Wooldridge - Análisis de regresión múltiple: Estimación (Selección)[1]

En el capítulo 2 se vio cómo usar el análisis de regresión simple para explicar una variable dependiente, *y*, como función de una sola variable independiente, *x*. El principal inconveniente del análisis de regresión simple en el trabajo empírico es que es muy difícil obtener conclusiones *cæteris paribus* de cómo afecta *x* a *y*: el supuesto clave RLS.4 —de que todos los demás factores que afectan a *y* no están correlacionados con *x*— a menudo no es realista.

El *análisis de regresión múltiple* es más adecuado para un análisis *cæteris paribus* debido a que permite controlar de manera explícita muchos otros factores que afectan en forma simultánea a la variable dependiente. Esto es importante tanto para probar teorías económicas como para evaluar los efectos de una política cuando hay que apoyarse en datos no experimentales. Debido a que los modelos de regresión múltiple pueden aceptar diversas variables explicativas que tal vez estén correlacionadas, puede esperarse inferir causalidad en casos en los que el análisis de regresión simple podría no dar buenos resultados.

Si al modelo se le agregan factores que pueden ser útiles para explicar *y*, entonces puede explicarse más de la variación en *y*. Por tanto, el análisis de regresión múltiple puede emplearse para construir mejores modelos para predecir la variable dependiente.

Otra ventaja del análisis de regresión múltiple es que puede incorporar relaciones con formas funcionales muy generales. En el modelo de regresión simple, en la ecuación únicamente puede aparecer una función de una sola variable explicativa. Como se verá, el modelo de regresión múltiple permite más flexibilidad.

En la sección 3.1 se introduce de manera formal el modelo de regresión múltiple y se analizan las ventajas de la regresión múltiple sobre la simple. En la sección 3.2 se demuestra cómo estimar los parámetros del modelo de regresión múltiple usando el método de mínimos cuadrados ordinarios. En las secciones 3.3, 3.4 y 3.5 se describen varias propiedades estadísticas de los estimadores de MCO, como el insesgamiento y la eficiencia.

El modelo de regresión múltiple sigue siendo el vehículo más empleado para el análisis empírico en la economía y en otras ciencias sociales. Asimismo, el método de mínimos cuadrados ordinarios se usa de manera general para estimar los parámetros del modelo de regresión múltiple.

3.1 *Motivación para la regresión múltiple*

El modelo con dos variables independientes

Se empezará con algunos ejemplos sencillos para mostrar el uso del análisis de regresión lineal múltiple para resolver problemas que no es posible resolver mediante regresión simple.

---

[1] Tomado en forma parcial de Jeffrey M. Wooldridge - *Introducción a la econometría. Un enfoque moderno*. 4a. edición, 2009, Capítulo 3. 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.

El primer ejemplo es una sencilla variación de la *ecuación del salario*, presentada en el capítulo 2, para obtener el efecto de la educación sobre el salario por hora:

[3.1] $$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u,$$

donde *exper* es años de experiencia en el mercado de trabajo. Por tanto, *wage* (salario) está determinada por las dos variables independientes o explicativas, educación y experiencia, y por otros factores no observados, contenidos en *u*. El interés principal sigue siendo el efecto de *educ* (educación) sobre *wage* (salario), manteniendo constantes todos los otros factores que afectan a *wage*; es decir, lo que interesa es el parámetro β_1.

Comparada con un análisis de regresión simple, en el que se relaciona *wage* con *educ*, la ecuación (3.1) extrae *exper* del término del error y la coloca de manera explícita en la ecuación. Dado que *exper* aparece en la ecuación, su coeficiente, β_2, mide el efecto *cæteris paribus* de *exper* sobre *wage*, que también es de cierto interés.

Como en la regresión simple, aquí también habrá que hacer supuestos acerca de la relación de *u* en la ecuación (3.1) con las variables independientes *educ* y *exper*. Pero, como se verá en la sección 3.2, hay algo de lo que se puede estar seguro: como en la ecuación (3.1) aparece la experiencia de manera explícita, se podrá medir el efecto de la educación sobre el salario, manteniendo constante la experiencia. *Con un análisis de regresión simple —en el cual exper forma parte del término del error— hay que suponer que la experiencia no está correlacionada con la educación, un supuesto cuestionable*.

Como segundo ejemplo, considere el problema de explicar *el efecto del gasto por estudiante (expend) sobre la calificación promedio en el examen estandarizado (avgscore) a nivel de bachillerato*. Suponga que la calificación promedio en el examen depende del financiamiento, del ingreso familiar promedio (*avginc*) y de otros factores no observables:

[3.2] $$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u.$$

El coeficiente de interés para los propósitos de las políticas es β1, el efecto *cæteris paribus* de *expend* sobre *avgscore*. Incluir *avginc* de manera explícita en el modelo permite controlar su efecto sobre avgscore. Esto puede ser importante porque el ingreso familiar promedio tiende a estar correlacionado con el gasto por estudiante, el cual suele estar determinado tanto por el impuesto sobre las propiedades inmuebles como por el impuesto local sobre la renta. En un análisis de regresión simple, *avginc* quedaría incluido en el término de error, que es posible que esté correlacionado con *expend*, lo que ocasionaría que en el modelo de dos variables el estimador MCO de β1 sea sesgado.

En los dos ejemplos anteriores se muestra cómo incluir en el modelo de regresión otros factores observables [*educ* en la ecuación (3.1) y *expend* en la ecuación (3.2)], además de la variable de principal interés. Un modelo con dos variables independientes puede expresarse en general como

[3.3] $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

donde

β_0 es el intercepto.

β_1 mide el cambio en y respecto a x_1, manteniendo constantes todos los demás factores.

β_2 mide el cambio en y respecto a x_2, manteniendo constantes todos los demás factores.

El análisis de regresión múltiple es útil también para generalizar relaciones funcionales entre variables. Por ejemplo, suponga que el consumo familiar (*cons*) sea una función cuadrática del ingreso familiar (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u, \quad [3.4]$$

donde *u* contiene otros factores que afectan el consumo. En este modelo, el consumo sólo depende de un factor observado, el ingreso, por lo que parece que puede tratarse en el marco de la regresión simple. Pero este modelo cae fuera de la regresión simple, porque contiene dos funciones del ingreso, *inc* e inc^2 (y por tanto tres parámetros: β_0, β_1 y β_2). Sin embargo, la función consumo puede expresarse de manera sencilla como un modelo de regresión con dos variables independientes haciendo $x_1 = inc$ y $x_2 = inc^2$.

De forma mecánica, no habrá ninguna diferencia al usar el método de mínimos cuadrados ordinarios (presentado en la sección 3.2) para estimar ecuaciones tan diferentes como la (3.1) y la (3.4). Cada una de ellas puede escribirse como la ecuación (3.3), que es lo único que interesa para los cálculos. Sin embargo, hay una diferencia importante en la interpretación de los parámetros. En la ecuación (3.1), β_1 es el efecto *cæteris paribus* de *educ* sobre *wage*. En la ecuación (3.4) no es esta la interpretación del parámetro β_1. En otras palabras, no tiene sentido medir el efecto de *inc* sobre *cons* cuando inc^2 se mantiene constante, porque si *inc* cambia, ¡también cambia inc^2! En lugar de esto, el cambio en consumo respecto al cambio en ingreso —la propensión marginal a consumir — se aproxima mediante

$$\Delta cons/\Delta inc \approx \beta_1 + 2\beta_2 inc.$$

Vean en el apéndice A el cálculo requerido para obtener esta ecuación. En otras palabras, el efecto marginal del ingreso sobre el consumo depende tanto de β_2 como de β_1 y del nivel de ingreso. *Este ejemplo muestra que, en cualquier aplicación particular, la definición de las variables independientes es crucial*. Pero para el desarrollo teórico de la regresión múltiple, no es necesario ser tan preciso acerca de estos detalles. Ejemplos como éste se estudiarán de manera más cabal en el capítulo 6.

En el modelo con dos variables independientes, el supuesto clave acerca de cómo está relacionado u con x_1 y x_2 es

$$E(u \mid x_1, x_2) = 0. \quad [3.5]$$

La interpretación de la condición (3.5) es similar a la del supuesto RLS.4 en el análisis de regresión lineal simple. Esta condición significa que, para cualesquiera valores de x_1 y x_2 en la población, el promedio del efecto de los factores no observables es igual a cero. Como en la regresión simple, la parte importante de este supuesto es que el valor esperado de *u* es el mismo para todas las combinaciones de x_1 y x_2; que este valor

común es cero no es ningún supuesto siempre que el intercepto βo se incluya en el modelo (vean la sección 2.1).

¿Cómo puede interpretarse el supuesto de media condicional cero en los ejemplos anteriores? En la ecuación (3.1), este supuesto es $E(u \mid educ, exper) = 0$. Esto significa que los otros factores que afectan *wage* no están relacionados en promedio con *educ* y *exper*. *Por tanto, si se piensa que la capacidad innata es parte de u, entonces se necesita que los niveles promedio de capacidad sean iguales para todas las combinaciones de educación y experiencia en la población trabajadora. Esto puede ser cierto o no, pero como se verá en la sección 3.3, hay que formular esta pregunta para determinar si el método de mínimos cuadrados ordinarios produce estimadores insesgados*.

El ejemplo en el que se mide el desempeño de un estudiante [ecuación (3.2)] es parecido a la ecuación del salario. El supuesto de media condicional cero es $E(u \mid expend, avginc) = 0$, lo cual significa que los otros factores que afectan las calificaciones —características de la escuela o del estudiante— no están, en promedio, relacionadas con el financiamiento por estudiante y con el ingreso familiar promedio. Aplicado a la función cuadrática de consumo en (3.4), el supuesto de media condicional cero tiene una interpretación un poco diferente. Expresada en forma literal, la ecuación (3.5) se convierte en $E(u \mid inc, inc^2) = 0$. Como cuando se conoce *inc*, también se conoce inc^2, resulta redundante incluir *inc2* en la esperanza: $E(u \mid inc, inc^2) = 0$ es lo mismo que $E(u \mid inc) = 0$. No hay problema si en la esperanza se coloca *inc* e inc^2 al establecer el supuesto, pero $E(u \mid inc) = 0$ es más concisa.

Modelo con k variables independientes

Una vez en el contexto de la regresión múltiple, no es necesario quedarse con sólo dos variables independientes. El análisis de regresión múltiple permite muchos factores observados que afecten a *y*. En el ejemplo del salario también pueden incluirse cantidad de capacitación laboral, años de antigüedad en el empleo actual, mediciones de la capacidad e incluso variables demográficas como cantidad de hermanos o educación de la madre. En el ejemplo del financiamiento de la escuela, otras variables pueden ser mediciones de la calidad de los maestros y tamaño de la escuela.

El *modelo general de regresión lineal múltiple* (también llamado *modelo de regresión múltiple*) poblacional puede expresarse como

[3.6] $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

donde

β_0 es el intercepto.
β_1 es el parámetro asociado con x_1.
β_2 es el parámetro asociado con x_2, y así sucesivamente.

Como hay *k* variables independientes y un intercepto, la ecuación (3.6) contiene k + 1 parámetros poblacionales (desconocidos). Por brevedad, a los parámetros distintos del intercepto se los llamará *parámetros de pendiente*, incluso aunque no siempre es esto lo que literalmente son. [Vean la ecuación (3.4), en donde ni β_1 ni β_2 son pendientes, pero juntos determinan la pendiente de la relación entre consumo e ingreso.]

En la regresión múltiple, la terminología es similar a la de la regresión simple y se presenta en la tabla 3.1. Como en la regresión simple, la variable *u* es el *término de error o perturbación*. Este término contiene los otros factores distintos de $x_1, x_2,\ldots, x_k$ que afectan a y. *No importa cuántas variables explicativas se incluyan en el modelo, siempre habrá factores que no se pueden incluir y todos ellos juntos están contenidos en u*.

Terminología de la regresión múltiple

y	$x_1, x_2, \ldots, x_k$
Variable dependiente	Variables independientes
Variable explicada	Variables explicativas
Variable de respuesta	Variables de control
Variable predicha	Variables predictoras
Regresando	Regresores

Tabla 3.1

Cuando se emplea el modelo general de regresión múltiple, hay que saber cómo interpretar los parámetros. En este capítulo y en los siguientes se obtendrá suficiente práctica, pero en este punto es útil recordar algunas cosas ya sabidas. Supongan que el *sueldo* (salary) de un director general o CEO está relacionado con las *ventas* de la empresa (sales) y su *antigüedad* en la organización (ceoten) mediante

[3.7] $$log\ (salary) = \beta_0 + \beta_1\ log\ (sales) + \beta_2\ ceoten + \beta_3\ ceoten^2 + u.$$

Esta ecuación encaja en el modelo de regresión múltiple (con k = 3) definiendo y = log(salary), x_1 = log(sales), x_2= ceoten y x_3 = ceoten^2. Como se sabe, por el capítulo 2, el parámetro β_1 es la elasticidad (*cæteris paribus*) del sueldo (salary) respecto a las ventas (sales). Si $\beta_3 = 0$, entonces $100\beta_2$ es aproximadamente el incremento porcentual *cæteris paribus* de *salary* cuando *ceoten* aumenta en un año. Cuando $\beta_3 \neq 0$, el efecto de *ceoten* sobre *salary* es más complicado. El tratamiento más general de modelos con términos cuadráticos se pospondrá hasta el capítulo 6.

La ecuación (3.7) proporciona un aviso importante acerca del análisis de regresión múltiple. *La palabra "lineal" en el modelo de regresión lineal múltiple* significa que la ecuación (3.6) es lineal en los parámetros, β_j. La ecuación (3.7) es un ejemplo de modelo de regresión múltiple que, aunque lineal en las β_j, es una relación no lineal entre *salary* y las variables *sales* y *ceoten*. En muchas aplicaciones de la regresión lineal múltiple hay relaciones no lineales entre las variables subyacentes.

El *supuesto clave* en el modelo general de regresión múltiple se establece con facilidad en términos de una esperanza condicional:

[3.8] $$E\ (u \mid x_1, x_2,\ldots, x_k) = 0.$$

Como mínimo, la ecuación (3.8) requiere que ninguno de los factores en el término de error no observado esté correlacionado con las variables explicativas. También significa que se ha entendido de manera correcta la relación funcional entre la variable explicada y las variables explicativas. Cualquier problema que cause que *u* esté correlacionada con cualquiera de las variables independientes hace que (3.8) no se satisfaga. En la sección 3.3 se mostrará que el supuesto (3.8) implica que los estimadores de MCO son insesgados y se obtendrá el sesgo que surge cuando una variable clave se omite de la ecuación. En los capítulos 15 y 16 se estudiarán otras razones que pueden hacer que (3.8) no se satisfaga y se mostrará qué puede hacerse en los casos en que no se satisface.

3.2 *Mecánica e interpretación de los mínimos cuadrados ordinarios (Selección)*

Ejemplo 3.1 Determinantes del promedio en la universidad

Las variables en la base de datos GPA1.RAW incluyen el promedio general de calificaciones en la universidad (*colGPA*), el promedio general de calificaciones en el bachillerato (*hsGPA*) y la puntuación en el examen de admisión a la universidad (*ACT*) para una muestra de 141 estudiantes de una universidad grande; los promedios generales de calificaciones tanto del bachillerato como de la universidad se dan en una escala de cuatro puntos. Para predecir el promedio general de calificaciones en la universidad, a partir del promedio general de calificaciones en el bachillerato y de la calificación en el examen de admisión se obtiene la siguiente línea de regresión de MCO:

[3.15] $$colGPA = 1.29 + .453\, hsGPA + .0094\, ACT.$$

¿Cómo se interpreta esta ecuación? Primero, el intercepto 1.29 es la predicción del promedio general de calificaciones en la universidad si *hsGPA* y *ACT* son ambos cero. Dado que ninguna persona que asista a la universidad tiene cero como promedio general de calificaciones de bachillerato ni cero en el examen de admisión a la universidad, el intercepto, en este caso, no tiene en sí ningún significado.

Estimaciones más interesantes son las de los coeficientes de pendiente de *hsGPA* y *ACT*. Como era de esperarse, existe una relación parcial positiva entre *colGPA* y *hsGPA*: con *ACT* constante, cada punto más en hsGPA se relaciona con .453 adicional en el promedio general de la universidad, es decir, casi medio punto. En otras palabras, si se eligen dos estudiantes, A y B, y éstos tienen la misma puntuación en el examen de admisión (*ACT*), pero el promedio general en el bachillerato del estudiante A es un punto superior al del estudiante B, entonces se predice que en la universidad el estudiante A tendrá un promedio general de calificaciones .453 más alto que el estudiante B. (Esto no dice nada acerca de dos personas reales, es sólo la mejor predicción.)

El signo de *ACT* implica que, si *hsGPA* permanece constante, un cambio de 10 puntos en el examen de admisión (*ACT*) —un cambio muy grande, ya que en la muestra la puntuación promedio es de 24 con una desviación estándar menor a tres— tendrá un efecto sobre *colGPA* de menos de una décima de punto. Este es un efecto pequeño que indica que, una vez que se ha tomado en cuenta el promedio general del bachillerato, la puntuación en el examen de admisión (*ACT*) no es un fuerte predictor del promedio general en la universidad. (Naturalmente, hay muchos otros factores que contribuyen al promedio general de calificaciones en la universidad, pero aquí nos concentramos en los

estadísticos disponibles para los estudiantes de bachillerato.) Más adelante, después de que se analice la inferencia estadística, se mostrará que el coeficiente de *ACT* no sólo es pequeño para fines prácticos, sino que es estadísticamente insignificante.

Centrándose en el análisis de regresión simple que relaciona *colGPA* sólo con *ACT* se obtiene

$$colGPA = 2.40 + .0271\, ACT\ ;$$

de manera que el coeficiente de *ACT* es casi el triple del estimado en (3.15). Pero esta ecuación no permite comparar dos personas con el mismo promedio general en el bachillerato; esta ecuación corresponde a otro experimento. Después se comentará más acerca de las diferencias entre la regresión múltiple y la simple.

Ejemplo 3.2 Ecuación para el salario por hora

Empleando las 526 observaciones sobre trabajadores en la base de datos WAGE1.RAW, las variables *educ* (años de educación), *exper* (años de experiencia en el mercado laboral) y *tenure* (años de antigüedad en el empleo actual) se incluyen en una ecuación para explicar *log(wage)*. La ecuación estimada es

[3.19] $$log(wage) = .284 + .092\, educ + .0041\, exper + .022\, tenure.$$

Como en el caso de la regresión simple, los coeficientes tienen una interpretación porcentual. Aquí la única diferencia es que también tienen una interpretación *cæteris paribus*. El coeficiente .092 significa que manteniendo exper y tenure constantes, se predice que un año más de educación incrementa *log(wage)* en .092, lo que se traduce en un aumento aproximado de 9.2% [100(.092)] en *wage*. *Es decir, si se toman dos personas con los mismos niveles de experiencia y antigüedad laboral, el coeficiente de educ es la diferencia proporcional con el salario predicho cuando en sus niveles de educación hay una diferencia de un año*.

El significado de "mantener todos los demás factores constantes" en la regresión múltiple

La interpretación del efecto parcial de los coeficientes de pendiente en el análisis de regresión múltiple puede causar cierta confusión, por lo que a continuación se presenta un análisis más amplio.

En el ejemplo 3.1, se observó que el coeficiente de ACT mide la diferencia que se predice para colGPA cuando hsGPA se mantiene constante. El poder del análisis de regresión múltiple es que proporciona esta interpretación *cæteris paribus* incluso cuando los datos no hayan sido recolectados de manera *cæteris paribus*. Al darle al coeficiente de ACT una interpretación de efecto parcial, puede parecer que se salió y se muestrearon personas con el mismo promedio general en el bachillerato pero con puntuaciones diferentes en el examen de admisión (ACT). Este no es el caso. Los datos son una muestra aleatoria tomada de una universidad grande: para obtener los datos no se pusieron restricciones sobre los valores muestrales de hsGPA o de ACT. Es muy raro que al obtener una muestra pueda uno darse el lujo de mantener constantes ciertas variables. Si se pudiera obtener una muestra de individuos con un mismo promedio general en el bachillerato, entonces se podría realizar un análisis de regresión simple relacionando

colGPA con ACT. *La regresión múltiple permite imitar esta situación sin restringir los valores de ninguna de las variables independientes*.

El poder del análisis de regresión múltiple es que permite hacer en un ambiente no experimental, lo que en las ciencias naturales puede hacerse con experimentos controlados de laboratorio: mantener constantes otros factores.

Cambiar de manera simultánea más de una variable independiente

Algunas veces se desea cambiar a la vez más de una variable independiente para determinar el efecto resultante sobre la variable dependiente. Esto es fácil de hacer usando la ecuación (3.17). Por ejemplo, en la ecuación (3.19) se puede obtener el efecto estimado sobre *wage* cuando una persona permanece un año más en una misma empresa: tanto *exper* (experiencia general en la fuerza laboral) como *tenure* (antigüedad en el empleo actual) aumentan en un año. El efecto total (manteniendo *educ* constante) es

$$\Delta \log(wage) = .0041\,\Delta exper + .022\,\Delta tenure = .0041 + .022 = .0261,$$

es decir, aproximadamente 2.6%. Dado que tanto *exper* como *tenure* aumentan un año, simplemente se suman los coeficientes de *exper* y *tenure* y se multiplica por 100 para convertir el efecto en un porcentaje.

3.3 *Valor esperado de los estimadores de MCO*

Ahora se verán las propiedades estadísticas del método de MCO para estimar los parámetros del modelo poblacional. En esta sección se obtienen los valores esperados de los estimadores de MCO. En particular, se establecen y se analizan cuatro supuestos, que son extensiones directas de los supuestos del modelo de regresión simple, bajo el cual los estimadores de MCO de los parámetros poblacionales son insesgados. Se obtiene también de manera explícita el sesgo de MCO cuando se omite una variable importante para la regresión.

Hay que recordar que las propiedades estadísticas no tienen nada que ver con la muestra de que se trate, sino con la propiedad de los estimadores cuando el muestreo aleatorio se hace repetidas veces. Así, las secciones 3.3, 3.4 y 3.5 son un poco abstractas. Aunque se dan ejemplos de la obtención del sesgo en modelos específicos, no tiene sentido hablar de las propiedades estadísticas de un conjunto de estimaciones obtenidas de una sola muestra.

El primer supuesto sólo define el modelo de regresión lineal múltiple (RLM).

Supuesto RLM.1 Lineal en los parámetros

El modelo poblacional puede expresarse como

[3.31] $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros (constantes) desconocidos de interés y u es un error aleatorio o término de perturbación no observable.

La ecuación (3.31) expresa formalmente el *modelo poblacional*, llamado algunas veces el *modelo verdadero*, para permitir la posibilidad de estimar un modelo que difiera de

(3.31). La característica clave es que este modelo es lineal en los parámetros $\beta_0, \beta_1, \ldots, \beta_k$. Como se sabe, (3.31) es bastante flexible porque tanto y como las variables independientes pueden ser funciones arbitrarias de las variables de interés, tales como logaritmos naturales y cuadrados [vean, por ejemplo, la ecuación (3.7)].

Supuesto RLM.2 Muestreo aleatorio

Se tiene una muestra aleatoria de n observaciones, $\{(x_{i1}, x_{i2},\ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, que sigue el modelo poblacional del supuesto RLM.1.

Algunas veces se necesita dar la ecuación de una determinada observación i: dada una observación obtenida de manera aleatoria de la población, se tiene

[3.32] $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i.$$

Recuerden que i se refiere a una observación y que el segundo subíndice de las x es el número de la variable. Por ejemplo, se puede escribir la ecuación del sueldo del director general o CEO para un determinado CEO i como

[3.33] $$log\,(salary_i) = \beta_0 + \beta_1\, log(sales_i) + \beta_2\, ceoten_i + \beta_3\, ceoten_i^2 + u_i\,.$$

El término u_i contiene los factores no observados del CEO i que afectan su sueldo. Para las aplicaciones, suele ser más fácil dar el modelo en forma poblacional, como en (3.31). Este modelo contiene menos desorden y hace énfasis en el hecho de que interesa estimar una relación poblacional.

A la luz del modelo (3.31), los estimadores de MCO $\beta^{\circ}_0, \beta^{\circ}_1, \ldots, \beta^{\circ}_k$ de la regresión de y sobre $x_1, \ldots, x_k$ se consideran como estimadores de $\beta_0, \beta_1, \ldots, \beta_k$. En la sección 3.2 se vio que, dada una muestra, MCO elige las estimaciones de intercepto y de las pendientes de manera que el promedio de los residuales sea cero y que la correlación muestral entre cada variable independiente y los residuales sea cero. Sin embargo, no se han dado las condiciones bajo las cuales, dada una muestra, las estimaciones de MCO están bien definidas. El supuesto siguiente llena esta brecha.

Supuesto RLM.3 No hay colinealidad perfecta

En la muestra (y por tanto en la población), ninguna de las variables independientes es constante y no hay ninguna relación *lineal* exacta entre las variables independientes.

El supuesto RLM.3 es más complicado que su contraparte para la regresión simple, porque ahora hay que considerar la relación entre todas las variables independientes. Si una variable independiente en (3.31) es una combinación lineal exacta de las otras variables independientes, entonces se dice que el modelo sufre de *colinealidad perfecta* y que no puede ser estimado por el método de MCO.

Es importante observar que el supuesto RLM.3 sí permite que las variables independientes estén correlacionadas; lo único que no permite es que estén perfectamente correlacionadas. *Si no se permitiera ninguna correlación entre las variables independientes, entonces la regresión múltiple sería de muy poca utilidad para el análisis econométrico*. Por ejemplo, en el modelo en el que se relacionan las puntuaciones de exámenes con los gastos en educación y el ingreso familiar promedio,

$$avgscore = \beta o + \beta_1 \, expend + \beta_2 \, avginc + u,$$

se espera que *expend* y *avginc* estén correlacionados: los distritos escolares en los que el ingreso familiar promedio es alto tienden a gastar más en educación por estudiante. De hecho, la principal motivación para incluir *avginc* en la ecuación es que se sospecha que está relacionado con *expend*, y por esto se desea mantenerlo constante en el análisis. El supuesto RLM.3 sólo descarta la correlación perfecta, en nuestra muestra, entre *expend* y *avginc*. Sería muy mala suerte obtener una muestra en la que los gastos por estudiante estuvieran correlacionados de manera perfecta con el ingreso familiar promedio. Pero una cierta correlación, quizá en una cantidad importante es esperada y en efecto permitida.

El caso más sencillo en que dos variables independientes pueden estar correlacionadas de manera perfecta es aquel en el que una variable sea un múltiplo constante de otra. *Esto puede ocurrir cuando el investigador, en forma inadvertida, coloca en una ecuación de regresión la misma variable medida en diferentes unidades*. Por ejemplo, al estimar la relación entre consumo e ingreso no tiene sentido incluir como variables independientes ingreso medido en dólares e ingreso medido en miles de dólares. Una de estas dos variables es redundante. ¿Qué sentido tendría mantener constante el ingreso medido en dólares y variar el ingreso medido en miles de dólares?

Como se sabe, entre los regresores *puede* haber diferentes funciones lineales de una misma variable. Por ejemplo, el modelo $cons = \beta_0 + \beta_1 \, inc + \beta_2 \, inc^2 + u$ no viola el supuesto RLM.3: aun cuando $x_2 = inc^2$ es una función exacta de $x_1 = inc$, inc^2 no es una función lineal exacta de *inc*. Incluir inc^2 en el modelo es una manera útil de generalizar la forma funcional, a diferencia de incluir el ingreso medido en dólares y en miles de dólares.

El sentido común indica no incluir en una misma ecuación de regresión la misma variable explicativa medida en diferentes unidades. Existen también situaciones más sutiles en las que una variable independiente puede ser múltiplo de otra. Suponga que se desea estimar una extensión de una función de consumo de elasticidad constante. Parecería natural especificar un modelo como el siguiente

[3.34] $$log(cons) = \beta_0 + \beta_1 \, log(inc) + \beta_2 \, log(inc^2) + u,$$

donde $x_1 = log(inc)$ y $x_2 = log(inc^2)$. Al utilizar las propiedades básicas de los logaritmos naturales (véase apéndice A), $log(inc^2) = 2*log(inc)$. Es decir, $x_2 = 2x_1$, y esto es válido para todas las observaciones de la muestra. Esto viola el supuesto RLM.3. En lugar de esto hay que incluir $[log(inc)]^2$, y no $log(inc^2)$, junto con $log(inc)$. Esta es una extensión razonable del modelo de elasticidad constante, y en el capítulo 6 se verá cómo interpretar tales modelos.

Los ejemplos anteriores muestran que el supuesto RLM.3 puede no satisfacerse si se descuida especificar el modelo. El supuesto RLM.3 tampoco se satisface si el tamaño de la muestra, n, es demasiado pequeño en relación con el número de parámetros que se estiman. En general, en el modelo de regresión de la ecuación (3.31), hay $k + 1$ parámetros y RLM.3 no se satisface si $n < k + 1$. De manera intuitiva, esto es razonable: para estimar $k + 1$ parámetros, se necesitan por lo menos $k + 1$ observaciones. Es claro que

es mejor tener tantas observaciones como sea posible, cosa que se notará al ver el cálculo de la varianza en la sección 3.4.

Si el modelo se ha especificado con cuidado y $n \geq k + 1$, el supuesto RLM.3 puede no satisfacerse en casos raros debido a mala suerte al recolectar la muestra. Por ejemplo, en una ecuación para el salario en que la educación y experiencia sean las variables, es posible que se obtenga una muestra aleatoria en la que cada individuo tenga exactamente el doble de años de educación que años de experiencia. Esta situación hará que el supuesto RLM.3 no se satisfaga, pero esta situación es muy poco probable a menos que se tenga un tamaño de muestra en extremo pequeño.

El último supuesto, y el más importante, para el insesgamiento es una extensión directa del supuesto RLS.4.

Supuesto RLM.4 Media condicional cero

El valor esperado del error u, dados los valores de las variables independientes, es cero. En otras palabras

[3.36] $E(u \mid x_1, x_2, \ldots, x_k) = 0.$

El supuesto RLM.4 puede no satisfacerse si en la ecuación (3.31) la relación funcional entre las variables explicada y explicativas está mal especificada: por ejemplo, si se olvida incluir el término cuadrático inc^2 en la función de consumo $cons=\beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ al estimar el modelo. Otra especificación errónea de la forma funcional se presenta cuando se emplea una variable en su nivel original siendo que en el modelo poblacional se emplea el logaritmo de la variable, o viceversa. Por ejemplo, si el verdadero modelo tiene log(wage) como variable dependiente, pero en el análisis de regresión se usa wage como variable dependiente, entonces los estimadores estarán sesgados. De manera intuitiva esto es bastante claro. En el capítulo 9 se analizarán maneras de detectar formas funcionales mal especificadas.

Omitir un factor importante correlacionado con cualquiera de las $x_1, x_2, \ldots, x_k$ ocasiona también que el supuesto RLM.4 no se satisfaga. En el análisis de regresión múltiple pueden incluirse muchos factores entre las variables explicativas y es menos probable que las variables omitidas sean un problema en comparación con el análisis de regresión simple. De cualquier manera en toda aplicación, hay factores que, debido a las limitaciones de los datos o a ignorancia, no pueden incluirse. Si se cree que estos factores deben controlarse y están correlacionados con una o más de las variables independientes, se violará el supuesto RLM.4. Más adelante se verán estos sesgos.

Hay otras maneras en las que u puede estar correlacionada con una variable explicativa. En el capítulo 15 se analizará el problema del error de medición en una variable explicativa. En el capítulo 16 se verá el problema, conceptualmente más complicado, en el que *una o más de las variables explicativas se determina conjuntamente con y*. El estudio de estos problemas se pospondrá hasta que se tenga una comprensión más firme del análisis de regresión múltiple bajo un conjunto ideal de supuestos.

Cuando se satisface el supuesto RLM.4 se suele decir que se tienen variables explicativas exógenas. Si por alguna razón x_j está correlacionada con u, entonces se dice que x_j es una variable explicativa *endógena*. Los términos “exógena” y “endógena” son origina-

rios del análisis de ecuaciones simultáneas (vea el capítulo 16), pero *el término "variable explicativa endógena" ha evolucionado para abarcar cualquier caso en el que una variable explicativa esté correlacionada con el término del error*.

Teorema 3.1 Insesgamiento de los estimadores de MCO

Bajo los supuestos RLM.1 a RLM.4,

[3.37] $E(\beta^{\circ}_j) = \beta_j, j = 0, 1, \ldots, k,$

para cualquier valor del parámetro poblacional β_j. En otras palabras, los estimadores de MCO son *estimadores insesgados de los parámetros poblacionales*.

En los ejemplos empíricos anteriores, el supuesto RLM.3 se ha satisfecho (porque se han podido calcular las estimaciones de MCO). Además, en su mayoría, las muestras han sido tomadas de manera aleatoria de una población bien definida. Si se cree que los modelos especificados son correctos bajo el supuesto clave RLM.4, entonces se puede concluir que en estos ejemplos el modelo de MCO es insesgado.

Como estamos llegando a un punto en el que se puede usar la regresión múltiple en trabajos empíricos serios, es útil recordar el *significado del insesgamiento*. Uno se siente tentado, en ejemplos como el de la ecuación del salario en (3.19), a decir algo así como "9.2% es una estimación insesgada del rendimiento de la educación". Como se sabe, una estimación no puede ser insesgada: una estimación es un número fijo, obtenido a partir de una determinada muestra que, por lo general, no es igual al parámetro poblacional. Cuando se dice que los estimadores de MCO son insesgados bajo los supuestos RLM.1 a RLM.4, en realidad se quiere decir que el procedimiento mediante el cual se obtienen las estimaciones de MCO es insesgado cuando se le considera aplicado a todas las muestras aleatorias posibles. Se espera haber obtenido una muestra que dé una estimación cercana al valor poblacional pero, por desgracia, esto no puede asegurarse. Lo que se asegura es que no hay razón para creer ni que sea probablemente muy grande ni que sea probablemente muy pequeño.

3.4 *Varianza de los estimadores de MCO*

Ahora se obtendrá la varianza de los estimadores de MCO de manera que, además de conocer la tendencia central de los β°_j también se tendrá una medida de dispersión en su distribución de muestreo. Antes de hallar la varianza, se agregará un supuesto de homocedasticidad como en el capítulo 2. Esto se hace por dos razones. Primero, imponiendo el supuesto de varianza constante del error, se simplifican las fórmulas. Segundo, en la sección 3.5 se verá que si se agrega el supuesto de homocedasticidad, el método de MCO tiene una importante propiedad de *eficiencia*.

Supuesto RLM.5 Homocedasticidad

Dado cualquier valor de las variables explicativas, el error u tiene la misma varianza. En otras palabras, $Var(u \mid x_1, \ldots, x_k) = \sigma^2$.

El supuesto RLM.5 significa que la varianza en el término del error, u, condicional en las variables explicativas, es la misma para todas las combinaciones de valores de las

variables explicativas. Si este supuesto no se satisface, entonces el modelo muestra *heterocedasticidad*, como ocurre en el caso de dos variables.

En la ecuación

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u,$$

la homocedasticidad requiere que la varianza del error no observado u no dependa de la educación, la experiencia o la antigüedad. Es decir,

$$Var(u \mid educ, exper, tenure) = \sigma^2.$$

Si esta varianza cambia de acuerdo con alguna de las tres variables explicativas, se tiene heterocedasticidad.

A los supuestos RLM.1 a RLM.5 se los conoce como ***supuestos de Gauss-Markov*** (para regresiones de corte transversal). Estos supuestos, como se han dado hasta ahora, sólo son adecuados para el análisis de corte transversal con muestreo aleatorio. Como se verá, los supuestos de Gauss-Markov para el análisis de series de tiempo y para otras situaciones como el análisis de datos de panel, son más difíciles de expresar, aunque hay muchas semejanzas.

Teorema 3.2 Varianza de muestreo de los estimadores de pendiente de MCO

Bajo los supuestos RLM.1 a RLM.5, condicionales en los valores muestrales de las variables independientes,

[3.51] $$Var(\hat{\beta}_j) = \sigma^2 / [STC_j (1 - R_j^2)]$$

para $j = 1, 2, \ldots, k$, donde $STC_j = \sum_{i=1}^{n} (x_{ij} - x^m_j)^2$ es la variación muestral total en x_j y R^2_j es la R cuadrada de regresión de x_j sobre todas las otras variables independientes (e incluyendo un intercepto).

Antes de estudiar con más detalle la ecuación (3.51), es importante saber que para obtener esta fórmula se usan todos los supuestos de Gauss-Markov. Mientras que el supuesto de homocedasticidad no se necesitó para concluir que los estimadores de MCO son insesgados, sí se necesita para demostrar la ecuación (3.51).

La magnitud de $Var(\hat{\beta}_j)$ tiene importancia práctica. Una varianza grande significa un estimador menos preciso y esto se traduce en intervalos de confianza grandes y pruebas de hipótesis menos exactas (como se verá luego). En la sección siguiente se analizan los elementos que comprende (3.51).

Los componentes de las varianzas de los estimadores de MCO: multicolinealidad

La ecuación 3.51 muestra que la varianza de $\hat{\beta}_j$ depende de tres factores: σ^2, STC_j y R_j^2. Recuerden que el subíndice j denota una de las variables independientes (por ejemplo,

educación o tasa de pobreza). A continuación se considerarán cada uno de los factores que afectan $Var\ (\beta^{\circ}_j)$.

La varianza del error, σ^2

De acuerdo con la ecuación (3.51), una σ^2 más grande significa varianzas más grandes para los estimadores de MCO. Esto no es nada sorprendente: más "ruido" en la ecuación (una σ^2 más grande) dificulta más estimar el efecto parcial de cualquier variable independiente sobre y, y esto se refleja en varianzas más grandes para los estimadores de pendiente de MCO. Como σ^2 es una característica de la población, no tiene nada que ver con el tamaño de la muestra. El único componente de (3.51) que es desconocido es σ^2. *Más adelante se verá cómo obtener un estimador insesgado* de σ^2.

Dada una variable dependiente y sólo hay, en realidad, una manera de reducir la varianza del error: agregar más variables explicativas a la ecuación (*extraer algunos factores del término del error*). Por desgracia, no siempre es posible hallar factores adicionales justificados que afecten a y.

La variación muestral total en x_j, STC_j

De acuerdo con la ecuación (3.51) se observa que cuanto mayor sea la variación total en x_j, menor será $Var\ (\beta^{\circ}_j)$. Por tanto, manteniendo constante todo lo demás, para estimar β_j se prefiere tener tanta variación muestral en x_j como sea posible. Esto ya se descubrió en el capítulo 2 en el caso de la regresión simple. Aunque es difícil que se puedan elegir los valores muestrales de las variables independientes, *hay una manera de aumentar la variación muestral en cada una de las variables independientes: aumentar el tamaño de la muestra*. En efecto, al muestrear de manera aleatoria una población, STC_j aumenta sin límite a medida que la muestra se hace más grande. Este es el componente de la varianza que depende sistemáticamente del tamaño de la muestra.

Si STC_j es pequeño, $Var\ (\beta^{\circ}_j)$ puede volverse muy grande, pero una STC_j pequeña no viola el supuesto RLM.3. Técnicamente, a medida que STC_j se aproxima a cero, $Var(\beta^{\circ}_j)$ se aproxima a infinito. El caso extremo en el que no hay variación muestral en x_j, STC_j=0 no es permitido por el supuesto RLM.3.

Relaciones lineales entre las variables independientes, R_j^2

En la ecuación (3.51), el término R_j^2 es tal vez el más difícil de entender de los tres. Este término no aparece en el análisis de regresión simple porque en tales casos sólo hay una variable independiente. Es importante ver que esta R-cuadrada es distinta de la R-cuadrada de la regresión de y sobre $x_1, x_2,\ldots, x_k$: esta R_j^2 se obtiene de una regresión en la que sólo intervienen las variables independientes del modelo original y donde x_j interviene como si fuera una variable dependiente.

Consideren primero el caso $k = 2$: $y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Entonces, $Var\ (\beta^{\circ}_1) = \sigma^2/[STC_1(1 - R_1^2)]$, donde R_1^2 es la R-cuadrada de la regresión simple de x_1 sobre x_2 (y, como siempre, un intercepto). Como la R-cuadrada mide la bondad de ajuste, un valor de R_1^2 cercano a uno indica que x_2 explica gran parte de la variación de x_1 en la muestra. Esto significa que x_1 y x_2 están fuertemente correlacionadas.

A medida que R_1^2 se aproxima a uno, $Var\ (\beta^\circ_1)$ se hace cada vez más grande. Por tanto, un alto grado de relación lineal entre x_1 y x_2 puede conducir a varianzas grandes en los estimadores de pendiente de MCO. (Un argumento similar aplica a $Var\ (\beta^\circ_2)$.) En la figura 3.1 se muestra la relación entre $Var\ (\beta^\circ_1)$ y la R-cuadrada de la regresión de x_1 sobre x_2.

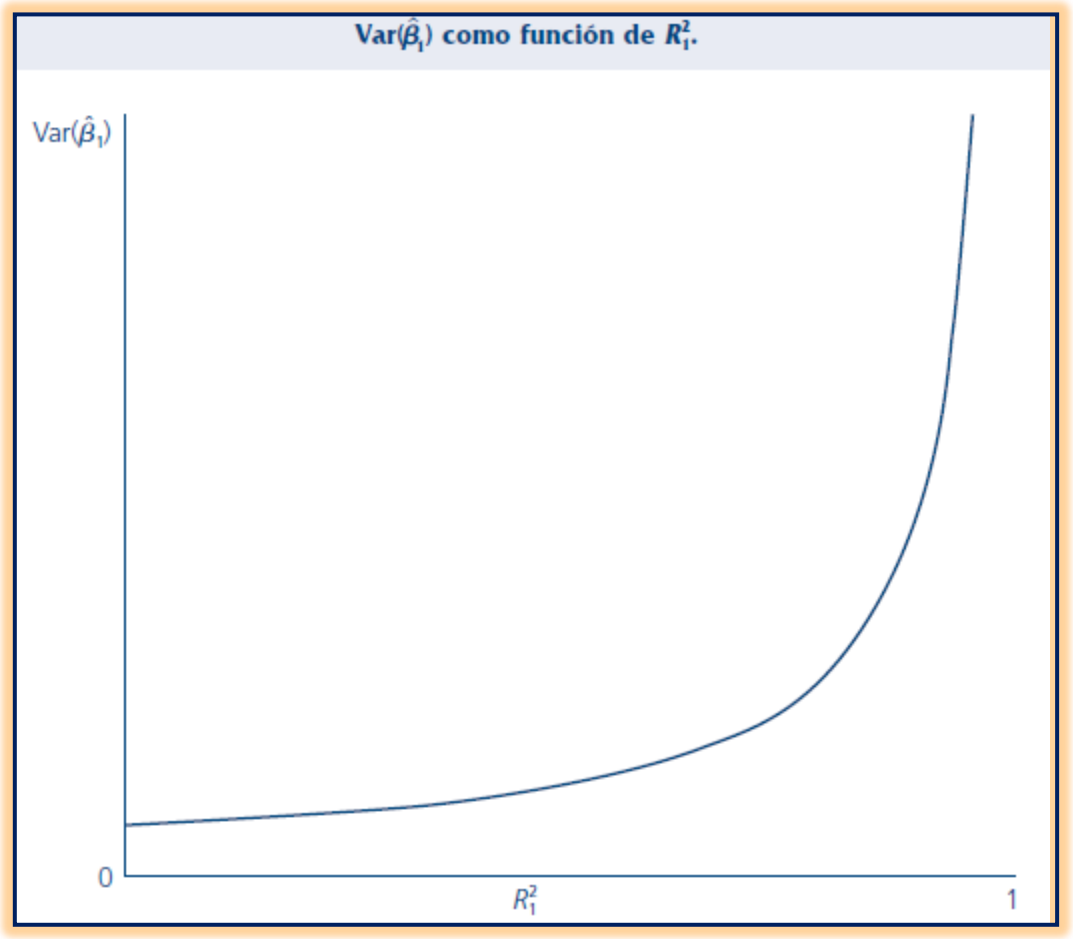

Figura 3.1

En el caso general, R_j^2 es la proporción de la variación total en x_j que puede ser explicada por las otras variables independientes que aparecen en la ecuación. Para σ^2 y STC_j dadas, la menor $Var\ (\beta^\circ_j)$ se obtiene cuando $R_j^2 = 0$, lo cual ocurre si y sólo si, x_j tiene correlación muestral cero con cada una de las otras variables independientes. Este es el mejor caso para estimar β_j, pero rara vez sucede.

El otro caso extremo, $R_j^2 = 1$, queda excluido por el supuesto RLM.3, porque esto significa que, en la muestra, x_j es una *combinación lineal perfecta de algunas de las demás variables independientes de la regresión*. Un caso más interesante es cuando el valor de R_j^2 es "cercano" a uno. De acuerdo con la ecuación (3.51) y con la figura 3.1, se ve que esto puede ocasionar que $Var\ (\beta^\circ_j)$ sea grande: $Var\ (\beta^\circ_j) \rightarrow \infty$ a medida que $R_j^2 \rightarrow 1$. A una correlación fuerte (pero no perfecta) entre dos o más variables independientes se la llama *multicolinealidad*.

Antes de analizar de manera más amplia el problema de la multicolinealidad es muy importante tener clara una cosa: el caso en que R_j^2 es cercana a uno *no viola* el supuesto RLM.3.

Como la multicolinealidad no viola ninguno de los supuestos, el "problema" de la multicolinealidad no está, en realidad, bien definido. Cuando se dice que la multicolinealidad surge al estimar β_j cuando R_j^2 es "cercana" a uno, "cercana" se pone entre comillas porque no hay un número absoluto que se pueda citar para concluir que la multicolinealidad es un problema. Por ejemplo, $R_j^2 = .9$ significa que 90% de la variación muestral en x_j puede ser explicada por las demás variables independientes del modelo de regresión. Sin duda, esto significa que x_j tiene una fuerte relación lineal con las demás variables independientes. Pero que esto se traduzca en que $Var\ (\beta^o_j)$ sea demasiado grande para ser útil depende de las magnitudes de σ^2 y de STC_j. Como se verá al hablar sobre inferencia estadística, lo que al final importa es qué tan grande es β^o_j en relación con su desviación estándar.

Aunque el problema de la multicolinealidad no está bien definido, una cosa está clara: permaneciendo todo lo demás constante, para estimar β_j, lo mejor es tener poca correlación entre x_j y las demás variables independientes. Esta observación suele conducir a la discusión de *cómo "resolver" el problema de multicolinealidad*. En las ciencias sociales, donde por lo común se es recolector pasivo de los datos, no hay otra manera de reducir la varianza de los estimadores insesgados que *recolectar más datos*. Dado un conjunto de datos, uno puede tratar de eliminar otras variables independientes del modelo con objeto de reducir la multicolinealidad. Por desgracia, eliminar una variable que pertenece al modelo poblacional puede llevar a sesgo, como se vio en la sección 3.3 (no incluida aquí).[2]

Estimación de σ^2: errores estándar de los estimadores de MCO

Ahora se mostrará cómo elegir un estimador insesgado de σ^2, el cual permitirá después obtener estimadores insesgados de $Var\ (\beta^o_j)$.

Como $\sigma^2 = E(u^2)$, un estimador "insesgado" de σ^2 es el promedio muestral de los errores cuadrados: $n^{-1} \sum_{i=1}^{n} u_i^2$. Por desgracia, éste no es un verdadero estimador porque los u_i no se pueden observar. Sin embargo, recuerden que es posible expresar los errores como $u_i = y_i - \beta_o - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}$, y la razón por la que no se pueden observar los u_i es que no se conocen los β_j. Cuando se sustituyen los β_j por sus estimadores de MCO, se obtienen los residuales de MCO:

$$u^o_i = y_i - \beta^o_o - \beta^o_1 x_{i1} - \beta^o_2 x_{i2} - \ldots - \beta^o_k x_{ik}.$$

Parece natural estimar σ^2 sustituyendo las u_i por las u^o_i. El estimador insesgado de σ^2 en el caso general de la regresión múltiple es

[3.56] $$\sigma^{o2} = (\sum_{i=1}^{n} u^o_i)/(n-k-1) = SRC/(n-k-1).$$

Este estimador ya se encontró en el caso k = 1 de la regresión simple.

---

[2] Para comentarios adicionales sobre el problema de multicolinealidad, sugiero acudir al texto de Wooldridge, capítulo 3.

El término $n - k - 1$ en (3.56) son los *grados de libertad* (gl) para el problema general de MCO con n observaciones y k variables independientes. Como en un modelo de regresión con k variables independientes y un intercepto hay $k + 1$ parámetros, se puede escribir

[3.57] $gl = n - (k+1) =$ *(número de observaciones) - (cantidad de parámetros estimados).*

La manera más sencilla de calcular los grados de libertad en una determinada aplicación es contar la cantidad de parámetros, incluyendo al intercepto, y restar esta cantidad del número de observaciones. (En el raro caso de que no se estime la intersección, la cantidad de parámetros disminuyen uno.)

Técnicamente, la división por $n - k - 1$ en (3.56) se debe a que el valor esperado de la suma de los residuales cuadrados es $E(SRC) = (n - k - 1)\sigma^2$. Se resume esto en el siguiente teorema.

Teorema 3.3 Estimación insesgada de σ^2

Bajo los supuestos RLM.1 a RLM.5 de Gauss-Markov, $E(\hat{\sigma}^2) = \sigma^2$.

A la raíz cuadrada positiva de $\hat{\sigma}^2$, que se denota $\hat{\sigma}$, se lo llama *error estándar de la regresión* (EER). El EER es un estimador de la desviación estándar del término de error. Los paquetes de software que corren regresiones suelen dar esta estimación, aunque le dan distintos nombres. (Además de EER, también se lo llama error estándar de la estimación y raíz cuadrática medio del error).

Observen que cuando se agrega otra variable independiente a la regresión (para una muestra dada) $\hat{\sigma}$ puede aumentar o disminuir. Esto se debe a que, aunque SRC debe disminuir cuando se agrega otra variable explicativa, los grados de libertad también disminuyen en uno. Como SRC está en el numerador y gl en el denominador, no se puede decir de antemano cuál será el efecto que domine.

En el capítulo 4, para construir intervalos de confianza y realizar pruebas, se necesitará estimar la desviación estándar de $\hat{\beta}_j$, que es la raíz cuadrada de la varianza:

$$de(\hat{\beta}_j) = \sigma/[STC_j(1-R_j^2)]^{1/2}$$

Como σ no se conoce, se sustituye por su estimador, $\hat{\sigma}$. Esto da el error estándar de $\hat{\beta}_j$:

[3.58] $$ee(\hat{\beta}_j) = \hat{\sigma}/[STC_j(1-R_j^2)]^{1/2}.$$

Al igual que las estimaciones de MCO, los errores estándar pueden obtenerse de cualquier muestra. Como $ee(\hat{\beta}_j)$ depende de $\hat{\sigma}$, el error estándar tiene una distribución de muestreo, que será un tema importante en el capítulo 4.

Hay que resaltar un punto acerca de los errores estándar. Como (3.58) se obtiene directamente de la fórmula de la varianza en (3.51) y dado que esta última se apoya en el supuesto de homocedasticidad RLM.5, se sigue que la fórmula del error estándar en (3.58) no es un estimador válido de $de(\hat{\beta}_j)$ si los errores muestran heterocedasticidad. Por tanto, mientras que la presencia de heterocedasticidad no causa sesgo en las $\hat{\beta}_j$, sí conduce a un sesgo en la fórmula usual para $Var(\hat{\beta}_j)$, lo cual invalida los errores

estándar. Esto es importante porque los paquetes para regresión, si no se les indica otra cosa, calculan (3.58) como el error estándar predeterminado de cada coeficiente (con una representación un poco diferente para el intercepto). Si se sospecha de heterocedasticidad, entonces los errores estándar "usuales" de MCO no son válidos y se deberán tomar medidas para corregir el problema. En el capítulo 8 se verán los métodos para el problema de la heterocedasticidad.

3.5 *Eficiencia de MCO: el teorema de Gauss-Markov*

En esta sección se enuncia y analiza el importante teorema de Gauss-Markov, el cual justifica el uso del método de MCO en lugar de otros diversos estimadores. Ya se conoce una justificación para el método de MCO: bajo los supuestos RLM.1 a RLM.4, el método de MCO es insesgado. Sin embargo, bajo estos supuestos hay muchos estimadores insesgados de las β_j. ¿Hay otros estimadores cuyas varianzas sean menores que las de los estimadores de MCO?

Si se limita la clase de los posibles estimadores apropiados, entonces se puede demostrar que el método de MCO es el mejor dentro de su clase. En concreto, se argumentará que, bajo los supuestos RLM.1 a RLM.5, el estimador de MCO β°_j es el *mejor estimador lineal insesgado (MELI)* para β_j. Para enunciar el teorema es necesario entender cada componente del acrónimo "MELI". Primero, ya se sabe qué es un estimador: es una regla que puede aplicarse a cualquier muestra de datos para obtener una estimación. Ya se sabe qué es un estimador insesgado: en el contexto presente, un estimador de β_j, por ejemplo, $\tilde{\beta}_j$, es un estimador insesgado de β_j si $E(\tilde{\beta}_j) = \beta_j$ para toda $\beta_0, \beta_1, \ldots, \beta_k$.

¿Qué significa el término "lineal"? En el presente contexto, un estimador $\tilde{\beta}_j$ de β_j es *lineal* si, y sólo si, se puede expresar como una función lineal de los datos de la variable dependiente:

$$[3.59] \qquad \tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i,$$

donde cada w_{ij} puede ser una función de los valores muestrales de todas las variables independientes. Como se puede ver de acuerdo con la ecuación (3.22), los estimadores de MCO son lineales.

Por último, ¿cómo se define "mejor"? En el presente teorema, *mejor* se define como menor varianza. Dados dos estimadores insesgados, es lógico preferir el que tenga menor varianza (véase el apéndice C).

Ahora, sean $\beta^{\circ}_0, \beta^{\circ}_1, \ldots, \beta^{\circ}_k$ los estimadores de MCO del modelo (3.31) bajo los supuestos RLM.1 a RLM.5. El teorema de Gauss-Markov dice que: dado cualquier estimador $\tilde{\beta}_j$ que sea lineal e insesgado, $Var(\tilde{\beta}_j) \geq Var(\beta^{\circ}_j)$ y esta desigualdad es, por lo general, estricta. Es decir, en la clase de los estimadores lineales insesgados, los estimadores de MCO tienen la mínima varianza (bajo los cinco supuestos de Gauss-Markov). En realidad, el teorema dice más aún. Si se quiere estimar una función lineal de los β_j, entonces la correspondiente combinación lineal de los estimadores de MCO proporciona la menor varianza entre todos los estimadores lineales insesgados. Se concluye con un teorema que se demuestra en el apéndice 3A.

Teorema 3.4 Teorema de Gauss-Markov

Bajo los supuestos RLM.1 a RLM.5, $\beta^{o}_{o}, \beta^{o}_{1},..., \beta^{o}_{k}$ son los mejores estimadores lineales insesgados (MELI) de $\beta_o, \beta_1,..., \beta_k$, respectivamente.

A este teorema se debe que los supuestos RLM.1 a RLM.5 se conozcan como los supuestos de Gauss-Markov (en el análisis de corte transversal).

La importancia del teorema de Gauss-Markov es que, si el conjunto estándar de supuestos se satisface, no es necesario buscar otros estimadores insesgados de la forma (3.59): ninguno será mejor que los estimadores de MCO. Esto es equivalente a decir que para cualquier otro estimador que sea lineal e insesgado, su varianza será por lo menos tan grande como la varianza de los estimadores de MCO; no es necesario hacer ningún cálculo para saber esto.

Para los propósitos presentes, el teorema 3.4 justifica el uso del método de MCO para estimar los modelos de regresión múltiple. Si no se satisface alguno de los supuestos de Gauss-Markov, entonces este teorema no es válido. Se sabe ya que si el supuesto de media condicional cero no se satisface (supuesto RLM.4), esto ocasiona que los estimadores de MCO sean sesgados, con lo que el teorema 3.4 ya no es válido. También se sabe ya que la heterocedasticidad (insatisfacción del supuesto RLM.5) no ocasiona sesgo. *Sin embargo, en presencia de heterocedasticidad, los estimadores de MCO ya no son los de menor varianza entre los estimadores lineales insesgados*. En el capítulo 8 se analiza un estimador que perfecciona los estimadores de MCO en presencia de heterocedasticidad.