

El modelo de regresión lineal simple

Alfonso Novales
Departamento de Economía Cuantitativa
Universidad Complutense

Septiembre 2008

Contents

1	El modelo de regresión lineal	1
1.1	El modelo de regresión lineal simple.	3
1.2	Componentes del modelo de regresión	4
1.3	Supuestos del modelo de regresión lineal	8
2	El estimador de Mínimos Cuadrados Ordinarios	11
2.1	Esperanza matemática	16
2.2	Matriz de covarianzas	19
3	El modelo de regresión lineal en desviaciones respecto de la media	21
4	Estimación de la varianza del término de error o perturbación aleatoria del modelo	22
5	Eficiencia	24
6	Propiedades adicionales del coeficiente de determinación	28
6.1	Expresión alternativa:	28
6.2	Relación con el coeficiente de correlación lineal en un modelo de regresión lineal simple:	28

1 El modelo de regresión lineal

El objeto básico de la Econometría consiste en especificar y estimar un modelo de relación entre las variables económicas relativas a una determinada cuestión conceptual. Por ejemplo, para conocer en profundidad el comportamiento del consumo privado agregado de un país, será preciso especificar y estimar un modelo de relación entre observaciones temporales de consumo privado y renta disponible. De modo similar, para analizar si la expansión monetaria

en un país ha sido inflacionista, será preciso especificar y estimar un modelo de relación entre las tasas de inflación y las tasas de crecimiento históricas de algún agregado monetario. En su forma más general y, por tanto, más abstracta, tal modelo de relación puede representarse como:

$$Y = f(X_1, X_2, X_3, \dots, X_k; \beta)$$

donde Y es la variable cuyo comportamiento se pretende explicar, y X_1, X_2, \dots, X_k son las distintas variables que se suponen potencialmente relevantes como factores explicativos de la primera. El vector denota una lista de parámetros que recogen la magnitud con que las variaciones en los valores de las variables X_i se transmiten a variaciones en la variable Y .

Vamos a limitarnos aquí al estudio de modelos de relación o modelos de regresión lineales, es decir, del tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k$$

en el que resulta evidente que los parámetros transmiten directamente efectos inducidos por los valores de las variables X_i sobre la variable Y , que se pretende explicar.

La estimación de tales relaciones se efectúa a partir de información muestral acerca de los valores tomados por Y, X_1, X_2, \dots, X_k , y trata de cuantificar la magnitud de la dependencia entre ellas.

Con objeto de ganar precisión y aprender más acerca del proceso de relación entre las variables consideradas queremos evaluar críticamente la validez de las hipótesis propuestas por la Teoría Económica acerca de las relaciones estimadas que, en ocasiones, consistirán en si determinada variable explicativa entra o no en la relación que se analiza, o si aparece con un determinado coeficiente, por ejemplo, 1 ó -1. Ejemplos del primer tipo son las cuestiones:

- 1) ¿Influye el déficit sobre los tipos de interés?
 - 2) ¿Afecta el precio de la competencia a la demanda de nuestro producto?
- mientras que ejemplos del segundo tipo son:
- 3) ¿Es el crecimiento monetario neutral, es decir, incide con coeficiente unitario sobre la inflación?
 - 4) ¿Tiene la demanda de nuestro producto elasticidad-precio unitaria? es decir, ¿el efecto de un aumento de un 10% en el precio es una caída del 10% en la demanda?

Estos son problemas de inferencia estadística, similares a los que resolvimos para contrastar hipótesis acerca de la esperanza o la varianza, desconocidas, de una determinada distribución de probabilidad. Por último, especialmente en cuestiones macroeconómicas, estaremos interesados en efectuar un ejercicio de seguimiento coyuntural y de previsión de las variables analizadas. Todo ello puede realizarse de modo riguroso mediante la utilización de procedimientos econométricos que vamos a estudiar en éste y en los dos próximos capítulos.

Así, mediante métodos econométricos, el analista económico puede tratar de responder a preguntas como:

- 1) ¿cuáles son los determinantes de la tasa de inflación?

2) sobre la base de la información histórica disponible, ¿cuál es la importancia cuantitativa de cada uno de dichos determinantes?

3) ¿podemos contrastar algunas de las implicaciones de la Teoría Económica acerca del efecto que variables como el crecimiento monetario tienen sobre la tasa de inflación?

4) ¿qué sugiere el modelo que hemos estimado para la tasa de inflación acerca del comportamiento de esta variable durante el próximo año?

Es crucial que el analista económico:

a) comience delimitando muy claramente la cuestión teórica que va a ser el centro de su ejercicio empírico,

b) a continuación, debe tratar de identificar cuál es la variable cuyo comportamiento pretende explicar, y cuáles son sus determinantes potenciales. Denominamos a este proceso especificación de un modelo de relación entre variables económicas. Como parte del proceso de especificación, el investigador toma posición acerca de qué variable influye sobre cuál, es decir, propone una relación causal. A diferencia del análisis que pudo efectuarse mediante un coeficiente de correlación, que no descansa en una determinada dirección en la relación entre dos variables, un análisis de regresión en Econometría supone que una variable X influye sobre otra variable Y , y no al revés;

c) luego, el analista debe escoger cuidadosamente la información estadística relevante para cuantificar tal relación, y

d) debe proceder a su cuantificación, es decir, debe estimar los parámetros desconocidos que aparecen en la relación antes especificada;

e) por último, utilizará el modelo de relación estimado, ya sea a efectos de contrastación de algún supuesto teórico, mediante un proceso de inferencia, o como elemento de análisis y seguimiento de la variable cuyo comportamiento escogió explicar.

1.1 El modelo de regresión lineal simple.

Vamos a limitarnos inicialmente al estudio del denominado modelo de regresión lineal simple, que considera una sola variable explicativa X :

$$Y = \beta_0 + \beta_1 X \quad (1)$$

En aplicaciones prácticas disponemos de una muestra de observaciones de ambas variables, y el modelo anterior sugiere que la relación entre las dos variables se satisface para cada una de las observaciones correspondientes. En algunas ocasiones especificaremos modelos de relación como (1) con el objeto de estimar el comportamiento de determinados agentes económicos. Un ejemplo importante consiste en entender la evolución del consumo agregado del sector privado de una economía real. En algunos casos se tratará de una muestra de datos temporales, y tendremos una relación del tipo (1) para cada instante de tiempo. Para ello, consideraríamos el modelo:

$$C_t = \beta_0 + \beta_1 Y_t, \quad t = 1, 2, \dots, T$$

donde Y_t denota el PIB del país, o la renta disponible del sector privado (renta total, menos impuestos, más transferencias), según el alcance que se quiera dar al análisis. Los subíndices t hacen clara referencia al hecho de que éste será un modelo a estimar con datos de series temporales. El coeficiente β_1 indica la variación que experimenta el consumo privado del país al variar, a lo largo del ciclo económico, la variable renta que hayamos incorporado como variable explicativa en (1).

En otros casos se dispondrá de una muestra de sección cruzada o de datos transversales, y tendremos una relación como (1) para cada una de las unidades muestrales que, en datos transversales, están constituidas por familias, empresas, países, comunidades autónomas, etc.. Por ejemplo, si disponemos de datos de observaciones de consumo y renta disponible de un conjunto de familias, podríamos especificar:

$$C_i = \beta_0 + \beta_1 Y_i, \quad i = 1, 2, \dots, n \quad (2)$$

siendo éste un modelo en que la interpretación del coeficiente β_1 sería ahora diferente de la que hicimos con datos de series temporales; en tal caso, β_1 nos proporciona el incremento que se produce en el gasto en consumo de una familia cuando aumenta su renta. No tendría ninguna connotación temporal, pues no hemos utilizado datos de tal tipo. De hecho, si dispusiésemos de dos muestras de sección cruzada, de las mismas familias, pero obtenidas en distintos momentos de un ciclo económico, bien podría ocurrir que la estimación del coeficiente β_1 variase significativamente entre ambas muestras.

En otras ocasiones, se pretende estimar una relación que no es de comportamiento, sino que refleja, más bien, un determinado proceso económico, como pueda ser la producción de bienes. Así, un modelo como:

$$C_t = \beta_0 + \beta_1 K_t + \beta_2 L_t, \quad t = 1, 2, \dots, T$$

podría interpretarse como la linealización de una función de producción agregada del tipo Cobb-Douglas para una determinada economía real, en la que los coeficientes β_1 y β_2 serían las elasticidades de producción de ambos inputs. En este caso, necesitaríamos un modelo de regresión algo más complejo que el modelo de regresión simple, que incluya varias variables explicativas.

El problema que nos interesa en economía estriba en la estimación de los valores numéricos de los dos coeficientes del modelo de regresión, por ejemplo, β_0 y β_1 en (2), así como en la posibilidad de contrastar hipótesis acerca de sus verdaderos valores numéricos, que son desconocidos.

1.2 Componentes del modelo de regresión

Por razones de exposición, y sin pérdida alguna de generalidad, suponemos en lo sucesivo que disponemos de una muestra de sección cruzada, y mantenemos el criterio notacional que venimos utilizando, designando con mayúsculas las variables genéricas con las que trabajamos: Y, X , y por minúsculas las observaciones numéricas incluidas en las muestras: $y_1, y_2, \dots, y_n, x_1, x_2, \dots, x_n$. Deno-

tamos el modelo de regresión, como relación entre las variables: $Y = \beta_0 + \beta_1 X$, mientras que denotamos la relación entre cada par de observaciones por: $y_i = \beta_0 + \beta_1 x_i$. Resulta evidente que es imposible que una relación como (1) se satisfaga para todas y cada una de las observaciones: $i = 1, 2, \dots, n$. Si ello ocurriese, podríamos sustituir las dos primeras observaciones muestrales de ambas variables en (1), y determinar exactamente los valores de los coeficientes β_0 y β_1 :

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \end{aligned}$$

obteniendo las estimaciones de dichos coeficientes con tan sólo estas dos observaciones muestrales. Sin embargo, no debe sorprendernos que al incorporar los valores numéricos de ambos coeficientes, junto con los de las variables Y y X correspondientes a la tercera observación en (1), $y_3 = \beta_0 + \beta_1 x_3$, la relación no se cumpla, salvo por una enorme casualidad.

Queda claro, por tanto, que no es obvio cómo obtener estimaciones de los coeficientes del modelo lineal simple a partir de una determinada muestra de T observaciones temporales, o n observaciones de sección cruzada. A ello dedicaremos algunas de las siguientes secciones. En cualquier caso, nos enfrentamos a una aparente paradoja: el modelo (1) no se satisfará para todas las observaciones muestrales, no importa qué valores numéricos asignemos a sus coeficientes β_0 y β_1 . Por ello, no consideramos exactamente el modelo (1), sino una variante del mismo:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, 3, \dots, n$$

donde la última variable, u_i , denominada perturbación estructural o término de error del modelo de regresión no es observable, y permite explicar las diferencias entre los dos miembros de la igualdad en (1). El problema de interés estriba en la estimación de los dos coeficientes en el modelo (2), cuando se dispone de una muestra de observaciones para las variables Y_i y X_i , aunque sin disponer de observaciones de la variable u_i .

La variable cuyo comportamiento se pretende explicar, Y_i , recibe el nombre de variable dependiente, mientras que la variable X_i recibe el nombre de variable independiente. En ocasiones, también se denomina a Y_i variable endógena o variable a explicar, mientras que a X_i se le denomina variable exógena o explicativa. Los coeficientes β_0 y β_1 se denominan término constante y pendiente del modelo de regresión simple, respectivamente.

La perturbación aleatoria, variable no observable para la que, en consecuencia, no dispondremos nunca de observaciones muestrales, se supone incorrelacionada con la variable X_i . Su interpretación es diversa:

- a) en primer lugar, puede contener otras variables explicativas que, aun siendo relevantes, no acertamos a especificar;
- b) también pudiera ser que, aun siendo conscientes de la existencia de tales variables, no dispusiéramos de observaciones muestrales para las mismas;

c) por último, el término de error puede estar reflejando errores de medida en la variable dependiente Y_i , que suelen surgir porque las variables que utilizamos en la estimación reflejan aproximadamente, pero no exactamente, los conceptos que querríamos incorporar en el modelo.

En el caso de la función de consumo anterior, es difícil en la práctica disponer de datos precisos acerca de los gastos en consumo de una determinada familia: en primer lugar, el consumo es un flujo, y la recogida de datos en un determinado instante de tiempo puede producir todo tipo de distorsiones en dicha variable. Para evitar este tipo de dificultades, en ocasiones, se utiliza como variable de consumo el resultado de sustraer de los ingresos declarados por la familia, el ahorro realizado durante el período.

Una vez estimados los coeficientes β_0 y β_1 en (2), tendríamos una ecuación lineal, una recta, entre el gasto en consumo y la renta de un conjunto de familias, denominada recta de regresión.

La recta de regresión proporciona la mejor relación existente entre las variables Y y X , en el caso de una regresión simple, o entre la variable dependiente, Y y el conjunto de variables explicativas, en una regresión lineal múltiple. Es tentador interpretar la recta de regresión como si nos proporcionase el valor *esperado* de Y *condicional* en los valores que pueda tomar la variable X . El concepto de *esperanza condicional* es, desde luego, muy importante en el análisis estadístico de datos económicos. Por ejemplo, un banco central puede estar interesado en un determinado momento en estimar la trayectoria que seguiría la tasa de inflación condicional a que dicho banco siga una política monetaria restrictiva. Querría asimismo caracterizar la trayectoria esperada de la inflación condicional a que se ponga en práctica una política monetaria expansiva, y así comparar ambas trayectorias esperadas, y escoger la política monetaria acorde a la senda de inflación preferible. De modo simple, este es un ejemplo del importante problema de diseño de política monetaria. Los modelos econométricos pueden ayudar en este tipo de situaciones. Una vez estimados los coeficientes β , disponemos de valores numéricos para ellos, y fijando una senda numérica para X (tasa de crecimiento monetario) podemos calcular una senda numérica para Y (tasa de inflación). Este ejercicio también se conoce como *predicción por escenarios*. Se trata de establecer sendas o escenarios alternativos para X , cuyos efectos se quieren comparar entre sí, estimar la senda de Y bajo cada uno de dichos escenarios, y calcular el resultado económico o de cualquier otro tipo.

El mismo esquema aplica a la gestión de la empresa, o en muchos contextos financieros. Por ejemplo, una empresa se está planteando la conveniencia de dos políticas de publicidad alternativa, una de bajo y otra de alto coste. Si, utilizando datos históricos, estima un modelo de regresión que explique las cifras de ventas (Y) utilizando el gasto en publicidad (X) durante los últimos 40 años, puede utilizar el modelo estimado para calcular aproximadamente las ventas que puede esperar bajo cada una de las dos políticas de publicidad. A continuación, un sencillo cálculo, aplicando los márgenes con que opera a las cifras de ventas estimadas y sustrayendo el coste de la campaña publicitaria, podrá decidir la preferencia por una u otra de las dos campañas.

Existe una limitación, sin embargo, y es que si recordamos el concepto de esperanza condicional, sabemos que dicha esperanza condicional es, en general, una función no lineal. Es decir, para calcular el valor esperado de Y para un determinado valor numérico de X , deberíamos utilizar la esperanza de la distribución de Y condicional en X , y ésta es, en general, una función no lineal. Cuando ambas variables, Y y X , tienen una distribución conjunta Normal, entonces, la esperanza condicional es una función lineal, pero no lo es en cualquier caso. Si no aceptamos la Normalidad de la distribución conjunta, entonces la regresión sólo se puede entender como una *aproximación* a la esperanza condicional de Y , dado X .

Por tanto, en este capítulo imponemos una forma funcional lineal para la dependencia de Y respecto de X y no hay ningún razón para pensar que la recta de regresión es una esperanza condicional. Para cada nivel de renta concreto como y^* , la recta estimada nos proporciona una estimación o predicción de gasto en consumo. Si hay alguna familia en la muestra con dicha renta, su gasto en consumo observado no coincidirá, salvo por casualidad, con el nivel *previsto* por la recta estimada. La diferencia:

$$\hat{u}_i = C_i - (\hat{\beta}_0 - \hat{\beta}_1 X_i),$$

que puede ser positiva, si el gasto en consumo excede del estimado por la recta, o negativa, si el gasto observado es inferior al estimado, se conoce como residuo de dicha observación muestral, denotado por \hat{u}_i y, como veremos en la sección 2, juega un papel fundamental en la estimación del modelo de regresión. Es importante observar que la recta de regresión estimada proporciona el nivel de consumo que deberíamos prever para cualquier nivel de renta, incluso si y^* no coincide con el de ninguna familia en la muestra. En tal caso tenemos un verdadero ejercicio de predicción.

En resumen, cuando se lleva a cabo un ejercicio empírico como la estimación del modelo de consumo (2), se tiene en mente un argumento del siguiente tipo: con el modelo (2) no se pretende explicar el comportamiento de la renta disponible de las familias, sino de su nivel de gastos en consumo. Para ello, consideramos las observaciones de la variable explicativa, la renta Y_i , como fijas: es decir, creemos que si hubiésemos entrevistado a otras n familias, hubiéramos generado los mismos datos para dicha variable. Sin embargo, las observaciones muestrales de la variable dependiente, el consumo C_i , habrían sido diferentes, como consecuencia de: a) aspectos específicos, no observables, de las familias encuestadas, b) errores de medida de diferente cuantía a aquellos en los que hemos incurrido en la muestra actualmente disponible, etc., y que aparecen recogidos en la perturbación aleatoria. El término de error es una variable aleatoria, diferente para cada observación muestral, y su realización no es observable. Por el contrario, el residuo es observable, puesto que se construye a partir de las estimaciones y de los datos de las variables dependiente e independiente. Término de error y residuo son entes de diferentes naturaleza.

Desde el punto de vista puramente estadístico, el modelo de regresión no tiene necesariamente una connotación de causalidad en la relación entre

variables. Del mismo modo que podemos estimar una regresión de una variable Y sobre otra variable X , podemos estimar una regresión en el orden inverso. Sin embargo, el análisis de este modelo elemental no trata a ambas variables de igual modo: las variables explicativas se consideran deterministas, mientras que la variable dependiente se considera aleatoria. El papel que juega cada una de las variables debe decidirse en función del aspecto teórico que está siendo objeto de estudio. En el ejemplo de consumo y renta, es evidente que queremos explicar los gastos en consumo en función de la renta, y no al revés; el consumo es la variable dependiente, y la renta es la variable independiente. Por eso, el investigador debe decidir de antemano el papel que juega cada una de estas dos variables, porque el tratamiento estadístico del modelo de regresión no concluye nada a este respecto. Sin embargo, su utilización en Econometría se efectúa condicional en una determinada hipótesis acerca de la dirección de la relación, y no al revés.

El modelo de regresión presupone que los valores numéricos de la variable dependiente gastos de consumo, C_i , se generan, en la realidad, a partir de los valores tomados por la variable renta Y_i y precisamente a través de la relación (2). En general, creemos que los procesos económicos son algo más complejos, y que se precisa más de una causa para explicar adecuadamente el comportamiento de una variable como el consumo, C_i , o bien formas funcionales más complicadas que la lineal. Sin embargo, el modelo de regresión simple es también una herramienta útil, al menos en una primera aproximación, desde la que no es muy complejo pasar al análisis del modelo de regresión lineal múltiple, cuyo estudio en profundidad dejamos para el capítulo siguiente, así como para cursos superiores.

Comentemos un poco más en detalle estos aspectos:

1.3 Supuestos del modelo de regresión lineal

1. Linealidad en las variables: en algunos casos, el supuesto de que la determinación de los valores del gasto en consumo, C_i , a partir de los de la renta, Y_i , se produce a través de un modelo lineal es excesivamente restrictiva, pues creemos que el modelo de relación es más bien no lineal. Examinaremos en el próximo capítulo una variedad de modelos alternativos al lineal que aquí analizamos. Sin embargo, en la mayoría de estos casos, el modelo lineal es nuevamente una buena aproximación al verdadero modelo, no lineal, de relación entre variable dependiente e independiente. El caso quizá más paradigmático de no linealidad, surge cuando se cree que el porcentaje de aumento en renta disponible que se transmite a consumo, no es constante, sino que decrece con el nivel de renta. Nótese que el modelo lineal tiene la propiedad de que el cociente de incrementos consumo/renta disponible o, si se prefiere, la derivada del consumo con respecto a la renta disponible, es 1, constante y, por ello, independiente del nivel de renta. Se tendría una relación muy distinta con un modelo del tipo:

$$C_i = \beta_0 + \beta_1 Y_i - \beta_2 Y_i^2 + u_i, \quad i = 1, 2, \dots, n$$

Este tipo de no linealidad en las variables puede incorporarse al análisis sin gran dificultad, del modo que veremos en el próximo capítulo,

2. Linealidad en los parámetros: muy diferente es la situación en que los parámetros entran en la relación entre variable dependiente e independientes de modo no lineal. El tratamiento que requieren tales modelos, con excepción de algunos casos sencillos, es sustancialmente más complejo, por lo que no es discutido en este texto,
3. Esperanza matemática nula: suponemos que la esperanza matemática del término de error u_i del modelo es cero: $E(u_i) = 0, i = 1, 2, \dots, n$. Si, por el contrario, tuviésemos: $E(u_i) = a \neq 0$, éste sería un efecto constante sobre Y_i y, por ello, determinista, y debería incluirse como parte del término constante β_0 en (1). Una situación en que este supuesto no se cumpliría es cuando el investigador, por error, omite del modelo una variable explicativa relevante. Así, supongamos que en vez de especificar el modelo:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{2t} + u_t, \quad t = 1, 2, 3, \dots, T$$

se especifica el modelo:

$$y_t = \beta_0 + \beta_1 x_t + v_t, \quad t = 1, 2, 3, \dots, T$$

en el que, inadvertidamente, se ha omitido la variable explicativa X_2 . En este último modelo, erróneamente especificado, el término de error v_t sería igual a: $v_t = \beta_2 x_{2t} + u_t$, y su esperanza matemática: $E(v_t) = E(\beta_2 x_{2t} + u_t) = E(\beta_2 x_{2t}) + E(u_t) = \beta_2 E(x_{2t}) + 0$, donde $E(X_2)$ denota la esperanza matemática de los valores que toma la variable omitida, X_2 , que suponemos constante a través del tiempo. Como consecuencia, $E(v_t)$ será distinta de cero en general,

4. Varianza constante del término de error (Homocedasticidad): suponemos que la varianza del término de error, que denotamos por $Var(u_i) = \sigma_u^2$ para todo $i = 1, 2, \dots, n$, es la misma para todas las observaciones muestrales, ya sean éstas de naturaleza temporal o de sección cruzada,
5. Ausencia de autocorrelación: además, suponemos que los términos de error correspondientes a dos observaciones muestrales cualesquiera, que son dos variables aleatorias diferentes, son estadísticamente incorrelacionadas (autocorrelación espacial en un corte transversal de datos ordenados geográficamente).
6. Estabilidad temporal: otro supuesto incorporado en el modelo es que sus coeficientes, β_0 y β_1 , son constantes en el tiempo; igualmente, creemos que el modelo es el mismo para todas las observaciones muestrales. Si disponemos de datos de series temporales, no hay submuestras de tiempo en las cuales los modelos sean diferentes; si estamos explicando los hábitos

de consumo de las familias españolas, creemos que la dependencia consumo/renta es igual para familias de renta alta y renta baja, o para familias que habitan en un medio rural y para las que viven en un medio urbano,

7. Causalidad unidireccional: también suponemos que existe una relación causal desde la variable explicativa X hacia la variable endógena Y , es decir, cambios en X influyen sobre cambios en Y , pero no al revés. Ello debe basarse en la naturaleza de la cuestión conceptual que se está analizando, y el investigador siempre debe tener buenos argumentos al respecto, pues ésta no es una cuestión empírica, sino teórica. De aquí surge la denominación de variable exógena para X , es decir, determinada fuera del modelo, y variable endógena, es decir, determinada dentro del modelo, para Y .

En el ejemplo de relación entre inflación y crecimiento monetario, si durante el período muestral se ha seguido una política monetaria consistente en fijar un determinado crecimiento anual para la cantidad de dinero y seguirlo estrictamente, el crecimiento monetario será una variable exógena en el modelo que pretende explicar la tasa de inflación. Si, por el contrario, se ha seguido una política monetaria en la que el crecimiento monetario se ha decidido en cada período como función de las tasas de inflación que hasta entonces se han registrado, entonces, no estaría justificado calificar de exógeno al crecimiento monetario a la inflación de endógena; quizá ambas deberían ser consideradas variables endógenas, para lo que necesitamos otro tipo de modelos

8. Variables explicativas deterministas: el modelo incorpora el supuesto, claramente restrictivo, acerca de que la variable explicativa X es determinista. La variable endógena Y no lo es, pues depende de la evolución de una variable aleatoria: el término de error del modelo, u .

En el ejemplo de relación entre expansión monetaria e inflación, este supuesto significa la creencia de que, si pudiésemos volver al año inicial en las mismas condiciones económicas entonces existentes, y recoger otra muestra para el mismo período, obtendríamos los mismos valores del crecimiento monetario. Desde este punto de vista, las tasas de crecimiento de la oferta monetaria que se han observado en este período son las únicas que pudieron haber ocurrido, con independencia de la información de que dispuso la autoridad monetaria, y de los objetivos de política económica que se trazaron. Sin embargo, nótese que, en esta hipotética situación, las tasas de inflación observadas para el período serían diferentes entre distintas muestras, debido a su componente estocástica u_t .

Enlazando con la discusión que mantuvimos en el punto anterior, podría tener sentido mantener el supuesto de una tasa de crecimiento monetario determinista bajo una política monetaria de fijación de una tasa de crecimiento constante todos los años; por el contrario, no podría mantenerse dicho supuesto bajo una política en que el crecimiento de la oferta monetaria se hace depender del "estado" de la economía y, en particular, de la

evolución de la tasa de inflación. De este modo, la clasificación de las variables explicativas en "exógenas" o endógenas" está ligada a que podamos mantener el supuesto de que son de naturaleza determinista.

En un análisis más general del modelo de regresión, que precisa de un instrumental técnico más complejo que el que presentamos en este texto, se considera que las variables explicativas son también estocásticas, como sin duda queremos creer en la realidad. En estas condiciones más generales, el modelo de regresión lineal simple está plenamente justificado bajo el supuesto de que las dos variables que en él aparecen, X e Y , tienen una distribución de probabilidad conjunta de carácter Normal o Gaussiano. En efecto, ya vimos al estudiar esta familia de distribuciones que la esperanza de la variable Y condicional en la variable X , es una expresión del tipo (1), donde las constantes β_0 y β_1 están relacionadas con los momentos de primer y segundo orden de la distribución bivalente Normal. De hecho, en tal caso, trabajamos generalmente bajo el supuesto de distribución Normal conjunta de todas las variables que aparecen en el modelo de regresión, e interpretamos éste como la esperanza condicional ya mencionada, lo cual puede extenderse al caso de varias variables explicativas.

2 El estimador de Mínimos Cuadrados Ordinarios

Supongamos que queremos estimar el modelo:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, 3, \dots, n$$

donde suponemos que: 1) u_i es una variable aleatoria con $E(u_i) = 0$ y $Var(u_i) = \sigma_u^2$ para todo i , 2) los valores x_i son fijos, 3) β_0 y β_1 son constantes desconocidas. Esta es la especificación del modelo de regresión lineal simple. Para ello, el investigador dispone de una muestra de 16 observaciones acerca de dos variables X e Y , la última de las cuales queremos explicar por medio de la primera:

Cuadro 1

n	Y	X	X ²	XY	Y-ajustada	u	Xu	u ²	Producto de	
									Residuo cuadrado	Desviaciones en X al cuadrado
1	16	15	225	240	16.3	-0.33	-5.0	0.11	20.8	15.1
2	18	13	169	234	14.7	3.26	42.4	10.66	6.6	13.6
3	8	11	121	88	13.1	-5.14	-56.5	26.39	0.3	-2.6
4	9	8	64	72	10.7	-1.74	-13.9	3.03	5.9	9.0
5	9	6	36	54	9.1	-0.14	-0.9	0.02	19.7	16.4
6	10	8	64	80	10.7	-0.74	-5.9	0.55	5.9	6.6
7	12	9	81	108	11.5	0.46	4.1	0.21	2.1	1.0
8	14	12	144	168	13.9	0.06	0.8	0.00	2.4	2.1
9	13	10	100	130	12.3	0.66	6.6	0.44	0.2	-0.1
10	10	5	25	50	8.3	1.66	8.3	2.75	29.6	14.6
11	7	9	81	63	11.5	-4.54	-40.9	20.60	2.1	8.2
12	15	12	144	180	13.9	1.06	12.8	1.13	2.4	3.6
13	16	13	169	208	14.7	1.26	16.4	1.60	6.6	8.5
14	18	18	324	324	18.7	-0.73	-13.1	0.53	57.2	40.2
15	15	10	100	150	12.3	2.66	26.6	7.09	0.2	-1.0
16	13	8	64	104	10.7	2.26	18.1	5.11	5.9	-0.8
Sumas :	203	167	1911	2253	203.00	0.00	0.00	80.22	167.94	134.19
Medias :	12.69	10.44	119.44	140.81	12.69	0.00	0.00	5.01	10.50	8.39
Varianzas:	11.71	10.50			6.70	5.01				

11

Así, tenemos un sistema de ecuaciones:

$$\begin{aligned}
 16 &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{u}_1, \\
 18 &= \hat{\beta}_0 + \hat{\beta}_1 X_2 + \hat{u}_2, \\
 8 &= \hat{\beta}_0 + \hat{\beta}_1 X_3 + \hat{u}_3, \\
 &\dots \\
 13 &= \hat{\beta}_0 + \hat{\beta}_1 X_{16} + \hat{u}_{16}
 \end{aligned}$$

que no puede resolverse, pues contiene 18 incógnitas, β_0 y β_1 , junto con los 16 residuos \hat{u}_i pero sólo 16 ecuaciones. Podríamos fijar los residuos igual a cero en dos ecuaciones y utilizarlas para obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$. Pero dichas estimaciones dependerán del par de ecuaciones seleccionadas, por lo que tal procedimiento no es adecuado. El método apropiado consiste en obtener valores numéricos para β_0 y β_1 que satisfagan de la manera más aproximada posible, simultáneamente, las 16 ecuaciones del sistema anterior.

Una vez estimados los coeficientes , se puede calcular para cada observación i :

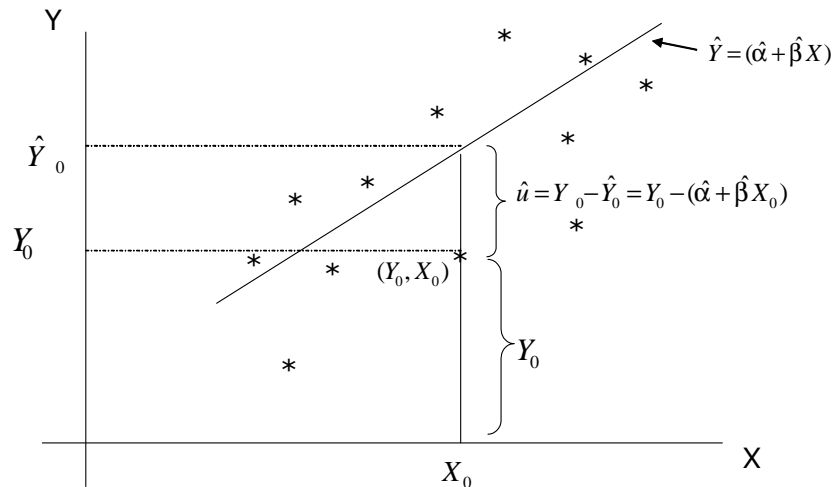
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (3)$$

en el que las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ han sustituido a los verdaderos valores, desconocidos. La expresión (3) representa la estimación, de acuerdo con el modelo econométrico, del valor que debía haber tomado la variable dependiente Y . Habrá siempre una discrepancia entre el valor realmente observado y_i y la estimación anterior, el residuo correspondiente a dicha observación muestral:

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),$$

Gráfico 1

Nube de puntos, recta de regresión, valores ajustados, residuos



Parece razonable que un posible criterio que defina a un estimador sea la minimización de la magnitud de los residuos que dicho estimador genera. Tal idea es correcta, pero hay varias dificultades para hacerla práctica: en primer lugar, tenemos no un residuo, sino un conjunto de n residuos, por lo que no se trata de minimizar un residuo determinado, sino una medida conjunta del tamaño global de todos ellos.

Una vez obtenidas unas estimaciones numéricas de los coeficientes, podría pensarse en sumar los n residuos generados: $\sum_{i=1}^n \hat{u}_i$, y escoger como estimación el par de valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que produce la menor suma de residuos. Una dificultad con tal procedimiento es la cancelación de residuos negativos con

residuos positivos. Además, si realmente se pretendiese minimizar la suma de residuos, bastaría generar residuos de tamaño muy grande, pero negativos, lo cual no es adecuado.

El estimador de mínimos cuadrados que introducimos en esta sección utiliza como criterio la minimización de la Suma de los Cuadrados de los Residuos (*SCR*), o también Suma Residual, aunque hay que recordar que es una suma de cuadrados. Se trata, por tanto, de seleccionar valores de los coeficientes β_0 y β_1 que resuelvan el problema:

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{Minimizar}} SCR = \sum_{i=1}^n \hat{u}_i^2$$

Nótese que el residuo asociado a cada observación $i, i = 1, 2, \dots, n$, depende de los valores de los coeficientes escogidos, porque:

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

de modo que el problema anterior puede escribirse:

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{Minimizar}} SR = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

La solución a este problema de optimización se denota por: $\hat{\beta}_0, \hat{\beta}_1$, y se denomina estimador de Mínimos Cuadrados Ordinarios (que abreviaremos como MCO) de los coeficientes del modelo de regresión lineal simple. El estimador MCO escoge, de entre todas las posibles, la recta que minimiza la suma de los cuadrados de las distancias entre cada punto de la nube generada por las observaciones muestrales y el asignado por la recta.

Derivando *SR* con respecto a ambas variables (β_0 y β_1) e igualando dichas derivadas a cero, tenemos:

$$\frac{\partial SR}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (4)$$

$$\frac{\partial SR}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (5)$$

con matriz de derivadas segundas:

$$\frac{\partial^2 SR}{\partial \beta_0 \partial \beta_1} = \begin{matrix} 2n & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2 \sum_{i=1}^n x_i^2 \end{matrix}$$

que tiene por determinante:

$$DET = 4 \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) = n^2 \left(\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 \right) = n^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = n^2 S_x^2$$

que es positiva. Como el primer menor, el elemento (1,1) de esta matriz, que es $2n$, es también positivo, podemos afirmar que la solución al sistema de ecuaciones (4) y (5) serán, los valores numéricos de los coeficientes β_0 y β_1 que, efectivamente, alcanzan un mínimo de la Suma Residual.

Si resolvemos dicho sistema, obtenemos:

$$\sum_{i=1}^n y_i = n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i \quad (6)$$

$$\sum_{i=1}^n y_i x_i = \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \quad (7)$$

que constituyen un par de ecuaciones simultáneas en las incógnitas, $\hat{\beta}_0$, $\hat{\beta}_1$. Este sistema se conoce como sistema de ecuaciones normales.

Utilizando los estadísticos que aparecen en la última fila del Cuadro 1, tendríamos:

$$\begin{aligned} 203 &= 16\beta_0 + 167\beta_1 \\ 134,2 &= 167\beta_0 + \beta_1 \end{aligned}$$

que resuelto, proporciona las estimaciones MCO:

$$\hat{\beta}_0 = 4,35; \hat{\beta}_1 = 0,799$$

con dichos datos. La sexta columna del cuadro presenta los valores previstos por el modelo para la variable dependiente. La columna siguiente muestra los residuos, es decir, la diferencia entre los valores de Y y los valores previstos por el modelo.

En general, si primero despejamos $\hat{\beta}_0$ en (6), tenemos:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (8)$$

que podremos utilizar para obtener el estimador MCO de β_0 , una vez que tengamos el estimador de 1. Substituyendo en (7), tenemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \rho_{xy} \frac{S_y}{S_x} \quad (9)$$

donde S_{xy} , S_x^2 , S_y^2 , S_x , S_y , denotan, respectivamente, la covarianza, varianzas y desviaciones típicas muestrales de X e Y . Las expresiones (8) y (9) son útiles, pues proporcionan directamente las estimaciones MCO como función de estadísticos muestrales, sin necesidad de resolver el sistema de ecuaciones normales. Primero se calcula $\hat{\beta}_1$ y, luego, se obtiene: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Ello demuestra una propiedad del estimador MCO: la recta estimada pasa por el punto (\bar{y}, \bar{x}) .

Nótese que las ecuaciones anteriores pueden escribirse también:

$$\begin{aligned}\sum_{i=1}^n \hat{u}_i &= 0 \\ \sum_{i=1}^n x_i \hat{u}_i &= 0\end{aligned}$$

que son dos propiedades del estimador de mínimos cuadrados:

- 1) la suma de los residuos que genera el estimador de mínimos cuadrados es igual a cero, lo que no necesariamente ocurre con otro procedimiento de estimación [ver suma de la columna 7 del Cuadro 1], y
- 2) los residuos de mínimos cuadrados están incorrelacionados con la variable explicativa del modelo. Cuando se considera un modelo de regresión lineal general o múltiple, que incluye no una, sino k variables explicativas, los residuos de mínimos cuadrados están incorrelacionados con todas las variables explicativas del modelo [ver suma de la columna 8 del Cuadro 1].

2.1 Esperanza matemática

La expresión del estimador MCO de la pendiente del modelo de regresión lineal simple puede escribirse:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) y_i = \sum_{i=1}^n \alpha_i y_i \quad (10)$$

como una combinación lineal ponderada de las observaciones de la variable endógena, con ponderaciones:

$$\alpha_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

En esta cadena de igualdades hemos utilizado el hecho de que la suma de las desviaciones de una variable con respecto a su media muestral, es siempre igual a cero. Las ponderaciones en esta expresión suman cero:

$$\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

Además:

$$\begin{aligned}\sum_{i=1}^n \alpha_i x_i &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) x_i = \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i}{\sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2)} = \\ &= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = 1\end{aligned}$$

Recordemos que estamos suponiendo que los valores x_1, x_2, \dots tomados por la variable X son fijos, es decir, no están sujetos a ninguna incertidumbre, de modo que, si volviésemos a tomar otra muestra de igual tamaño, tendríamos para dicha variable las mismas observaciones numéricas, una por una, que las que ya disponemos. Tan sólo las observaciones y_1, y_2, \dots de la variable endógena Y diferirían de las actualmente disponibles, debido a que las realizaciones muestrales de la perturbación aleatoria u_i , el único componente aleatorio de Y , serían diferentes de las actuales. Vamos a utilizar ahora repetidamente el carácter determinista no aleatorio, de la variable X .

Si sustituimos en (10) y_i por su expresión a través del modelo de regresión, tenemos:

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^n \alpha_i (\beta_0 + \beta_1 x_i + u_i) = \sum_{i=1}^n \alpha_i \beta_0 + \sum_{i=1}^n \alpha_i \beta_1 x_i + \sum_{i=1}^n \alpha_i u_i = (11) \\ &= \beta_0 \sum_{i=1}^n \alpha_i + \beta_1 \sum_{i=1}^n \alpha_i x_i + \sum_{i=1}^n \alpha_i u_i = \beta_0 \cdot 0 + \beta_1 \cdot 1 + \sum_{i=1}^n \alpha_i u_i = \\ &= \beta_1 + \sum_{i=1}^n \alpha_i u_i \end{aligned}$$

donde hemos utilizado las dos propiedades antes demostradas. Esta es una representación muy útil, que presenta el estimador de mínimos cuadrados de la pendiente como una combinación lineal de las perturbaciones del modelo, con coeficientes α_i , más una constante desconocida, el verdadero valor de dicha pendiente. Los coeficientes α_i en dicha combinación lineal varían de una muestra a otra con los valores de la variable explicativa, X , por lo que el valor numérico del estimador de mínimos cuadrados también variaría si dispusiéramos de distintas muestras recogidas en distintos períodos de tiempo, por ejemplo.

Es importante recordar que suponemos que la variable explicativa es determinista. Es decir, que los valores numéricos observados en la muestra para dicha variable son los únicos posibles, dadas las unidades de observación muestral, sean individuos, empresas, familias, o un conjunto de observaciones de determinada frecuencia (diaria, mensual, trimestral anual) a lo largo de un determinado intervalo de tiempo. Recordemos que de una muestra a otra, cambiarían los valores observados de la variable dependiente, y_i porque cambiaría la realización numérica de las perturbaciones u_i , pero no porque cambiaran los valores de la variable explicativa x_i , que serían los mismos entre distintas muestras extraídas de las mismas unidades de observación.

A continuación, vamos a obtener la esperanza matemática y la varianza de los estimadores de mínimos cuadrados de $\hat{\beta}_0$ y $\hat{\beta}_1$. Esto es necesario para poder proceder a contrastar hipótesis acerca de sus verdaderos valores que, recordemos, son desconocidos. Disponemos de una estimación numérica, obtenida con la muestra disponible, que sería diferente si pudiésemos calcularla con otra muestra distinta.

Tomando esperanzas, y notando que:

$$E(\alpha_i u_i) = \alpha_i 0 = 0$$

tenemos:

$$E(\hat{\beta}_1) = \beta_1 + E\left(\sum_{i=1}^n \alpha_i u_i\right) = \beta_1 + \sum_{i=1}^n E(\alpha_i u_i) = \beta_1 + \sum_{i=1}^n \alpha_i E(u_i) = \beta_1$$

lo que prueba que el estimador MCO del parámetro β_1 es *insesgado*, puesto que su esperanza matemática coincide con el verdadero valor del parámetro que se pretende estimar, que es desconocido.

Notemos que el supuesto de que la variable explicativa no es aleatoria es crucial para probar la ausencia de sesgo del estimador de mínimos cuadrados. En las expresiones anteriores nos hemos encontrado con $E(\alpha_i u_i)$, y cada α_i depende de todas las observaciones $x_j, j = 1, 2, \dots, n$. Si fuese aleatoria, no sabríamos decir nada acerca de la esperanza matemática $E(x_i u_i)$, salvo haciendo supuestos específicos acerca de la covarianza entre ambas variables aleatorias, x_i y u_i , pero mucho menos acerca de la esperanza $E(\alpha_i u_i)$.

Recordando que la expresión del estimador MCO del término independiente β_0 es:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

notemos que:

$$E(\bar{y}) = \beta_0 + E(\hat{\beta}_1 \bar{x}) + E(\bar{u}) = \beta_0 + \beta_1 \bar{x}$$

por lo que:

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1 \bar{x}) = (\beta_0 + \beta_1 \bar{x}) - E(\hat{\beta}_1) \cdot \bar{x} = (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0$$

de modo que, al igual que ocurría con la estimación de β_1 , el estimador MCO de β_0 es también *insesgado*.

La recta de regresión estimada pasa por el punto (\bar{x}, \bar{y}) . Es decir, el valor numérico que la recta de regresión estimada asocia a la variable dependiente Y cuando $X = \bar{x}$ es, precisamente, $Y = \bar{y}$. En efecto:

$$y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

El punto (\bar{x}, \bar{y}) se conoce en ocasiones como el *centro de gravedad* de la nube de puntos $(x_i, y_i), i = 1, 2, \dots, N$.

2.2 Matriz de covarianzas

Todo estimador puntual debe ir siempre acompañado de una medida de dispersión del mismo, generalmente su varianza, de modo que podamos juzgar el grado en que se aproxima al verdadero valor del parámetro que pretendemos estimar. Pero además, para poder llevar a cabo un análisis de inferencia estadística, es decir, para poder contrastar si alguno de los coeficientes β_0 ó β_1 , o ambos, toman determinados valores teóricos, es preciso disponer de desviaciones típicas de sus estimaciones. Estos no son sino un caso particular de los problemas de estimación e inferencia estadísticos, y los resolvemos de modo similar, mediante la construcción de intervalos de confianza, al nivel deseado, alrededor del valor hipotético que se pretende contrastar.

Recordemos el supuesto de que las perturbaciones aleatorias del modelo correspondientes a todas las unidades muestrales tienen la misma varianza, σ_u^2 . Por tanto, si partimos de la expresión (11) que antes obtuvimos para el estimador de β_1 , tenemos:

$$Var(\alpha_i u_i) = \alpha_i^2 Var(u_i) = \alpha_i^2 \sigma_u^2$$

para cualquier $i = 1, 2, \dots$. Entonces, puesto que la covarianza entre u_i y u_j es igual a cero, se tiene:

$$\begin{aligned} E\left(\sum_{i=1}^n \alpha_i u_i\right) &= \sum_{i=1}^n E(\alpha_i u_i) = \sum_{i=1}^n \alpha_i E(u_i) = \sum_{i=1}^n \alpha_i 0 = 0 \\ Var\left(\sum_{i=1}^n \alpha_i u_i\right) &= \sum_{i=1}^n Var(\alpha_i u_i) = \sum_{i=1}^n \alpha_i^2 Var(u_i) = \sigma_u^2 \left(\sum_{i=1}^n \alpha_i^2\right) \end{aligned}$$

Si calculamos la suma de los cuadrados de las ponderaciones:

$$\sum_{i=1}^n \alpha_i^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{nS_x^2}$$

siendo S_x^2 la varianza muestral de X : $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$.

Como el estimador $\hat{\beta}_1$ es la suma de una constante (el verdadero valor β_1) y una variable aleatoria (la suma ponderada de las perturbaciones) [ver (11)], la varianza de $\hat{\beta}_1$ será igual tan sólo a la varianza de esta última suma:

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^n \alpha_i u_i\right) = \sigma_u^2 \left(\sum_{i=1}^n \alpha_i^2\right) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_u^2}{nS_x^2}$$

Para obtener la varianza del estimador MCO de $\hat{\beta}_0$, notemos que:

$$Var(\hat{\beta}_0) = Var(\bar{y}) + Var(\hat{\beta}_1 \bar{x}) - 2Cov(\bar{y}, \hat{\beta}_1 \bar{x}) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1)$$

donde aparece la varianza de la media muestral de la variable endógena, que podemos calcular, del siguiente modo: si sumamos la expresión (1) del modelo lineal simple para todas las observaciones muestrales, tenemos:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n (\beta_0 + \beta_1 x_i) + \sum_{i=1}^n u_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i + \sum_{i=1}^n u_i$$

y, dividimos por el tamaño muestral, n :

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

lo que puede utilizarse para probar que:

$$\begin{aligned} Var(\bar{y}) &= Var(\beta_0 + \beta_1 \bar{x} + \bar{u}) = Var(\beta_0) + Var(\beta_1 \bar{x}) + Var(\bar{u}) = 0 + 0 + \frac{\sigma_u^2}{n} = \frac{\sigma_u^2}{n} \\ Cov(\bar{y}, u_i) &= Cov(\beta_0 + \beta_1 \bar{x} + \bar{u}, u_i) = Cov(\beta_0, u_i) + \bar{x}Cov(\beta_1, u_i) + Cov(\bar{u}, u_i) = \\ &= 0 + 0 + \frac{1}{n} \sum_{j=1}^n Cov(u_j, u_i) = \frac{1}{n} \sigma_u^2 \\ Cov(\bar{y}, \hat{\beta}_1) &= Cov\left(\bar{y}, \beta_1 + \sum_{i=1}^n \alpha_i u_i\right) = Cov(\bar{y}, \beta_1) + \sum_{i=1}^n \alpha_i Cov(\bar{y}, u_i) = 0 + \frac{1}{n} \sigma_u^2 \sum_{i=1}^n \alpha_i = 0 \end{aligned}$$

por lo que tenemos:

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1) = \frac{\sigma_u^2}{n} + \bar{x}^2 \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \\ &= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \sigma_u^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

y, por tanto:

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = \bar{x}Cov(\bar{y}, \hat{\beta}_1) - \bar{x}Var(\hat{\beta}_1) = \\ &= 0 - \bar{x} \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = - \frac{\bar{x} \sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Argumento alternativo:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = (\beta_0 + \beta_1 \bar{x} + \bar{u}) - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u} \\
\hat{\beta}_0 - E\hat{\beta}_0 &= (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u} \\
Var(\hat{\beta}_0) &= E\left[(\hat{\beta}_1 - E\hat{\beta}_1) \bar{x} + \bar{u}\right]^2 = E\left[(\beta_1 - \hat{\beta}_1)^2 \bar{x}^2\right] + E(\bar{u}^2) + 2E\left[(\hat{\beta}_1 - E\hat{\beta}_1) \bar{x} \bar{u}\right] = \\
&= \bar{x}^2 Var(\hat{\beta}_1) + \frac{\sigma_u^2}{n} - 2\bar{x}E\left[(\hat{\beta}_1 - E\hat{\beta}_1) \bar{u}\right] \\
\text{Pero} &: E\left[(\hat{\beta}_1 - E\hat{\beta}_1) \bar{u}\right] = E\left[\left(\sum_{i=1}^n \alpha_i u_i\right) \left(\frac{1}{n} \sum_{j=1}^n u_j\right)\right] = \frac{\sigma_u^2}{n} \sum_{i=1}^n \alpha_i = 0 \\
\text{Luego} &: Var(\hat{\beta}_0) = \bar{x}^2 Var(\hat{\beta}_1) + \frac{\sigma_u^2}{n} = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \sigma_u^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
Cov(\hat{\beta}_0, \hat{\beta}_1) &= E\left[(\hat{\beta}_0 - \beta_0) (\hat{\beta}_1 - \beta_1)\right] = E\left[(\bar{u} - (\hat{\beta}_1 - \beta_1) \bar{x}) (\hat{\beta}_1 - \beta_1)\right] = \\
&= E\left[\bar{u} (\hat{\beta}_1 - \beta_1)\right] - \bar{x}E\left[(\hat{\beta}_1 - \beta_1)^2\right] = 0 - \bar{x}Var(\hat{\beta}_1) = -\frac{\bar{x}\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

que indica, entre otras cosas, que el signo de la covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ es el opuesto al signo de la media muestral de la variable X .

Supongamos que dicha media fuese positiva, y también que el error de estimación de β_1 fuese asimismo positivo, es decir, que hubiésemos estimado (sin saberlo), un valor $\hat{\beta}_1$ superior al teórico. Su producto por la media de X generaría, en promedio, una contribución positiva del error de estimación a la explicación de la variable Y :

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u} = [\beta_0 + \beta_1 \bar{x}] + [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) \bar{x}]$$

donde en el corchete de la derecha, el segundo sumando está teniendo una contribución positiva. Para compensarlo, la estimación MCO de β_0 estaría por debajo de su valor verdadero: $\beta_0 > \hat{\beta}_0$. Es decir, si el estimador de Mínimos Cuadrados sobreestima β_1 , entonces infraestima β_1 . Si infraestimamos β_1 , entonces sobreestimamos β_0 . Lo contrario ocurriría si la media muestral de X fuese negativa.

3 El modelo de regresión lineal en desviaciones respecto de la media

Como hemos visto en la sección anterior, a partir del modelo de regresión lineal:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, 3, \dots, n$$

se deduce que:

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

y, restando, tenemos un modelo en el que cada variable aparece en desviaciones respecto de su media muestral:

$$y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + (u_i - \bar{u}), \quad i = 1, 2, 3, \dots, n$$

Nótese que la primera y tercera ecuaciones son válidas para cada observación muestral y tenemos, en cada una de ellas, tantas relaciones como observaciones muestrales, mientras que la segunda ecuación aplica sólo a las medias muestrales y constituye, por tanto, una única relación.

En el modelo en desviaciones no hay término independiente, y el término de error es distinto del término de error del modelo original.

Si estimamos este modelo en diferencias por mínimos cuadrados, tendremos el mismo estimador de β_1 que en el modelo original, ya que:

$$\begin{aligned} \text{Var}(x_i - \bar{x}) &= \text{Var}(x_i) \\ \text{Cov}[(x_i - \bar{x}), (y_i - \bar{y})] &= \text{Cov}(x_i, y_i) \end{aligned}$$

Aunque no habremos estimado β_0 , puesto que dicho parámetro ha desaparecido del modelo, podemos utilizar la relación que obtuvimos antes para calcular $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

La varianza del término de error del modelo en diferencias es ligeramente distinta del modelo original, puesto que:

$$\begin{aligned} \text{Var}(u_i - \bar{u}) &= E[u_i(u_i - \bar{u})] = E(u_i^2) - E(u_i \bar{u}) = \\ &= E(u_i^2) - E\left(u_i \sum_{i=1}^n \frac{u_i}{n}\right) = \sigma_u^2 - \frac{1}{n} \sigma_u^2 = \frac{n-1}{n} \sigma_u^2 \end{aligned}$$

Los residuos del modelo estimado con las variables en desviaciones respecto de la media son:

$$\hat{v}_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) = y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i$$

y, por tanto, coinciden numéricamente, con los que se obtienen estimando el modelo con las variables originales.

4 Estimación de la varianza del término de error o perturbación aleatoria del modelo

Conociendo las expresiones analíticas de las varianzas de ambos estimadores, así como también de su covarianza, podremos contrastar hipótesis acerca de valores teóricos para alguno de los dos coeficientes, y también contrastar hipótesis

conjuntas, acerca de ambos simultáneamente. Pero en ellas aparece la varianza del término de error σ_u^2 , que es desconocida. Debemos, por tanto, estimar este parámetro, y utilizar su estimación en lugar de su verdadero valor, que es desconocido.

Por similitud, parece razonable utilizar la varianza muestral de los residuos como un estimador de la varianza poblacional σ_u^2 . Los residuos de mínimos cuadrados tienen media cero, como muestra la primera ecuación normal, por lo que su varianza muestral es: $S_{\hat{u}}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 = SCR/n$. Pero estimamos con una pequeña corrección:

$$\hat{\sigma}_u^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{n}{n-2} S_{\hat{u}}^2$$

Tomamos $n-2$ y no simplemente n en el denominador, pero que el estimador $\hat{\sigma}_u^2$ sea insesgado [ver Apéndice]. Una vez que se dispone de una estimación de la varianza, puede utilizarse en las expresiones de la varianza de los estimadores de los coeficientes, de manera que tenemos así estimaciones de las varianzas de los coeficientes estimados, lo que indicaremos con un circunflejo encima de la palabra "Varianza".

Ejemplo.- Con los datos del Cuadro 1, tenemos una Suma Residual, es decir, una suma de cuadrados de residuos, de 80,2. Ello nos lleva a la estimación de la varianza del término de error:

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{80,2}{16-2} = 5,729 \Rightarrow \hat{\sigma}_u = 2,393 \\ R^2 &= 1 - \frac{\text{Suma Cuadrados Residuos}}{\text{Suma Total}} = 1 - \frac{5,014}{11,715} = 1 - 0,428 = 0,572 \end{aligned}$$

Podemos utilizar ahora la estimación de σ_u^2 en las expresiones de las varianzas de los estimadores de Mínimos Cuadrados:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{5,729}{167,9} = 0,03417 \Rightarrow DT(\hat{\beta}_1) = 0,185 \\ \text{Var}(\hat{\beta}_0) &= \sigma_u^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{5,729}{16} \frac{1911}{167,9} = 4,075 \Rightarrow DT(\hat{\beta}_2) = 2,02 \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\bar{x} \sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = -\frac{5,729}{167,9} (10,4) = -0,354 \end{aligned}$$

Finalmente, el modelo estimado se representa escribiéndolo como la función lineal que es, anotando debajo de los coeficientes estimados sus desviaciones típicas que son, asimismo, estimadas, como acabamos de ver, pues sus verdaderos valores dependen de σ_u^2 :

$$y_i = 4,35 + 0.799x_i + u_i, \quad R^2 = 0,572; \quad \hat{\sigma}_u = 2,393$$

(2,02) (0,185)

Ejemplo.- Consideremos un modelo muy sencillo:

$$y_i = \beta_0 + u_i,$$

en el que aparece una constante como única variable explicativa, por lo que se denomina modelo constante de regresión. El estimador MCO será el estadístico muestral que minimice la suma de los residuos, que en este caso es:

$$SCR = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \beta_0)^2,$$

por lo que se trata de minimizar la suma de las desviaciones al cuadrado entre los valores muestrales de la variable Y y un estadístico. La solución a dicho problema de minimización está dada por la media muestral, y el valor minimizado es, por tanto, la varianza muestral. En consecuencia, el estimador del modelo constante de regresión es la media muestral. Ello significa que la media muestral es el estimador óptimo, cuando no se dispone de información acerca de ninguna otra variable. En tal situación, lo mejor que podemos hacer es aproximar cada valor potencialmente observable de la variable Y por la media muestral de que dispongamos. Es, desde luego, un estimador algo pobre, pero nos sirve de referencia a la que hay que mejorar; es decir, contando con información muestral acerca de alguna otra variable, hemos de conseguir estimaciones MCO de un modelo de regresión tales que la Suma de Cuadrados de Residuos que generan sea inferior a la varianza muestral de Y. Pero ello va a ocurrir siempre. Cuando se estima el modelo constante, la Suma de Cuadrados de Residuos, que es la varianza de Y, coincide con la Suma Total, por lo que el coeficiente de determinación es igual a cero. Ningún otro modelo tendrá un coeficiente de determinación inferior.

5 Eficiencia

En el modelo de regresión, la aleatoriedad proviene del término de error, de quien suponemos que tiene esperanza matemática nula y varianza σ_u^2 . La aleatoriedad se transmite a la variable y_i , que tiene esperanza $E(y_i) = \beta_0 + \beta_1 x_i$ y varianza σ_u^2 , igual a la de u_i , de quien se diferencia en una constante, $\beta_0 + \beta_1 x_i$. Por otra parte, (10) muestra que el estimador MCO de β_1 depende linealmente de las observaciones de la variable aleatoria Y. También $\hat{\beta}_0$ es una combinación lineal de las observaciones y_i :

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \left(\beta_1 + \sum_{i=1}^n \alpha_i u_i \right) \bar{x} = \bar{y} - \beta_1 \bar{x} - \bar{x} \left(\sum_{i=1}^n \alpha_i (y_i - \beta_0 - \beta_1 x_i) \right) = \\
&= \bar{y} - \beta_1 \bar{x} - \bar{x} \sum_{i=1}^n \alpha_i y_i + \beta_0 \bar{x} \sum_{i=1}^n \alpha_i + \beta_1 \bar{x} \sum_{i=1}^n \alpha_i x_i = \bar{y} - \beta_1 \bar{x} - \bar{x} \sum_{i=1}^n \alpha_i y_i + \beta_0 \bar{x} \cdot 0 + \beta_1 \bar{x} \cdot 1 = \\
&= \bar{y} - \bar{x} \sum_{i=1}^n \alpha_i y_i = \frac{1}{n} \sum_{i=1}^n y_i - \bar{x} \sum_{i=1}^n \alpha_i y_i = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} \alpha_i \right) y_i
\end{aligned}$$

Pues bien, el estimador MCO es de mínima varianza dentro de la clase de estimadores lineales:

Theorem 1 (*Teorema de Gauss-Markov*).- *Bajo los supuestos del modelo lineal de regresión, el estimador MCO es el estimador lineal insesgado de mínima varianza de los coeficientes del modelo de regresión.*

Proof. Consideremos un estimador lineal de la pendiente del modelo de regresión:

$$\tilde{\beta}_1 = \sum_{i=1}^n c_i y_i$$

que supondremos distinto del estimador de mínimos cuadrados, es decir, que no todas las constantes c_i son iguales a las α_i . Para que este estimador sea insesgado ha de cumplirse:

$$\begin{aligned}
E(\tilde{\beta}_1) &= E\left(\sum_{i=1}^n c_i y_i\right) = E\left(\sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i + u_i)\right) = E\left(\beta_0 \sum_{i=1}^n c_i\right) + \beta_1 E\sum_{i=1}^n c_i x_i + E\sum_{i=1}^n c_i u_i = \\
&= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i + 0
\end{aligned}$$

que será igual a β_1 y, con ello, el estimador $\tilde{\beta}_1$ será insesgado sólo si se cumple, simultáneamente:

$$\begin{aligned}
\sum_{i=1}^n c_i &= 0 \\
\sum_{i=1}^n c_i x_i &= 1
\end{aligned}$$

Suponemos, por tanto, que las constantes c_i satisfacen ambas condiciones. Teniendo en cuenta que tanto $\sum_{i=1}^n c_i$ como $\sum_{i=1}^n c_i x_i$ son constantes, la varianza de este estimador es:

$$Var(\tilde{\beta}_1) = Var\left(\sum_{i=1}^n c_i u_i\right) = \sum_{i=1}^n Var(c_i u_i) = \sigma_u^2 \sum_{i=1}^n c_i^2$$

de modo que, para probar que el estimador de mínimos cuadrados tiene menor varianza que este estimador lineal insesgado genérico, habremos de probar que:

$$\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \leq \sum_{i=1}^n c_i^2$$

con independencia de cuáles sean las constantes $c_i, i = 1, 2, \dots, n$.

Para ello, consideremos la expresión:

$$\begin{aligned} \sum_{i=1}^n \left(c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 &= \sum_{i=1}^n c_i^2 - 2 \sum_{i=1}^n c_i \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 = \\ &= \sum_{i=1}^n c_i^2 - 2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \sum_{i=1}^n c_i^2 - 2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \\ &= \sum_{i=1}^n c_i^2 - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \geq 0 \end{aligned}$$

donde la última desigualdad proviene del hecho de que el punto de partida es una suma de cuadrados y por tanto, necesariamente positiva.

Pero esto significa que, como queríamos mostrar:

$$\sum_{i=1}^n c_i^2 \geq \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

■

El teorema de Gauss-Markov es importante, por cuanto que afirma que la matriz de covarianzas del estimador MCO es inferior a la de cualquier otro estimador lineal e insesgado. Es decir, la diferencia entre ambas matrices, en el orden citado, es semidefinida negativa. Ello tiene implicaciones más útiles: la varianza del estimador MCO de β_0 es inferior a la de cualquier otro estimador lineal e insesgado de dicho coeficiente, y lo mismo ocurre con la varianza del estimador MCO de β_1 .

Cuando el término de error del modelo tiene una distribución Normal, tenemos un resultado aún más importante, que afirma que el estimador MCO es eficiente, es decir, tiene la menor varianza posible (la menor matriz de covarianzas), dentro de la clase de los estimadores insesgados, sean estos lineales o no.

Theorem 2 *Teorema de Rao.*- Si se cumplen las condiciones de la Sección 13.1 y, además, el término de error del modelo tiene distribución Normal, entonces el estimador MCO es el estimador insesgado de mínima varianza de los coeficientes del modelo de regresión.

Proof. La demostración se basa en probar que, cuando el término de error del modelo de regresión tiene distribución Normal, $u_i \sim N(0, \sigma_u^2)$, entonces el estimador de Mínimos Cuadrados coincide con el estimador de Máxima Verosimilitud. Como este último es siempre (bajo condiciones muy generales y, por tanto, fáciles de satisfacer) el estimador de mínima varianza o eficiente, habremos probado que, en este caso especial, el estimador de mínimos cuadrados también lo es. ■

Consideremos el modelo de regresión con término de error Normal:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad u_i \sim N(0, \sigma_u^2)$$

del que escribimos la función de verosimilitud:

$$L(\beta_0, \beta_1, \sigma_u^2 / y_1, x_1, y_2, x_2, \dots, y_n, x_n) = \prod_{i=1}^n \frac{1}{\sigma_u \sqrt{2\pi}} e^{-u_i^2 / 2\sigma_u^2}$$

y su logaritmo:

$$\begin{aligned} \ln L(\beta_0, \beta_1, \sigma_u^2 / y_1, x_1, \dots, y_n, x_n) &= -\frac{n}{2} \ln \sigma_u^2 - \frac{n}{2} \ln (2\pi) - \sum_{i=1}^n \frac{u_i^2}{2\sigma_u^2} = \\ &= -\frac{n}{2} \ln \sigma_u^2 - \frac{n}{2} \ln (2\pi) - \sum_{i=1}^n \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma_u^2} \end{aligned}$$

El estimador de Máxima Verosimilitud se obtiene derivando en la expresión anterior con respecto a los parámetros desconocidos: β_0 , β_1 , σ_u^2 , e igualando a cero dichas derivadas.

Pero sin necesidad siquiera de hacer dicho cálculo, ya apreciamos que los valores numéricos de β_0 y β_1 que maximizan $\ln L$ son los mismos que minimizan la Suma de Cuadrados de los Residuos, ya que ésta entra con signo menos en la expresión de $\ln L$. Por tanto, los estimadores de Mínimos Cuadrados y de Máxima Verosimilitud de ambos parámetros coinciden, y el teorema queda probado.

Este resultado es importante, porque justifica el uso del estimador de Mínimos Cuadrados, dado que es un estimador eficiente. Pero, como con cualquier teorema, es preciso entender el conjunto de condiciones bajo las que puede afirmarse la conclusión que se ha obtenido. En nuestro caso, es especialmente importante recordar que la eficiencia del estimador de Mínimos Cuadrados se obtiene si el término de error del modelo sigue una distribución Normal, pero no necesariamente en otro caso.

El estimador de Máxima verosimilitud de la varianza del término de error es:

$$\hat{\sigma}_{MV}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2$$

que es parecido, aunque no idéntico, al estimador MCO de dicho parámetro. De hecho, como sabemos [ver Apéndice] que el estimador MCO de σ_u^2 es insesgado, podemos asegurar que el estimador de máxima verosimilitud es sesgado:

$$E(\hat{\sigma}_{MV}^2) = E\left(\frac{n-2}{n}\hat{\sigma}_{MCO}^2\right) = \frac{n-2}{n}E(\hat{\sigma}_{MCO}^2) = \frac{n-2}{n}\sigma_u^2$$

Sin embargo, su sesgo desaparece al aumentar el tamaño muestral por cuanto que el factor $(n-2)/n$ tiende a uno. El estimador MV de la varianza es, por tanto, asintóticamente insesgado.

6 Propiedades adicionales del coeficiente de determinación

6.1 Expresión alternativa:

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

6.2 Relación con el coeficiente de correlación lineal en un modelo de regresión lineal simple:

Comenzamos obteniendo una expresión para la Suma de Cuadrados de los Residuos de la estimación de mínimos cuadrados:

$$\begin{aligned} SCR &= \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left[y_i - \left(\bar{y} + \frac{S_{xy}}{S_x^2} (x_i - \bar{x}) \right) \right]^2 = \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + \left(\frac{S_{xy}}{S_x^2} \right)^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2 \frac{S_{xy}}{S_x^2} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \\ &= nS_y^2 + n \frac{S_{xy}^2}{S_x^2} - 2n \frac{S_{xy}^2}{S_x^2} = n \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right) \end{aligned}$$

por lo que el coeficiente de correlación lineal puede escribirse:

$$R^2 = 1 - \frac{SCR}{nS_y^2} = 1 - \frac{S_y^2 - \frac{S_{xy}^2}{S_x^2}}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} = \rho_{xy}^2$$