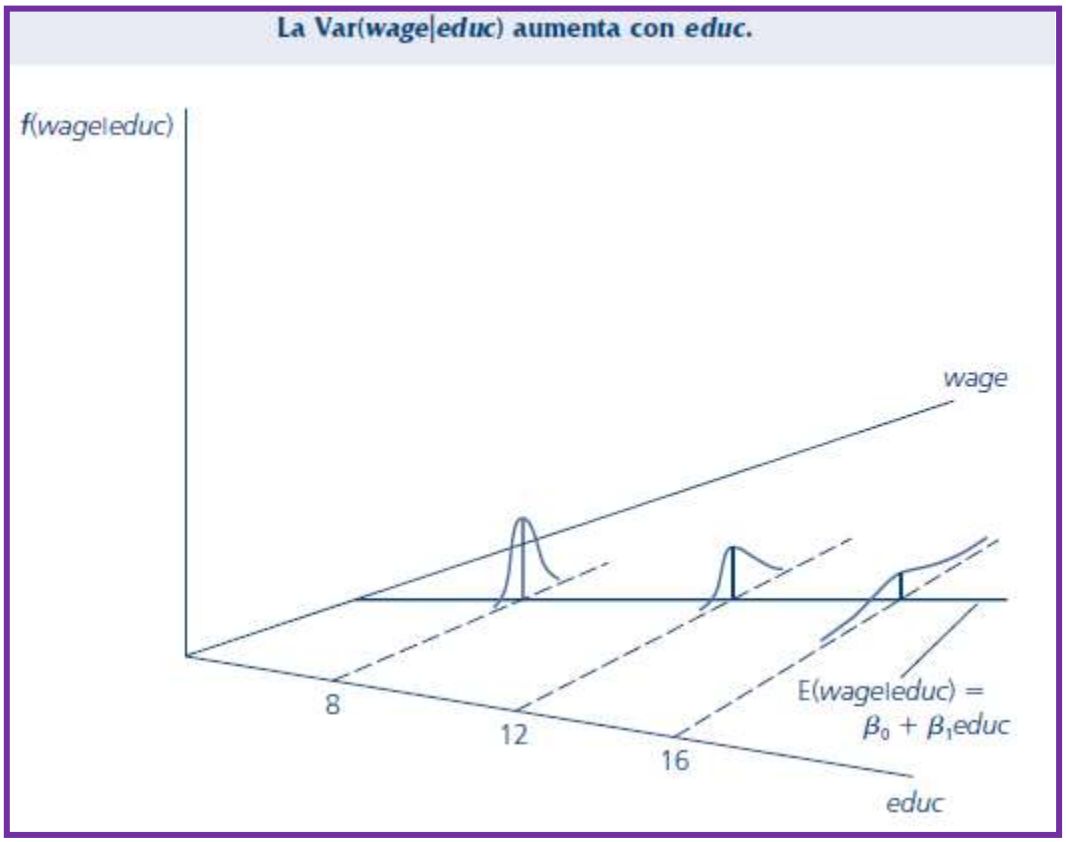

Figura 2.9 – Heterocedasticidad

Una vez que se asume la homocedasticidad se está listo para establecer lo siguiente:

Teorema 2.2 Varianza de muestreo de los estimadores de MCO

Bajo los supuestos RLS.1 a RLS.5,

[2.57] $Var\,(\beta^{\circ}_1) = \sigma^2 / \{\sum_{i=1}^{n} (x_i - x^m)^2\} = \sigma^2 / STC_x,$

[2.58] $Var\,(\beta^{\circ}_0) = \sigma^2\, n^{-1} \sum_{i=1}^{n} x_i^2 / \{\sum_{i=1}^{n} (x_i - x^m)^2\},$

donde éstos son condicionales sobre los valores muestrales $\{x_1, \ldots, x_n\}$. *Para su demostración, véase la página 55 del libro.*

Las ecuaciones (2.57) y (2.58) son las fórmulas "estándar" para el análisis de regresión simple, y no son válidas en presencia de heterocedasticidad. Esto será importante cuando se estudien los intervalos de confianza y las pruebas de hipótesis en el análisis de regresión múltiple.

En la mayoría de los casos, lo que interesa es la $Var\,(\beta^{\circ}_1)$. Se puede resumir fácilmente cómo esta varianza depende de la varianza del error, σ^2, y de la variación total en $\{x_1, x_2, \ldots, x_n\}$, STC_x. Primero, cuanto mayor sea la varianza del error, mayor es $Var\,(\beta^{\circ}_1)$. Esto tiene sentido, ya que una mayor variación en los factores no observables que afectan a y hace más difícil estimar β_1 con precisión. *Por otro lado, en la variable independiente se prefiere mayor variabilidad: a medida que aumenta la variabilidad de las x_i la varianza de β_1 disminuye. También esto es intuitivamente correcto, ya que cuanto más dispersa sea la muestra de las variables independientes, más fácil será hallar la relación entre $E(y|x)$ y x*. Es decir, será más sencillo estimar β_1. Si hay poca variación en las x_i, entonces puede ser difícil hallar cómo varía $E\,(y|x)$ con la variación en x. A medida que se incrementa el tamaño de la muestra, también aumenta la variación total de las x_i. Por tanto, un tamaño de muestra mayor da como resultado una varianza menor de las β°_1.

Este análisis muestra que, si lo que interesa es β_1 y si se tiene la posibilidad de elegir, entonces se debe escoger que las x_i estén tan dispersas como sea posible. Esto es factible algunas veces, cuando se trata de datos experimentales, pero en las ciencias sociales rara vez se puede tener esos lujos: por lo general, hay que conformarse con las x_i que se obtengan mediante un muestreo aleatorio. Algunas veces, se tiene la oportunidad de contar con tamaños de muestra mayores, aunque esto puede ser costoso.

Para la construcción de intervalos de confianza y para la obtención de estadísticos de prueba, será necesario trabajar con las desviaciones estándar de β°_1 y β°_0, *de* (β°_1) y *de*(β°_0). Recuerden que éstas se obtienen al calcular la raíz cuadrada de las varianzas dadas en las ecuaciones (2.57) y (2.58). En particular, $de\,(\beta^{\circ}_1) = \sigma/\sqrt{STC_x}$, donde σ es la raíz cuadrada de σ^2, y $\sqrt{STC_x}$ es la raíz cuadrada de STC_x.

Estimación de la varianza del error

Las fórmulas de las ecuaciones (2.57) y (2.58) permiten aislar los factores que contribuyen a la $Var\,(\beta^{\circ}_1)$ y a la $Var\,(\beta^{\circ}_0)$. Pero estas fórmulas son desconocidas, salvo en el raro caso en que se conozca σ^2. No obstante, usando los datos puede estimarse σ^2, lo que entonces permite estimar la $Var\,(\beta^{\circ}_1)$ y la $Var\,(\beta^{\circ}_0)$.

Esta es una buena ocasión para hacer hincapié en la diferencia entre los *errores* (o perturbaciones) y los *residuales*, ya que es crucial para construir un estimador de σ^2. La ecuación (2.48) muestra cómo escribir el modelo poblacional en términos de una observación muestral aleatoria como $y_i = \beta_o + \beta_1 x_i + u_i$,, donde u_i es el error en la observación i. También se puede expresar y_i en términos de su valor ajustado y su residual como en la ecuación (2.32): $y_i = y^\circ_i + u^\circ_i = \beta^\circ_o + \beta^\circ_1 x_i + u^\circ_i$. Comparando estas dos ecuaciones, se ve que el error aparece en la ecuación que contiene los parámetros poblacionales, β_o y β_1. Por otro lado, los residuales aparecen en la ecuación estimada con β°_o y β°_1. *Los errores jamás son observables, mientras que los residuales se calculan a partir de los datos*.

Usando las ecuaciones (2.32) y (2.48) se pueden expresar los residuales en función de los errores:

[2.59] $$u^\circ_i = y_i - \beta^\circ_o - \beta^\circ_1 x_i = (\beta_o + \beta_1 x_i + u_i) - \beta^\circ_o - \beta^\circ_1 x_i =$$

$$= u_i - (\beta^\circ_o - \beta_o) - (\beta^\circ_1 - \beta_1) x_i.$$

Aunque el valor esperado de β°_o es igual a β_o, y lo mismo ocurre con β°_1, u°_i no es lo mismo que u_i. *Pero el valor esperado de la diferencia entre ellos sí es cero*.

Una vez comprendida la diferencia entre los errores y los residuales, se puede volver a la estimación de σ^2. Primero, $\sigma^2 = E(u^2)$, de manera que un "estimador" insesgado de σ^2 es $n^{-1} \Sigma_{i=1}^{n} u_i^2$. Por desgracia, éste no es un verdadero estimador, ya que los errores u_i no pueden observarse. Pero, se tienen estimaciones de las u_i, a saber, los residuales de MCO u°_i. *Sustituyendo los errores por los residuales de MCO, se tiene* $n^{-1} \Sigma_{i=1}^{n} u^{\circ 2}_i = SRC/n$. *Éste es un verdadero estimador, porque da una regla de cómo calcular su valor a partir de cualquier muestra de datos de x e y*. Una pequeña desventaja de este estimador es que es sesgado (aunque si n es grande el sesgo es pequeño). Como es fácil obtener un estimador insesgado, se usará uno de este tipo.

El estimador SRC/n es sesgado debido esencialmente a que no toma en cuenta dos restricciones que deben satisfacer los residuales de MCO. Estas restricciones están dadas por las dos condiciones de primer orden de MCO:

[2.60] $$\sum_{i=1}^{n} u^\circ_i = 0, \quad \sum_{i=1}^{n} x_i u^\circ_i = 0.$$

Una manera de ver estas restricciones es: si se conocen $n - 2$ residuales, entonces los otros dos pueden obtenerse empleando las restricciones que implican las condiciones de primer orden de las ecuaciones en (2.60). *Así, para los residuales de MCO hay sólo n-2 grados de libertad, a diferencia de los n grados de libertad de los errores*. Si en las ecuaciones en (2.60) se sustituye u°_i por u_i las restricciones ya no son válidas. En el estimador insesgado de σ^2 que se usará aquí, se hace un ajuste para tener en cuenta los grados de libertad:

[2.61] $$\sigma^{\circ 2} = (n - 2)^{-1} \sum_{i=1}^{n} u^{\circ 2}_i = SRC / (n-2).$$

Teorema 2.3 Estimación insesgada de σ^2

Bajo los supuestos RLS.1 a RLS.5,

$$E(\hat{\sigma}^2) = \sigma^2.$$

(Ver demostración en pág. 58 del libro).

Si en las fórmulas de la varianza (2.57) y (2.58) se sustituye $\hat{\sigma}^2$, se obtienen estimadores insesgados de $Var(\hat{\beta}_1)$ y $Var(\hat{\beta}_0)$. Más tarde se necesitarán estimadores de las desviaciones estándar de $\hat{\beta}_1$ y $\hat{\beta}_0$ y, para esto, se necesita estimar σ. El estimador natural de σ es

[2.62] $$\hat{\sigma} = \sqrt{\hat{\sigma}^2},$$

al que se le llama *error estándar de la regresión* (EER). (Otros nombres para $\hat{\sigma}$ son error estándar de la estimación y raíz del error cuadrático medio, pero no se usaran aquí.) Aunque $\hat{\sigma}$ no es un estimador insesgado de σ, se puede mostrar que es un estimador consistente de σ (vean el apéndice C) y será útil para nuestros propósitos.

La $\hat{\sigma}$ estimada es interesante porque es una estimación de la desviación estándar de los factores no observables que afectan a y; de manera equivalente, es una estimación de la desviación estándar de y después de haber eliminado el efecto de x. La mayoría de los paquetes para regresión dan el valor de $\hat{\sigma}$ junto con la R-cuadrada, el intercepto, la pendiente y otros estadísticos MCO (bajo alguno de los nombres dados arriba). Por ahora, el interés primordial es usar $\hat{\sigma}$ para estimar las desviaciones estándar de $\hat{\beta}_1$ y $\hat{\beta}_0$. Como $de(\hat{\beta}_1) = \sigma/\sqrt{STC_x}$, el estimador natural de $de(\hat{\beta}_1)$ es

$$de(\hat{\beta}_1) = \hat{\sigma}/\sqrt{STC_x} = \hat{\sigma}/(\sum_{i=1}^{n}(x_i - x^m)^2)^{1/2};$$

al que se le llama *error estándar de $\hat{\beta}_1$*. Observen que $ee(\hat{\beta}_1)$ se ve como una variable aleatoria cuando se piensa en aplicar MCO a diferentes muestras de y; esto es cierto porque $\hat{\sigma}$ varía en las distintas muestras. En una muestra dada, $se(\hat{\beta}_1)$ es un número, de la misma manera que $\hat{\beta}_1$ es sólo un número cuando se calcula a partir de los datos dados.

De manera similar, $ee(\hat{\beta}_0)$ se obtiene de $de(\hat{\beta}_0)$ sustituyendo σ por $\hat{\sigma}$. El error estándar de cualquier estimación da una idea de qué tan preciso es el estimador. Los errores estándar son de gran importancia en todo este libro; se usarán para construir estadísticos de prueba e intervalos de confianza para cada uno de los procedimientos econométricos que se estudien, a partir del capítulo 4.