CAPÍTULO 3-1
EL MODELO LINEAL GENERAL Y SU ESTIMACIÓN CON EL MÉTODO DE LOS MÍNIMOS CUADRADOS

3-1.1 EL MODELO LINEAL EN SU FORMA GENERAL

Para un modelo teórico determinista, la forma general del modelo lineal[139] se define con

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

donde el índice suscrito i designa un individuo en la población o una observación en la muestra, y_i es la variable dependiente y x_{i1}, x_{i2}, ..., x_{ik} son las variables independientes. Los coeficientes β_j son los parámetros desconocidos del modelo que se pretende estimar.

Por lo general, una de las variables independientes y a menudo la primera es una constante: $x_{i1} = 1$ para todo i

Es posible escribir entonces el modelo de la manera siguiente:

[139] Para ser precisos, tendríamos que referirnos a un modelo lineal general con variable dependiente única, puesto que el modelo lineal general puede contener varias variables dependientes.

$$y_i = \sum_{j=1}^{k} \beta_j \, x_{ij} = \beta_1 + \beta_2 \, x_{i2} + \cdots + \beta_k \, x_{ik}$$

Se usa la expresión "constante del modelo" para designar al mismo tiempo la variable independiente cuyo valor es constante (x_{i1}) y el parámetro que se le asocia ((β_1).

3-1.1.1 Ejemplo de un modelo lineal

Pretendemos estudiar la relación que existe entre el tamaño de la ciudad más grande (variable dependiente denotada *PLAR*) y la población total y el PIB per cápita del país (variables independientes denotadas *PTOT* y *GNPC* respectivamente). Uno de los modelos que podríamos considerar, sería:

$$\text{PLAR}_i = \beta_1 + \beta_2 \, \text{PTOT}_i + \beta_3 \, \text{GNPC}_i$$

Si fijamos los valores de los parámetros β_1, β_2 y β_3 y si se conoce la población total y el PIB per cápita de un país, es posible calcular lo que predice el modelo en cuanto a la población de su ciudad más grande. Por ejemplo, supongamos que fijamos[140]

$$\beta_1 = 3500$$
$$\beta_2 = 0.01$$
$$\beta_3 = 0.1$$

Presentamos los datos con relación a Brasil y Costa-Rica en 1990, extraídos de la tabla 1 de Lemelin y Polèse (1995):

[140] Estos valores son cercanos a los valores estimados con el método de los menores cuadrados ordinarios que se aplicó a los datos de 1990 que se presentaron en la tabla 1 de Lemelin y Polèse (1995). Los valores estimados exactos son $\beta_1 = 3431$, $\beta_2 = 0.01324$ y $\beta_3 = 0.09375$. El coeficiente de determinación múltiple de la regresión es de 0.26.

			PLAR (‘000)	*PTOT* (‘000)	*PURB* (‘000)	*GNPC* ($ US)
7	Brasil	Sao Paulo	17395	150368	112643	2680
13	Costa Rica	San José	1016	3015	1420	1900

A partir de estos datos, el modelo predice que, en 1980, la población de Sao Paulo era de, en miles:

$$3500 + (0.01 \times 150368) + (0.1 \times 2680) = 5272$$

y la población de San José,

$$3500 + (0.01 \times 3015) + (0.1 \times 1900) = 3720$$

Es fácil darse cuenta que estas predicciones son de muy mala calidad. La diferencia con los valores observados es de 12 123 en el primer caso y de –2704 en el segundo. Es todavía prematuro concluir que el modelo no sirve con sólo dos observaciones. Sin embargo, podemos sospechar que la relación lineal no es la más adecuada para el fenómeno que se estudia.

En el ejemplo anterior, el modelo cuenta con tres parámetros. Se asocia a cada parámetro una variable independiente. El parámetro β_1 es la constante del modelo, es decir que su variable independiente asociada es una constante. Así que si quisiéramos ser totalmente explícitos, tendríamos que presentan los datos de la forma siguiente:

			Constante	*PLAR* ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7	Brasil	Sao Paulo	1	17395	150368	112643	2680
13	Costa Rica	San José	1	1016	3015	1420	1900

Ahora, se escribe el cálculo de las “predicciones” como sigue:

$$(3500 \times 1) + (0.01 \times 150368) + (0.1 \times 2680) = 5272$$

$$(3500 \times 1) + (0.01 \times 3015) + (0.1 \times 1900) = 3720$$

Como los demás, se multiplica, entonces, el parámetro β_1 por el valor de la variable correspondiente. Es importante

siempre tener en claro que la constante es una de las variables del modelo, particularmente cuando se cuenta el número de variables independientes (en este caso, tres). Es de igual manera importante cuando el modelo cuenta con variables independientes dicotómicas (nombradas variables mudas) con el fin de no introducir redundancias en el modelo (vea capítulo 4-2).

Nota:

Algunos autores escriben

$$y_i = \sum_{j=0}^{h} \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_h x_{ih}$$

Es la razón por la cual al momento de dar el número de variables independientes, es necesario precisar si incluye o no la constante (hay $h + 1 = k$ variables contando la constante). Este detalle es importante al momento de contar el número los grados de libertad que se asocian a algunas variables-test. En este trabajo se incluye siempre la constante en el número de variables independientes (que indicamos con una k por lo general).

Cuando el modelo no tiene más que dos variables independientes incluyendo la constante, se trata de la regresión lineal simple:

$$y_i = \alpha + \beta x_i$$

Se estudiará aquí solamente el caso general de la regresión lineal múltiple cuyo caso particular es la regresión simple.

3-1.1.2 La representación de las relaciones no lineales en el modelo lineal

El modelo lineal general permite representar relaciones no lineales siempre y cuando sean lineales con relación a los parámetros o bien linealizables. Algunos ejemplos nos ayudarán a ilustrar lo que esto significa.

Ejemplo 1: la transformación logarítmica:

La relación exponencial

$$PLAR_i = K\, PURB_i^h$$

es lineal cuando tomamos los logaritmos:

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

Por lo tanto, en estas condiciones, las variables del modelo ya no son *PLAR* y *PURB* sino más bien ln *PLAR* y ln *PURB*.

Lemelin y Polèse (1995) estimaron los parámetros de esta relación; se efectuaron los cálculos con los logaritmos neperianos.[141] Se presentan los resultados en la tabla 2 del artículo: ln K = 2.067 y h = 0.636. A continuación, damos los valores redondeados para Brasil y Costa-Rica en 1990 que se calcularon a partir de la tabla 1 de Lemelin y Polèse (1995):

			ln *PLAR* (‘000)	ln *PURB* (‘000)
7	Brasil	Sao Paulo	9.76	11.63
13	Costa Rica	San José	6.92	7.26

[141] Como es posible observar al momento de aplicar los logaritmos neperianos en la relación exponencial. El valor estimado del parámetro h no tiene influencia de la selección de la base de los logaritmos (el número trascendental e para los logaritmos neperianos o 10 para los logaritmos comunes). La constante estimada, Log K o ln K, depende, sin embargo, de la selección de la base.

A partir de estos datos es posible calcular que el modelo "predice" que, en 1990, la población de Sao Paolo era de, en miles:

$$\text{EXP}[2.067 + (0.636 \times 11.63)] = \text{EXP}(9.46) = 12883$$

y aquella de San José,

$$\text{EXP}[2.067 + (0.636 \times 7.26)] = \text{EXP}(6.68) = 800$$

La linealización de un modelo con su transformación logarítmica es un procedimiento frecuente. Se examinó anteriormente al momento del ajuste de una curva de tendencia (vea 1-2.3):

$$y_t = y_0\,(1+r)^t$$

llega a ser

$$\log y_t = \log y_0 + t \log(1+r)$$

No obstante, en este caso el exponente t es una de las dos variables independientes del modelo (la otra es la constante) y log yt es la variable dependiente mientras que y0 y log(1+ r) son los parámetros que se pretende estimar.

En economía se aplica la transformación logarítmica a la función de producción Cobb-Douglas, que se define con:

$$Y_i = A\,K_i^B\,T_i^C$$

donde:

Y es la cantidad producida;

K es la cantidad de capital empleado;

T es la cantidad de mano de obra empleada.

A, B y C son los parámetros.

Al momento de aplicar la transformación logarítmica, el modelo llega a ser lineal:

$$\log Y_i = \log A + B \log K_i + C \log T_i$$

Ejemplo 2: el añadido de variables independientes:

La relación

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i + d (\ln GNPC_i)^2$$

encierra tres variables independientes (incluyendo la constante) pero una de ellas aparece, al mismo tiempo, en forma lineal y en forma cuadrática. No obstante, es posible tratar esta relación como si fuera una relación lineal. Para esto, sólo basta considerar que GNPC y (ln GNPC)2 son dos variables diferentes. Incluyendo la constante, el modelo cuenta, entonces, con cuatro variables independientes.[142]

Obviamente, se puede generalizar este procedimiento. Así, la relación cúbica

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3$$

llega a ser lineal cuando se define

$$z_{i1} = 1 \text{ (constante)}, z_{i2} = x_i, z_{i3} = x_i^2 \text{ et } z_{i4} = x_i^3$$

Se puede escribir, entonces, el modelo en la forma de una relación lineal:

$$y_i = \beta_1 + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} = \sum_{j=1}^{4} \beta_j z_{ij}$$

Este procedimiento permite, también, linealizar polinomios de cualquier grado, particularmente útiles para estimar superficies de tendencias. La estimación de superficie de tendencia puede servir para describir las variaciones en el espacio de los valores de una variable como, por ejemplo, el precio de las casas. Es efectivamente un modelo descriptivo,[143] puesto que la relación no se basa en ninguna teoría. Los

[142] Se puede comparar este procedimiento con el uso, en el teatro o en el cine, de más de un actor para representar un mismo personaje en edades diferentes.

[143] Del mismo modo que el ajuste de una curva de tendencia temporal es un modelo descriptivo. A menudo se usa un modelo descriptivo como complemento de un modelo teórico (vea, en particular, los trabajos de Francois Desrosiers y Marius Thériault de la Universidad Laval sobre los precios inmobiliarios en la región de Quebec).

datos que se requieren son los precios de venta de los inmuebles y su localización, en coordenadas XY[144]. Un polinomio de segundo grado se define como

$$Z_i = \beta_0 + X_i\,\beta_1 + Y_i\,\beta_2 + X_i^2\,\beta_3 + Y_i^2\,\beta_4 + X_i\,Y_i\,\beta_5$$

Este modelo cuenta con seis variables independientes: la constante, X_i, Y_i, X_i^2, Y_i^2, y X_iY_i.. En un polinomio de tercer grado, tendríamos las cuatro variables suplementarias siguientes: X_i^3, Y_i^3, $X_i^2Y_i$ y $X_iY_i^2$. Cuanto más alto es el grado del polinomio, más compleja puede ser la superficie que describe. Pero, por otro lado, más alto está el número de variables independientes. Veremos cómo el número de parámetros que se puede estimar es limitado por el número de observaciones.

3-1.2 ¿CUÁNDO INTERVIENE LO ALEATORIO?

En el análisis de regresión se busca conocer los parámetros de la relación entre los y_i y los x_{ij}. Ahora bien, si el modelo teórico determinista fuera verdadero, entonces cada observación se conformaría con exactitud al modelo; en estas condiciones, para conocer sus parámetros β_j, sólo bastaría recolectar, con relación a y_i y los x_{ij}, tantas observaciones como haya parámetros y resolver un sistema de k ecuaciones (una para cada observación i) con k desconocidas, o sea los β_j).[145]

---

[144] En un sistema de información geográfica (SIG) se registra la situación de los objetos en el espacio en la forma de coordenadas como la posición de un punto en el plano cartesiano; estas coordenadas son, a veces, la latitud y la longitud geográficas de la posición pero no necesariamente.

[145] Se emplean métodos similares en ciertas circunstancias. Se conoce, entonces, más bien como una "calibración" del modelo en lugar de una "estimación".

Así, en física, la velocidad v de un objeto en caída libre es igual al tiempo transcurrido t multiplicado por la constante de aceleración a[146].

$$v = at$$

De esta relación, se infiere que la distancia recorrida d es proporcional al cuadrado del tiempo de caída:

$$d = \frac{1}{2} a t^2$$

Puesto que la ley de aceleración de la gravedad es determinista, sólo basta una sola observación precisa del cuerpo en caída para conocer el valor de la constante *a*. Al medir *d* y *t*, es fácil calcular el valor de *a*.

De manera idéntica, si el modelo

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

fuera exacto, los valores observados para Brasil y Costa-Rica permitirían definir un sistema de dos ecuaciones lineales de dos desconocidas:

$$9.76 = \ln K + 11.63\ h \text{ (Brasil)}$$

$$6.92 = \ln K + 7.26\ h \text{ (Costa Rica)}$$

La solución de este sistema es

$$\ln k = 2.20$$

$$h = 0.65$$

¡Estaríamos en la gloria! Sin embargo, estamos concientes que los modelos, y particularmente en ciencias sociales, son demasiados simples, aún más si son lineales, para representar toda la complejidad de lo real. Nuestros modelos teóricos no son más que aproximaciones y, aunque sean buenos, solo de manera aproximativa, las observaciones se conforman a ellos. De modo que, si estimáramos los k parámetros con la ayuda de k observaciones, sería muy probable que una nueva observación (k + 1) fuera incompatible con el modelo (desde un

[146] La constante de aceleración de la gravedad es igual a 980.621 cm/s^2, o sea 32.1725 pies/s^2, al nivel del mar.

punto de vista determinista); dicho de otra manera, la $(k + 1)^{iesima}$ ecuación sería contradictoria con las demás.[147]

Por ejemplo, añadamos a las observaciones efectuadas sobre Sao Paolo y San José los datos relativos a Toronto (calculados a partir de la tabla 1 de Lemelin y Polèse, 1998):

			ln PLAR ('000)	ln PURB ('000)
7	Brasil	Sao Paulo	9.76	11.63
9	Canada	Toronto	8.15	9.93
13	Costa Rica	San Jose	6.92	7.26

Si aplicamos a Toronto los coeficientes que se calcularon anteriormente, obtenemos:

$$\ln PLAR = 2.20 + 0.65 \ln PURB$$

$$\ln PLAR = 2.20 + (0.65 \times 9.93) = 8.65 \neq 8.15$$

No se verifica la ecuación en el caso de Toronto. De hecho, sabemos que no existe solución para el sistema siguiente de tres ecuaciones y dos desconocidas:

$$9.76 = \ln K + 11.63\ h \text{ (Brésil)}$$

$$8.15 = \ln K + 9.93\ h \text{ (Can.)}$$

$$6.92 = \ln K + 7.26\ h \text{ (Costa Rica)}$$

Generalizando, con una muestra de n observaciones y k parámetros para estimar (uno para cada variable independiente), es posible construir un sistema de n ecuaciones con k desconocidas. En caso que el número de observaciones sea superior al número de parámetros para estimar, el número de ecuaciones es, entonces, superior al número de desconocidas.

[147] En las ciencias llamadas exactas como la física, es común enfrentarse con un problema muy similar. Los errores de medición introducen un elemento de inexactitud en las observaciones con que, lo mismo cuando los modelos son leyes "deterministas", subsiste un cierto grado de imprecisión con relación a los valores de los parámetros (como en el caso de la constante de aceleración de la gravedad).

Ahora bien, usualmente, un sistema de este tipo no tiene solución porque las ecuaciones son incompatibles entre sí.

Por lo tanto, aunque un modelo sea una buena aproximación de la realidad, subsiste, no obstante, una diferencia entre las predicciones del modelo y las observaciones. La ausencia entre las variables independientes de numerosos factores secundarios cuya influencia es pequeña (modelo incompleto y demasiado simple), es parte de la explicación de esta diferencia. Esta situación se traduce por un "error" que no parece ser sistemático sino, más bien, fruto del azar. Con el fin de tomar en cuenta este error, se añade una variable aleatoria en el modelo teórico:

$$y_i = \sum_{j=1}^{k} \beta_j \, x_{ij} + u_i$$

El término aleatorio u_i es, también, conocido como término de error, error estocástico o simplemente error o bien perturbación (disturbance term). Es importante entender que los valores que tome el término aleatorio son igual de inobservables que los parámetros de la relación, todo lo que podemos observar son los valores de la variable dependiente y las variables independientes.

Por ejemplo, el modelo

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

es un modelo teórico determinista. Sin embargo, el modelo en el cual se basa la estimación de los parámetros y los tests de hipótesis que aparecen en Lemelin y Polèse (1995) es, en realidad,

$$\ln PLAR_i = \ln K + h \ln PURB_i + u_i \,,$$

donde u_i es un término aleatorio.

Así, acabamos de descubrir la tercera "puerta" por la cual se introduce lo aleatorio en el análisis de regresión. Describimos estas tres puertas como sigue:

[...] existen tres 'puertas' por las cuales se introduce lo aleatorio en los modelos:

1. Para empezar, existe la naturaleza aleatoria que ya mencionamos del vínculo entre una muestra y la población de donde se obtuvo.
2. Las variables operacionales son medidas imperfectas de los conceptos y se puede considerar que el error de medición es aleatorio (o sea que se determina al azar). Es posible por lo tanto, representar con un modelo aleatorio la influencia de los errores de medición que intervienen al momento de traducir las hipótesis teóricas en hipótesis operacionales (entre los primeros modelos aleatorios, justamente hay que mencionar los modelos de la "teoría de los errores" en ciencias físicas).
3. Finalmente, algunos fenómenos son, por naturaleza, aleatorios y no pueden representarse adecuadamente con modelos teóricos no aleatorios. En estos modelos, el azar es el reflejo de, por un lado, una indeterminación fundamental (como en física de las partículas) y, por el otro, una multitud de factores inobservables (como suele suceder en ciencias sociales[148]) cuyas manifestaciones aparecen como reglas gracias a leyes de probabilidades". (Capítulo 2.2).[149]

[148] Pensemos, en particular, en los modelos de utilidad aleatoria (random utility) subyacentes a los modelos de selecciones discretas (discrete choice) logit, probit, etc. Vamos a encontrar este tipo de modelos en el apartado 4-3.

[149] Este pasaje es inspirado de Malinvaud, quien escribe: "Se sabe que se justifica el uso del cálculo de las probabilidades para el análisis de los datos de estadística con una u otra de las dos consideraciones siguiente. O bien, se asimila el fenómeno estudiado como un proceso que encierra una determinación aleatoria de algunas magnitudes; en este caso, se consideran, entonces, las magnitudes como aleatorias en el universo (NDLA: o sea en la población) así como en la muestra observada. O bien, la selección de las unidades observadas es el resultado de un sorteo aleatorio; la composición de la muestra es, entonces aleatoria y, por consiguiente, los datos obtenidos también aunque sean datos sobre magnitudes no aleatorias" (Malinvaud,

Según esta concepción, aunque los datos englobaran la totalidad de la población estudiada, el elemento aleatorio no desaparecería puesto que el aspecto aleatorio se debe no tanto por la relación entre la población y la muestra sino más bien, por la relación entre el modelo determinista (la ley matemática), cuyos parámetros son desconocidos, y las observaciones las cuales se alejan del modelo de manera aleatoria:[150] así, las observaciones dejan de ser incompatibles con el modelo para ser simplemente, desde el enfoque de la probabilidad, más o menos compatibles con el modelo. Agreguemos, sin embargo, que, en este contexto, la distinción entre población y muestra sigue subsistiendo pero esta distinción vale, primeramente, por el hecho que los valores que toman los términos aleatorios inobservables se sortean de la población infinita de los valores que el proceso aleatorio subyacente a cada uno de los términos aleatorios podría generar. Esto último nos permite entender que es posible que se engendren los valores de los términos aleatorios asociados a diferentes observaciones gracias a procesos aleatorios diferentes. Es para ilustrar esto que, en algunos contextos, se mencionan los términos aleatorios en plural.

Por lo tanto, en un primer nivel, la combinación de un término aleatorio y un modelo determinista permite acomodarse con el carácter aproximativo del acuerdo entre el modelo y las observaciones. Para ir más allá, es necesario caracterizar las distribuciones de probabilidad de los términos aleatorios u_i. Tendremos, entonces, un modelo aleatorio y se-

---

1969, p. 62). Malinvaud prosigue diciendo que el primer tipo de justificación le parece más apropiado en el contexto de la econométría.

[150] Hay algo de la caverna de Platón en esta concepción (ver en el anexo el texto de la alegoría). Tratamos con la realidad observable como si fuera el reflejo imperfecto (la sombra proyectada) del modelo teórico determinista (lo ideal). El término aleatorio del modelo representa las imperfecciones de la realidad observable. La inducción estadística busca discernir lo "ideal" (en el sentido que le daba Platón a esta palabra) a través de su reflejo.

remos capaces de aplicar los métodos de inducción estadística con el fin, en particular:

- de estimar los parámetros de las distribuciones de probabilidad de los términos aleatorios;
- de estimar los parámetros de las distribuciones de muestreo de los estimadores;
- de efectuar unos tests de hipótesis.

3-1.3 EL ESTIMADOR DE LOS MINIMOS CUADRADOS ORDINARIOS

Para complementar la simbología empleada, se conviene lo que sigue:

b_i: valor estimado del parámetro β_i.

$\hat{y}_i$: valor de y_i "predicho" o calculado por el modelo tal como se estimó.

Tenemos por definición:

$$\hat{y}_i \equiv \sum_j b_j x_{ij}$$

e_i: residuo calculado (o "error") de la regresión para la i^{esima} observacion

Tenemos por definición:

$$e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$$

NB: No se debe confundir e_i, el residuo calculado (observable), con el término aleatorio correspondiente u_i inobservable.

3-1.3.1 Definición

Aunque no se haya complementado la especificación del modelo aleatorio, es posible aplicar el método de los mínimos

cuadrados (vea el enunciado de este principio en el apartado 2-2.3). Sólo es necesario reconocer que el modelo no es más que una aproximación y que las observaciones no se conforman más que aproximadamente a él.

El principio de los mínimos cuadrados consiste en escoger los valores estimados b_j que minimizan la suma de los cuadrados de los residuos (o "errores"). Esto significa minimizar la suma de los cuadrados de las diferencias entre los valores observados de y_i y los valores "predichos" $\hat{y}_i$:

$$\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i \left(y_i - \sum_j b_j \, x_{ij} \right)^2 = \sum_i e_i^2$$

La expresión $\sum_i \left(y_i - \hat{y}_i\right)^2$ es, por lo tanto, el cuadrado de la distancia euclidiana generalizada entre los valores observados de la variable dependiente y los valores predichos. Se define la solución de este problema de minimización con el estimador de los mínimos cuadrados ordinarios.

Se presentan, con frecuencia, los resultados de la estimación en unas tablas: vea las tablas 2 y 4 de Lemelin y Polèse (1995), la tabla 1 de Heikkila et al. (1989) o la tabla 1 de Richardson et al. (1990).

3-1.3.2 Algunas propiedades del estimador de los mínimos cuadrados ordinarios

Estimador lineal

Este estimador es lineal, es decir que se calcula cada b_j como una función lineal de los y_i, o con más exactitud, como una suma ponderada de los y_i:

$$b_j = \sum_i w_{ji}\, y_i$$

donde cada uno de los coeficientes w_{ji} depende del conjunto de los xgh.[151]

Suma de los residuos nula

Cuando el modelo de regresión tiene una constante, como en la mayoría de los casos, la suma de los residuos la regresión es nula:

$$\sum_i e_i = 0$$

No se exhibe aquí la demostración porque se necesita, para el efecto, la escritura matricial.

Relación entre los promedios

Cuando el modelo tiene una constante, el promedio de los valores predichos es igual al valor predicho a partir de los valores promedios de las variables independientes y estos dos valores son iguales al valor promedio observado de la variable dependiente:

$$m_y = m_{\hat{y}} = \sum_j b_j\, m_{x_j}$$

Se deduce esta propiedad de la previa.

Demostración:

Sabemos que $\sum_i e_i = 0$

[151] Para ser más preciso, w_{ji} es el elemento j,i de la matriz $(X'X)^{-1}X'$. Observe que el hecho que el estimador sea lineal no es una consecuencia de que el modelo sea lineal.

Ahora bien $e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$

Esto implica, en particular, que la suma de los valores "predichos" es igual a la suma de los valores observados:

$$\sum_i y_i - \sum_i \hat{y}_i \equiv \sum_i (y_i - \hat{y}_i) \equiv \sum_i e_i \equiv 0$$

$$\sum_i y_i = \sum_i \hat{y}_i$$

Pero, puesto que

$$\sum_i \hat{y}_i \equiv \sum_i \left(\sum_j b_j x_{ij} \right) \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

entonces

$$\sum_i y_i = \sum_i \hat{y}_i$$

implica que

$$\sum_i y_i \equiv \sum_i \hat{y}_i \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

$$\left(\frac{1}{n}\right) \sum_i y_i \equiv \left(\frac{1}{n}\right) \sum_i \hat{y}_i \equiv \left(\frac{1}{n}\right) \sum_j b_j \left(\sum_i x_{ij} \right)$$

$$\left(\frac{1}{n}\right) \sum_i y_i \equiv \sum_j b_j \left[\left(\frac{1}{n}\right) \sum_i x_{ij} \right]$$

Ahora bien

$$m_y = \left(\frac{1}{n}\right) \sum_i y_i$$

$$m_{\hat{y}} = \left(\frac{1}{n}\right) \sum_i \hat{y}_i$$

$$m_{x_j} = \left(\frac{1}{n}\right) \sum_i x_{ij}$$

Por lo tanto, tenemos

$$m_y = m_{\hat{y}} = \sum_j b_j \, m_{x_j}$$

3-1.4 EL COEFICIENTE DE DETERMINACIÓN MÚLTIPLE Y EL ANÁLISIS DE LA VARIANZA

3-1.4.1 Construcción del coeficiente de determinación múltiple

El coeficiente de determinación múltiple es una medición de asociación que pertenece a la familia de las mediciones de similitud; con más precisión, es una medición del grado de acuerdo entre el modelo y las observaciones. En estadística, una medición de este tipo se llama "medición de ajuste" (goodness of fit measure).

El coeficiente de determinación múltiple se basa en un análisis de descomposición[152] de la variabilidad de la variable dependiente donde se calcula esta variabilidad con la suma de los cuadrados de las desviaciones con relación a la media:

$$\sum_i \left(y_i - m_y\right)^2 = (n-1)\, s_y^2$$

En estadística, este tipo de análisis de descomposición se llama una "análisis de varianza".[153]

---

[152] Sobre el análisis de descomposición, vea 1-2.2.

[153] Hay una forma más especializada de análisis de varianza que permite examinar la relación entre una variable dependiente y varias variables independientes categóricas, descomponiendo la varianza de la variable dependiente entre la varianza dentro los grupos (definididos por combinaciones de categorías de las variables independientes) y la varianza entre los grupos. Abordaremos este tema en el apartado 4-2.

Primera etapa: descomposición de la variabilidad

Cuando el modelo tiene una constante, es posible descomponer la variabilidad en dos componentes:[154]

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2$$

Principio de demostración:

$$\sum_i \left(y_i - m_y\right)^2 = (n-1)\, s_y^2$$

Si desarrollamos el miembro de la izquierda de esta expresión, obtenemos:

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left[\left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - m_y\right)\right]^2$$

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2 + 2\sum_i \left(y_i - \hat{y}_i\right)\left(\hat{y}_i - m_y\right)$$

Se puede mostrar que, si el modelo tiene una constante, el último término es nulo[155], de tal manera que:

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2$$

[154] En caso de que no haya constante, la descomposición ya no es válida. Puede hasta ocurrir que R^2 sea negativo.

[155] Se requiere de la escritura matricial para esta demostración.

Segunda etapa: interpretación de los elementos de la descomposición

En la expresión $\sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2$, el segundo término es una medición de la variabilidad de los valores predichos por el modelo: es la parte "explicada" de la variabilidad. El primer término es una medición de la variabilidad de los residuos: es la parte de la variabilidad que el modelo no prevé. Por lo tanto, tenemos:

Variabilidad total	=	Variabilidad residual	+	Variabilidad "explicada"

De esta interpretación, se infiere la simbología siguiente la cual se usa frecuentemente en las salidas de los paquetes de aplicaciones estadísticas:[156]

- *SST*: Suma de los Cuadrados Totales (*Sum of Squares Total*)
  $$= \sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$
- *SSM*: Suma de los Cuadrados del Modelo (*Sum of Squares Model*),
  ya que, cuando existe una constante, $m_y = m_{\hat{y}}$,
  $$= \sum_i (\hat{y}_i - m_{\hat{y}})^2 = \sum_i (\hat{y}_i - m_y)^2 = (n-1)\, s_{\hat{y}}^2$$

[156] No obstante, tenga cuidado porque es posible encontrar la simbología siguiente: *SSR* por *Sum of Squares Regression* en lugar de *SSM* y *SSE* por *Sum of Squares Errors* en lugar de *SSR*.

- *SSR*: Suma de los Cuadrados de los Residuos (*Sum of Squares Residuals*)

$$= \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$$

En resumen:

Variabilidad total		Variabilidad residual		Variabilidad "explicada"
SST		*SSR*		*SSM*
Sum of Squares, Total		Sum of Squares, Residuals		Sum of Squares, Model
$\sum_i \left(y_i - m_y\right)^2 = (n-1)s_y^2$	=	$\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$	+	$\sum_i \left(\hat{y}_i - m_{\hat{y}}\right)^2 = \sum_i \left(\hat{y}_i - m_y\right)^2 = (n-1)s_{\hat{y}}^2$

Tercera etapa: construcción de una medición de ajuste ("goodness of fit")

El coeficiente de determinación múltiple es la parte de la variabilidad "explicada" en la variabilidad total:

$$R^2 = \frac{\text{Variabilidad «explicada»}}{\text{Variabilidad total}} = \frac{SSM}{SST}$$

O bien, puesto que $SST = SSR + SSM$, tenemos $SSM = SST - SSR$ y

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{(n-1)s_y^2}$$

Recordemos que el método de los mínimos cuadrados consiste en tomar, como valores estimados de los parámetros, los valores de los coeficientes que minimizan $\sum_i e_i^2$. A la

luz de la formula enunciada arriba, entendemos que usar el método de los mínimos cuadrados es lo mismo que escoger los valores de los coeficientes que maximizan R^2 bajo la especificación, es decir en el marco del modelo seleccionado.[157]

> El valor del coeficiente de determinación múltiple es, por lo general, presentado en las tablas de resultados: vea las tablas 2 y 4 de Lemelin y Polèse (1995).

3-1.4.2 Campo de variación del coeficiente de determinación múltiple (valores extremos)

El coeficiente de determinación varía entre cero y uno. En efecto, matemáticamente hablando, *SST*, *SSM* y *SSR* son sumas de cuadrados y, por consiguiente, su valor no puede ser negativo. Además, *SST* = *SSR* + *SSM*, lo que nos permite deducir que ni *SSR*, ni *SSM* pueden exceder el valor de *SST*. Finalmente, puesto que $R^2 = \frac{SSM}{SST}$, se deduce de lo anterior que el coeficiente de determinación no puede ser inferior a cero o superior a uno. Examinemos ahora en que circunstancias R^2 podría alcanzar estos valores extremos.

El coeficiente de determinación es igual a uno cuando SSR = 0, es decir, cuando el modelo reproduce perfectamente las observaciones que sirvieron para estimar los parámetros de este mismo modelo. Es igual a cero cuando SSR = SST, es decir, cuando SSM = 0. Pero, ¿en qué circunstancias es posible tener SSM = 0? Bueno, para empezar, SSM es una suma de cuadrados:

---

[157] Como lo veremos más tarde, esto no es lo mismo que comparar los R^2 de diferentes modelos después de haber estimado los parámetros de cada uno de ellos con el propósito de obtener para cada modelo el R^2 más elevado posible con el método de los menores cuadrados.

$$SSM = \sum_i \left(\hat{y}_i - m_y\right)^2$$

Por lo tanto, sólo es posible tener $SSM = 0$ cuando todos los términos de la suma son nulos, o sea, si $\hat{y}_i = m_y$ para cada observación *i*. ¿Cómo puede ocurrir esto? Es posible mostrar que ocurre esta situación cuando, a partir de un modelo general

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

los coeficientes estimados con el método de los mínimos cuadrados son todos nulos con excepción de la constante. En estas condiciones,

$$b_2 = b_3 = ... = b_k = 0$$

y tenemos

$$\hat{y}_i = b_1 = m_y$$

donde m_y es un valor de b1 como se estimó con el método de los mínimos cuadrados ordinarios en este caso.

En resumen, que el coeficiente de determinación sea nulo muestra que no es posible detectar una relación entre la variable dependiente y las variables independientes; de hecho, al momento de estimar los parámetros del modelo, todas las variables independientes desaparecen con excepción de la constante, porque se multiplican por un coeficiente cuyo valor se estima en cero.

¿Es realmente el coeficiente de determinación múltiple una medición de similitud, como lo afirmamos al principio? Para convencerse, solo basta ver que SSR es el cuadrado de la distancia euclidiana generalizada entre el conjunto de los valores observados y el conjunto de valores predichos por el modelo. Es, por lo tanto, una medición de disimilitud. La razón $\frac{SSR}{SST}$ es, por consiguiente, una medición de disimilitud

normada cuyo campo de variación se extiende de cero a uno. Así que $R^2 = 1 - \frac{SSR}{SST} = \frac{SSM}{SST}$ es una medición de similitud cuyo campo de variación se extiende también de cero de uno.

3-1.4.3 Relación entre R^2 y el coeficiente de correlación simple

Es posible mostrar que el coeficiente de determinación, R^2, es igual al cuadrado del coeficiente de correlación simple entre los valores observados y_i y los valores predichos $\hat{y}_i$.

$$r_{\hat{y}y}^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y} \right)^2 = R^2$$

3-1.4.4 Coeficiente de determinación ajustado

Cuando se respetan las hipótesis clásicas (se definen estas hipótesis más abajo), el coeficiente de determinación ajustado

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) = 1 - \frac{SSR \big/ (n-k)}{SST \big/ (n-1)}$$

es un estimador no sesgado del "verdadero" coeficiente de determinación.

El coeficiente de determinación ajustado puede interpretarse como un modo de tomar en cuenta el número de variables independientes en la evaluación del desempeño de un modelo. En efecto, es posible, por lo general, aumentar el coeficiente de determinación R^2 con solo añadir variables independientes en el modelo aunque la presencia de variables suplementarias no se base en una hipótesis teórica.

Podemos observar en la formula del coeficiente de determinación ajustado $\overline{R}^2$ que, al momento de añadir variables, $\overline{R}^2$ puede disminuir a condición que R^2 no aumente lo suficiente para compensar el incremento de k. Existen, sin embargo, otros procedimientos aun más fiables para decidir hasta que punto es oportuno añadir o quitar tal o tal variable: estos otros procedimientos son los tests de hipótesis.

Se presenta, por lo general, el valor del coeficiente de determinación ajustado en las tablas de resultados vea las tablas 2 y 4 de Lemelin y Polèse (1995), la tabla 1 de Heikkila et al. (1989) o la tabla 1 de Richardson et al. (1990).