Errores de especificación, multicolinealidad y observaciones atípicas

Tomás del Barrio Castro
Miquel Clar López
Jordi Suriñach Caralt

PID_00160617

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Ya hemos presentado el modelo de regresión lineal múltiple (MRLM) estándar, el cual cumple unas hipótesis básicas que ya hemos estudiado. En este módulo analizaremos lo que sucede con respecto al modelo de regresión cuando se incumplen algunas de aquellas hipótesis. Por lo tanto, dejamos el modelo estándar y pasamos a otro modelo en el que deberemos valorar las consecuencias de estos incumplimientos. En concreto, estudiaremos lo siguiente:

Consultad el MRLM estándar y las hipótesis básicas de este modelo en el apartado 2 del módulo "Modelo de regresión lineal múltiple: especificación, estimación y contraste" de esta asignatura.

1) En el apartado 1 veremos lo que sucede si se presentan errores de especificación en las variables explicativas.

2) Posteriormente, en el apartado 2, veremos los problemas que se plantean en el MRLM cuando éstos se deben a la muestra de observaciones disponibles para realizar el análisis. Estudiaremos dos problemas. En primer lugar, el problema de la correlación entre las variables explicativas. En segundo lugar, la presencia de observaciones atípicas.

Objetivos

En este módulo el estudiante dispone de los materiales didácticos necesarios para alcanzar los objetivos siguientes:

1. Conocer las propiedades de las estimaciones de mínimos cuadrados ordinarios ante una especificación errónea del modelo.

2. Conocer las diferencias y las analogías entre las consecuencias de un problema de omisión de variables relevantes y uno de inclusión de variables irrelevantes.

3. Saber detectar un problema de mala especificación, tanto por lo que respecta al conjunto de regresores como a la forma funcional del modelo.

4. Conocer las consecuencias negativas asociadas a los modelos con diferentes grados de multicolinealidad.

5. Saber detectar si existe multicolinealidad o no, así como cuáles son las variables que la generan.

6. Seleccionar la mejor alternativa (en cuanto a especificación del modelo que presenta multicolinealidad) para alcanzar los objetivos inicialmente deseados del modelo econométrico.

7. Saber detectar cuándo una observación es atípica, presenta apalancamiento, tiene una influencia en el ajuste mayor que la del resto o es un *outlier*.

8. Conocer las características de los distintos tipos de observaciones mencionados en el punto anterior, así como las consecuencias que producen sobre la estimación del modelo.

1. Errores de especificación

El objetivo de este apartado es analizar las consecuencias para la estimación por mínimos cuadrados ordinarios (MCO) ante el incumplimiento de una de las hipótesis básicas del modelo de regresión lineal múltiple (MRLM) estándar, como la especificación correcta de la parte sistemática o determinista del modelo. Si el modelo está correctamente especificado, es decir, si se cumplen todas las hipótesis básicas, los estimadores MCO* de los parámetros β_j no tienen sesgo, son eficientes y consistentes en error cuadrático medio, y el estimador de la varianza del término de perturbación es no sesgado. No obstante, ¿qué ocurre con estas propiedades ante errores de especificación? Trataremos este problema en los subapartados siguientes.

Consultad el método de estimación por mínimos cuadrados ordinarios y recordad también que la parte determinista del modelo es **XB** en los subapartados 2.3 y 2.1, respectivamente, del módulo "Modelo de regresión múltiple: especificación..." de esta asignatura.

*** Abreviamos método de *mínimos cuadrados ordinarios* con la sigla *MCO*.**

1.1. Errores de especificación en las variables explicativas

Los **errores de especificación** hacen referencia a cualquier error que se cometa en el conjunto de hipótesis en que se apoya el modelo de regresión, tanto si son las que afectan a los regresores* como si son las que afectan al término de perturbación. Sin embargo, se acostumbra a utilizar el término anterior en un sentido más reducido, para referirse a los errores cometidos en la especificación de la parte sistemática del modelo.

*** Los regresores son las variables explicativas que se usan para especificar el MRLM.**

Así pues, diremos que hay un error en la especificación del modelo si hay un problema de:

a) Omisión de variables relevantes: se comete error cuando se incluyen menos variables explicativas de las que se deberían incluir.

La omisión de variables relevantes...

... puede deberse a causas muy heterogéneas, desde la falta de datos fiables sobre una determinada variable explicativa hasta el hecho de ignorar qué es relevante.

b) Inclusión de variables irrelevantes: se comete error cuando se añaden una o más variables explicativas que son irrelevantes en la explicación de la variable endógena.

c) Forma funcional incorrecta: se da cuando la relación especificada entre la variable endógena y las variables explicativas* no es correcta.

*** La relación especificada entre la variable endógena y las variables explicativas se supone lineal por hipótesis del modelo.**

1.2. Problemas asociados a la omisión de las variables relevantes

En este subapartado estudiaremos el primero de los tres casos a que hemos hecho referencia anteriormente. Así, en primer lugar plantearemos el problema de la omisión de variables relevantes, después estudiaremos sus consecuencias sobre los estimadores MCO y acabaremos con algunos ejemplos aclaratorios.

Consultad los tres casos de errores de especificación de las variables explicativas del modelo en el subapartado 1.1 de este módulo didáctico.

1.2.1. Especificación del modelo

Estamos ante un **error de especificación por omisión de variables relevantes** cuando hay como mínimo una variable explicativa que es significativa en la explicación del comportamiento de la variable endógena de interés que no aparece incluida en el modelo especificado.

De este modo se especifica (erróneamente) el modelo $\mathbf{Y} = \mathbf{X}^*\mathbf{B}^* + \mathbf{U}^*$ en lugar del verdadero $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, y la matriz de variables exógenas erróneamente especificadas, $\mathbf{X}^*$, es de rango inferior respecto a la matriz de variables explicativas del modelo correctamente especificado, $\mathbf{X}$, dado que el número de columnas $\mathbf{X}^*$ es menor que el de $\mathbf{X}$.

Si suponemos que el número correcto de variables explicativas que se tendrían que considerar en el modelo es k, sin pérdida de generalidad, podemos definir las matrices y los vectores siguientes:

$$\mathbf{X_1} = \begin{bmatrix} 1 & X_{21} & \dots & X_{r1} \\ 1 & X_{22} & \dots & X_{r2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2N} & \dots & X_{rN} \end{bmatrix}, \quad \mathbf{X_2} = \begin{bmatrix} X_{(r+1)1} & \dots & X_{k1} \\ X_{(r+1)2} & \dots & X_{k2} \\ \vdots & & \vdots \\ X_{(r+1)N} & \dots & X_{kN} \end{bmatrix},$$

$$\mathbf{B_1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{bmatrix}, \quad \mathbf{B_2} = \begin{bmatrix} \beta_{r+1} \\ \beta_{r+2} \\ \vdots \\ \beta_k \end{bmatrix}$$

donde en la matriz $\mathbf{X}_1$ se agrupan las primeras r variables explicativas, mientras que en la matriz $\mathbf{X}_2$ se agrupa el resto (las variables omitidas): $k - r$. Para simplificar la notación, a partir de ahora llamaremos $\mathbf{s}$ a la diferencia $k - r$. Además, $\mathbf{B}_1$ agrupa a los parámetros asociados a las variables de la matriz $\mathbf{X}_1$ y $\mathbf{B}_2$, los asociados a las variables de $\mathbf{X}_2$. Escribiremos $\mathbf{X} = [\mathbf{X_1} \mid \mathbf{X_2}]$ y $\mathbf{B}' = [\mathbf{B_1'} \mid \mathbf{B_2'}]$ para indicar la partición de la matriz $\mathbf{X}$ y del vector $\mathbf{B}'$.

Por lo tanto, $\mathbf{X}_1$ es de dimensión $N \times r$ y $\mathbf{X}_2$, de dimensión $N \times s$. Por tanto, las dimensiones de $\mathbf{B}_1$ y $\mathbf{B}_2$ son r y s, respectivamente.

Así pues, teniendo en cuenta la notación que acabamos de establecer, podemos escribir el modelo correctamente especificado (donde se consideran las k variables exógenas) de la manera siguiente:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} = [\mathbf{X_1} \mid \mathbf{X_2}] \begin{bmatrix} \mathbf{B_1} \\ - \\ \mathbf{B_2} \end{bmatrix} + \mathbf{U} = \mathbf{X_1B_1} + \mathbf{X_2B_2} + \mathbf{U}, \qquad (1.1)$$

y el modelo erróneamente especificado por omisión de variables relevantes (donde se consideran únicamente las primeras r variables explicativas) lo escribimos del modo siguiente:

$$\mathbf{Y} = \mathbf{X}^*\mathbf{B}^* + \mathbf{U}^* = \mathbf{X}_1\mathbf{B}_1 + \mathbf{U}^*. \qquad (1.2)$$

1.2.2. Propiedades de los estimadores $\hat{\mathbf{B}}^*$

En primer lugar, definimos el vector de estimadores MCO que obtendríamos si trabajásemos con el modelo con omisión de variables relevantes 1.2:

$$\hat{\mathbf{B}}^* = \hat{\mathbf{B}}_1 = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_r \end{bmatrix} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\,\mathbf{X}^{*\prime}\mathbf{Y} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} =$$

$$= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{XB} + \mathbf{U}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{XB} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{U}. \qquad (1.3)$$

En la expresión 1.3...

... hemos sustituido $\mathbf{Y}$ por $\mathbf{XB} + \mathbf{U}$, dado que esta igualdad es la que establece el modelo verdadero.

A continuación analizamos las propiedades de este vector con un poco de detalle.

Sesgo

Para ver si el vector de estimadores obtenido en 1.3 es o no un **vector de estimadores sesgados**, hay que analizar su valor esperado. Si calculamos la esperanza matemática de la expresión 1.3 y consideramos las hipótesis básicas del modelo, y en particular la que postula que $E[\mathbf{U}] = 0$, el valor esperado del segundo sumando de la expresión anterior se anula. Si, además, se sustituyen la matriz $\mathbf{X}$ y el vector $\mathbf{B}$ por sus verdaderas expresiones, llegamos a lo siguiente:

$$E[\hat{\mathbf{B}}^*] = E[\hat{\mathbf{B}}_1] = \begin{bmatrix} E[\hat{\beta}_1] \\ E[\hat{\beta}_2] \\ \vdots \\ E[\hat{\beta}_r] \end{bmatrix} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'[\mathbf{X}_1 \mid \mathbf{X}_2] \begin{bmatrix} \mathbf{B}_1 \\ - \\ \mathbf{B}_2 \end{bmatrix} =$$

$$= [(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1 \mid (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2] \begin{bmatrix} \mathbf{B}_1 \\ - \\ \mathbf{B}_2 \end{bmatrix} = [\mathbf{I}_r \mid (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2] \begin{bmatrix} \mathbf{B}_1 \\ - \\ \mathbf{B}_2 \end{bmatrix} =$$

$$= \mathbf{B}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{B}_2.$$

El resultado obtenido pone de manifiesto que los estimadores $\hat{\mathbf{B}}_1$ son sesgados, dado que:

$$E[\hat{\mathbf{B}}_1] = \mathbf{B}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{B}_2 \neq \mathbf{B}_1, \qquad (1.4)$$

siendo el sesgo:

$$\text{sesgo } [\hat{\mathbf{B}}_1] = E[\hat{\mathbf{B}}_1] - \mathbf{B}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{B}_2. \qquad (1.5)$$

Así, para un estimador concreto del vector de estimadores $\hat{\mathbf{B}}_1$, su valor esperado es el siguiente:

$$E[\hat{\beta}_j] = \beta_j + \phi_{j(r+1)}\beta_{r+1} + \phi_{j(r+2)}\beta_{r+2} + \ldots + \phi_{jk}\beta_k \quad \forall j = 1, 2, \ldots, r, \qquad (1.6)$$

donde los elementos $\phi_{j(r+1)}, \phi_{j(r+2)}, \ldots, \phi_{jk}$ son los elementos de la fila *j*-ésima de la matriz de dimensión $r \times s$, resultante de hacer $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$.

Factores de interés en el sesgo del estimador

De 1.4 ó 1.6 observamos que el sesgo del estimador depende de los dos elementos siguientes:

1) La magnitud de los parámetros asociados a las variables explicativas omitidas, es decir, de los parámetros $\beta_j \quad j = r + 1, \ldots, k$ en el modelo correctamente especificado (los parámetros del vector $\mathbf{B}_2$). Cuanto mayor sea la magnitud de estos parámetros, mayor será la relevancia de las variables erróneamente omitidas para explicar la variable endógena y, por tanto, mayor será el sesgo.

2) La correlación entre las variables explicativas incluidas en el modelo ($\mathbf{X}_1$) y las omitidas ($\mathbf{X}_2$). Cuanto mayor sea esta correlación, mayor será el sesgo.

Podéis observar que los ϕ_j son los elementos (por filas) de la matriz resultante de realizar la operación $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$. Es decir, son los r estimadores MCO de cada una de las regresiones en que las variables endógenas son las variables de la matriz $\mathbf{X}_2$ (las variables relevantes omitidas) y las variables explicativas son las variables de la matriz $\mathbf{X}_1$ (las variables consideradas en el modelo erróneamente especificado):

$$\mathbf{X}_2 = \mathbf{X}_1\Phi + \mathbf{V} \quad \rightarrow \quad \hat{\Phi} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2.$$

Un caso extremo de sesgo...

... se daría cuando todos los parámetros de $\mathbf{B}_2$ fuesen cero; en este caso, el sesgo también sería cero, pero obviamente ya no habría problema de omisión de variables.

En el caso extremo...

... de ortogonalidad entre $\mathbf{X}_1$ y $\mathbf{X}_2$, el sesgo será cero.

Matriz de varianzas y covarianzas

En lo referente a la **matriz de varianzas y covarianzas de los estimadores del modelo erróneamente especificado**, tenemos:

$$\text{VAR}[\hat{\mathbf{B}}^*] = \text{VAR}[\hat{\mathbf{B}}_1] = \sigma_u^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}. \qquad (1.7)$$

Puede demostrarse que los elementos de la diagonal principal de la matriz 1.7 serán menores, o como máximo iguales, que los que se habrían obtenido por los mismos estimadores con el modelo correctamente especificado. Sólo serán iguales si las variables omitidas son ortogonales a las variables consideradas en el modelo, es decir, si $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}_{r\times s}$.

Consultad la matriz de varianzas y covarianzas del modelo correctamente especificado en el subapartado 2.3.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Magnitud de los elementos de la diagonal principal de la matriz 1.7

Intuitivamente, se puede entender que los elementos de la diagonal principal de la matriz 1.7 son menores que los que se obtienen por los mismos estimadores con el modelo correctamente especificado, si el modelo erróneamente especificado con omisión de variables se interpreta como un modelo restringido donde se supone que los parámetros asociados a las variables omitidas son cero. Ya vimos que en un modelo restringido las varianzas de los estimadores eran menores, o como máximo iguales, que las asociadas a los estimadores del modelo no restringido. De acuerdo con esto, pues, es evidente que las varianzas de los estimadores en el modelo con omisión de variables relevantes serán menores, o iguales, que las de los mismos estimadores en el modelo correctamente especificado.

Consultad las varianzas de un modelo restringido en el subapartado 3.2.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Consistencia

Para acabar, sólo queda por analizar la **consistencia de los estimadores $\hat{\mathbf{B}}^*$** en un modelo con omisión de variables relevantes.

Recordad la expresión 1.3 de los estimadores $\hat{\mathbf{B}}^*$. Recordad también la consistencia de los estimadores en ECM en el subapartado 2.3.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Dado que los estimadores $\hat{\mathbf{B}}_1$ son estimadores sesgados y que el sesgo no tiende a cero cuando aumenta el tamaño muestral, son estimadores inconsistentes en términos de error cuadrático medio (ECM).

Excepción única a la inconsistencia en ECM de los estimadores $\hat{\mathbf{B}}_1$

Solamente hay una excepción: si las variables consideradas en el modelo erróneamente especificado (variables de la matriz $\mathbf{X}_1$) son ortogonales a las variables omitidas (variables de la matriz $\mathbf{X}_2$), entonces los estimadores $\hat{\mathbf{B}}_1$ son consistentes en ECM, dado que en este caso, al ser el sesgo cero, se tiene lo siguiente:

$$\text{ECM}[\hat{\mathbf{B}}_1] = (\text{sesgo}[\hat{\mathbf{B}}_1])^2 + \text{VAR}[\hat{\mathbf{B}}_1] = \text{VAR}[\hat{\mathbf{B}}_1], \tag{1.8}$$

y tomando límites a 1.8 se tiene:

$$\lim_{N\to\infty} \text{ECM}[\hat{\mathbf{B}}_1] = \lim_{N\to\infty} \text{VAR}[\hat{\mathbf{B}}_1] = \lim_{N\to\infty} \left[\frac{\sigma_u^2}{N}\left(\frac{\mathbf{X}'_1\mathbf{X}_1}{N}\right)^{-1}\right] = 0. \tag{1.9}$$

El límite es cero porque en la primera parte , $\frac{\sigma_u^2}{N}$, el numerador es un escalar y el denominador es un valor que tiende a infinito. La segunda parte del límite es finita. Así, el producto entre un límite que tiende a cero y otro que es finito es igual a cero.

1.2.3. Propiedades del estimador de mínimos cuadrados ordinarios de la varianza del término de perturbación

El estimador MCO de la varianza del término de perturbación en el modelo erróneamente especificado se obtiene realizando el cociente entre la suma de cuadrados de los errores (*SCE*) asociada al modelo con omisión de variables y los grados de libertad del modelo mencionado, es decir:

Consultad la *SCE* y el estimador MCO de la varianza del término de perturbación en el modelo correctamente especificado, respectivamente, en los subapartados 2.3.1 y 2.4.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

$$\hat{\sigma}_{u1}^2 = \frac{SCE_1}{N-r} = \frac{\mathbf{e}_1'\mathbf{e}_1}{N-r}. \tag{1.10}$$

El subíndice 1 indica que trabajamos con el modelo con omisión de variables relevantes, donde las variables explicativas consideradas son únicamente las variables de la submatriz $\mathbf{X}_1$.

Se puede demostrar que el estimador visto en la expresión 1.10 es un estimador sesgado de σ_u^2 del modelo con todas las variables y que el sesgo es positivo e igual a $\mathbf{B}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{B}_2$, es decir:

$$\text{E}[\hat{\sigma}_{u1}^2] > \sigma_u^2. \tag{1.11}$$

La conclusión anterior también es cierta cuando las variables de las matrices $\mathbf{X}_1$ y $\mathbf{X}_2$ son ortogonales, y en este caso particular el sesgo es el siguiente:

$$\text{sesgo}[\hat{\sigma}_{u1}^2] = \frac{1}{N-r}\mathbf{B}_2'\mathbf{X}_2'\mathbf{X}_2\mathbf{B}_2 > 0. \tag{1.12}$$

Además de las consecuencias que acabamos de estudiar, la omisión de variables relevantes puede conducir a incumplir algunas hipótesis básicas del modelo de regresión. Así, esta omisión puede aparentar un cambio estructural, o bien su efecto puede trasladarse de manera sistemática a los residuos del modelo, lo cual provoca que éstos muestren un cierto comportamiento determinista (no aleatorio) y, por tanto, parece que haya autocorrelación en el término de perturbación, o bien hace que el valor esperado del término de perturbación sea distinto de cero.

Consultad las hipótesis básicas del MRLM estándar en el subapartado 2.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

1.2.4. Ejemplo para el caso de dos variables explicativas

Supongamos que el modelo verdadero, donde las variables están expresadas en desviaciones respecto a sus medias muestrales, es el siguiente:

$$\tilde{Y}_i = \beta_2\tilde{X}_{2i} + \beta_3\tilde{X}_{3i} + u_i \quad \forall i = 1, ..., N. \tag{1.13}$$

No obstante, por error, se especifica:

$$\tilde{Y}_i = \beta_2\tilde{X}_{2i} + v_i \quad \forall i = 1, ..., N. \tag{1.14}$$

Para comprobar las propiedades de los estimadores MCO en el modelo erróneamente especificado mencionado, calculamos las expresiones siguientes:

a) En primer lugar, el estimador $\hat{\beta}_2$ de 1.14:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}\tilde{X}_{2i}\tilde{Y}_i}{\sum_{i=1}^{N}\tilde{X}_{2i}^2}. \tag{1.15}$$

Tomando las esperanzas matemáticas, vemos que el estimador es sesgado, ya que su esperanza matemática no es igual, en general, al valor de la población:

$$E[\hat{\beta}_2] = \beta_2 + \beta_3\frac{\sum_{i=1}^{N}\tilde{X}_{2i}\tilde{X}_{3i}}{\sum_{i=1}^{N}\tilde{X}_{2i}^2}. \tag{1.16}$$

Consultad el MRLM erróneamente especificado por omisión de variables y las propiedades de los estimadores de este modelo en los subapartados 1.2.1 y 1.2.2 de este módulo, respectivamente.

b) La varianza de la población del estimador $\hat{\beta}_2$ que se obtendría a partir del modelo erróneamente especificado del ejemplo sería:

$$\frac{\sigma_u^2}{\sum_{i=1}^{N}\tilde{X}_{2i}^2}, \tag{1.17}$$

mientras que la que se obtendría si se especificase correctamente el modelo sería:

$$\frac{\sigma_u^2\sum_{i=1}^{N}\tilde{X}_{3i}^2}{\sum_{i=1}^{N}\tilde{X}_{2i}^2\sum_{i=1}^{N}\tilde{X}_{3i}^2 - \left(\sum_{i=1}^{N}\tilde{X}_{2i}\tilde{X}_{3i}\right)^2}. \tag{1.18}$$

Podemos comprobar que 1.18...

... equivale a la expresión:

$$\frac{\sigma_u^2}{\sum_{i=1}^{N}\tilde{X}_{2i}^2(1 - r_{23}^2)},$$

donde r_{23} indica el coeficiente de correlación simple entre $\tilde{X}_2$ y $\tilde{X}_3$ Dado que $0 < r_{23}^2 < 1$, se comprueba que 1.17 es menor, o como máximo igual, que 1.18.

c) Para acabar, analizamos la consistencia en ECM del estimador $\hat{\beta}_2$ en el modelo erróneamente especificado. Cuando las variables $\tilde{X}_2$ y $\tilde{X}_3$ están correlacionadas, en el modelo especificado el estimador $\hat{\beta}_2$ es sesgado; por lo tanto:

$$\lim_{N\to\infty} \text{ECM}[\hat{\beta}_2] = \lim_{N\to\infty} \text{VAR}[\hat{\beta}_2] + \lim_{N\to\infty} \left(\text{sesgo}[\hat{\beta}_2]\right)^2 =$$
$$= 0 + \lim_{N\to\infty} \left(\beta_3 \frac{\sum \tilde{X}_{2i}\tilde{X}_{3i}}{\sum \tilde{X}_{2i}^2}\right)^2 \neq 0.$$

Entonces, si $r_{23} \neq 0$, $\hat{\beta}_2$ en el modelo erróneamente especificado no es consistente.

1.3. Efectos por la inclusión de variables superfluas

De manera análoga al caso de omisión de variables relevantes del modelo, ahora estudiaremos los efectos sobre el MRLM de estimar por MCO un modelo en el que se incluyen más variables explicativas de las necesarias. Es interesante ir comparando la contraposición de resultados obtenidos respecto al caso anterior.

Consultad los problemas asociados a la omisión de variables relevantes en el modelo en el subapartado 1.2 de este módulo didáctico.

1.3.1. Especificación del modelo

Supongamos ahora que el modelo verdadero es el siguiente:

Nota: no debéis confundir la matriz X_1 de este subapartado con la definida en el subapartado 1.2.1.

$$\mathbf{Y} = \mathbf{X_1}\mathbf{B_1} + \mathbf{U_1}, \tag{1.19}$$

donde la matriz de las variables explicativas, $\mathbf{X_1}$, agrupa únicamente a las variables que realmente determinan el comportamiento de la variable endógena; es decir, en la matriz $\mathbf{X_1}$ no hay ninguna variable relevante. Además, suponemos que el número de estas variables es k; por lo tanto, la matriz $\mathbf{X_1}$ es de dimensión $N \times k$:

$$\mathbf{X_1} = \begin{bmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2N} & \dots & X_{kN} \end{bmatrix}. \tag{1.20}$$

Por su parte, el vector $\mathbf{B_1}$ (de dimensión k) agrupa a los parámetros asociados a las variables de la matriz $\mathbf{X_1}$. Asimismo, suponemos que el vector de términos de perturbación del modelo correctamente especificado 1.19, $\mathbf{U_1}$, se distribuye según una normal multivariante de orden N con valor esperado cero y matriz de varianzas y covarianzas diagonal con un valor constante σ_u^2 en la diagonal:

$$\mathbf{U_1} \sim \text{N}(\mathbf{0}_{N\times 1}, \sigma_u^2 \mathbf{I}_N).$$

Para analizar los efectos de la inclusión de variables irrelevantes en la especificación del modelo, proponemos el modelo erróneo siguiente:

$$\mathbf{Y} = \mathbf{X}_0\mathbf{B}_0 + \mathbf{U}_0, \tag{1.21}$$

donde en la matriz $\mathbf{X}_0$, además de las k variables que realmente explican el comportamiento de la variable endógena, se añaden erróneamente s variables que son irrelevantes a la hora de explicar el comportamiento de la variable mencionada. Por lo tanto, la matriz $\mathbf{X}_0$ es de dimensión $N \times (k + s)$:

$$\mathbf{X}_0 = \left[\begin{array}{cccc|ccc} 1 & X_{21} & \dots & X_{k1} & X_{(k+1)1} & \dots & X_{(k+s)1} \\ 1 & X_{22} & \dots & X_{k2} & X_{(k+1)2} & \dots & X_{(k+s)2} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & X_{2N} & \dots & X_{kN} & X_{(k+1)N} & \dots & X_{(k+s)N} \end{array}\right]. \tag{1.22}$$

Observad que...

... bajo la hipótesis de la inclusión de variables superfluas, el rango de la matriz $\mathbf{X}_0$ siempre es mayor que el de $\mathbf{X}_1$:

$\rho(\mathbf{X}_0) = k + s > \rho(\mathbf{X}_1) = k.$

Así, también podemos escribir el modelo incorrectamente especificado por inclusión de variables irrelevantes del modo siguiente:

$$\mathbf{Y} = \mathbf{X}_0\mathbf{B}_0 + \mathbf{U}_0 = [\mathbf{X}_1 \mid \mathbf{X}_2]\begin{bmatrix}\mathbf{B}_1 \\ - \\ \mathbf{B}_2\end{bmatrix} + \mathbf{U}_0 = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \mathbf{U}_0, \tag{1.23}$$

donde $\mathbf{X}_1$ es la submatriz de las observaciones correspondientes a las k variables explicativas relevantes; $\mathbf{X}_2$, la correspondiente a las s variables explicativas irrelevantes incluidas en el modelo 1.21, y $\mathbf{B}_1$ y $\mathbf{B}_2$ son dos vectores de dimensiones k y s que reúnen, respectivamente, los parámetros asociados a las k variables relevantes y los asociados a las s variables irrelevantes consideradas en el modelo.

1.3.2. Propiedades de los estimadores $\hat{\mathbf{B}}_0$

En primer lugar, presentamos la expresión del estimador de $\mathbf{B}_0$. Si le aplicamos la fórmula de la estimación por mínimos cuadrados ordinarios en el modelo 1.21, obtenemos lo siguiente:

Consultad la fórmula de la estimación MCO en un modelo en el subapartado 2.3.1 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

$$\hat{\mathbf{B}}_0 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y} = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'(\mathbf{X}_1\mathbf{B}_1 + \mathbf{U}_1) =$$

$$= \left(\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}[\mathbf{X}_1 \mid \mathbf{X}_2]\right)^{-1}\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}(\mathbf{X}_1\mathbf{B}_1 + \mathbf{U}_1) =$$

Nota: no confundáis la matriz $\mathbf{X}_1$ de este subapartado con la definida en el subapartado 1.2.1.

$$= \left(\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}[\mathbf{X}_1 \mid \mathbf{X}_2]\right)^{-1}\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}\mathbf{X}_1\mathbf{B}_1 + \left(\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}[\mathbf{X}_1 \mid \mathbf{X}_2]\right)^{-1}\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}\mathbf{U}_1, \tag{1.24}$$

que es igual a:

$$\hat{\mathbf{B}}_0 = \begin{bmatrix}\mathbf{I}_k \\ - \\ \mathbf{0}_{s\times k}\end{bmatrix}\mathbf{B}_1 + \left(\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}[\mathbf{X}_1 \mid \mathbf{X}_2]\right)^{-1}\begin{bmatrix}\mathbf{X}_1' \\ - \\ \mathbf{X}_2'\end{bmatrix}\mathbf{U}_1. \tag{1.25}$$

El resultado 1.25 es lógico...

... porque, si se hace una regresión de una variable contra sí misma, independientemente de si se consideran o no otras variables explicativas en la regresión, el estimador MCO del parámetro en cuestión es 1, y el resto de los estimadores (si los hay) son cero.

A continuación analizamos las propiedades del estimador $\hat{\mathbf{B}}_0$.

Sesgo

Con el fin de averiguar si se trata de un **estimador no sesgado**, tomando esperanzas matemáticas en la expresión 1.25 y desarrollándolo, llegamos a:

Recordad que $E(U_1) = 0$.

$$E[\hat{\mathbf{B}}_0] = E\begin{bmatrix}\hat{\mathbf{B}}_1\\ -\\ \hat{\mathbf{B}}_2\end{bmatrix} = \begin{bmatrix}E[\hat{\mathbf{B}}_1]\\ -\\ E[\hat{\mathbf{B}}_2]\end{bmatrix} = \begin{bmatrix}\mathbf{B}_1\\ -\\ \mathbf{0}_{k\times s}\end{bmatrix}. \qquad (1.26)$$

El resultado 1.26 permite concluir que los estimadores MCO de los k coeficientes asociados a las variables explicativas relevantes ($\mathbf{X}_1$) no tienen sesgo, $E[\hat{\mathbf{B}}_1] = \mathbf{B}_1$, mientras que los de los s parámetros asociados a las variables explicativas irrelevantes incluidas en el modelo ($\mathbf{X}_2$) tienen una esperanza matemática igual a su valor de la población, que es cero ($E[\hat{\mathbf{B}}_2] = 0$); por lo tanto, sus estimadores tampoco tienen sesgo.

Matriz de varianzas y covarianzas

La **matriz de varianzas y covarianzas de los estimadores** en el modelo erróneamente especificado es la siguiente:

$$VAR[\hat{\mathbf{B}}_0] = \sigma_u^2(\mathbf{X}_0'\mathbf{X}_0)^{-1}, \qquad (1.27)$$

y concretamente, la submatriz para los estimadores $\hat{\mathbf{B}}_1$ en el modelo mal especificado:

$$VAR[\hat{\mathbf{B}}_1] = \begin{bmatrix} VAR[\hat{\beta}_1] & \dots & COV[\hat{\beta}_1,\hat{\beta}_k]\\ \vdots & \ddots & \vdots\\ COV[\hat{\beta}_k,\hat{\beta}_1] & \dots & VAR[\hat{\beta}_k]\end{bmatrix} =$$

$$= \sigma_u^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}, \qquad (1.28)$$

toma la matriz de varianzas y covarianzas de los estimadores del vector $\hat{\mathbf{B}}_1$ asociados a las **variables explicativas relevantes** (de la matriz $\mathbf{X}_1$) que se obtienen como consecuencia de especificar erróneamente el modelo de regresión por la inclusión de **variables irrelevantes** (las variables de la matriz $\mathbf{X}_2$). Por otra parte, como sabemos, si se especificase correctamente el modelo, la matriz de varianzas y covarianzas del estimador $\hat{\mathbf{B}}_1$ sería $\sigma_u^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$.

Consultad la matriz de varianzas y covarianzas del estimador en el modelo correctamente especificado en el subapartado 2.3.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

En consecuencia, la inclusión de variables irrelevantes en el modelo de regresión provoca un incremento en las varianzas de los estimadores asociadas a las variables explicativas relevantes respecto a las que se obtendrían por los mismos estimadores en el modelo correctamente especificado.

La única excepción a la conclusión anterior tiene lugar cuando las variables irrelevantes incluidas en el modelo ($\mathbf{X_2}$) y las variables relevantes ($\mathbf{X_1}$) están incorrelacionadas: en este caso, la inclusión de variables irrelevantes no provoca ningún incremento en las varianzas de los estimadores de las variables relevantes.

Consistencia

La tercera propiedad que hay que estudiar es la **consistencia de los estimadores** en el modelo erróneamente especificado. Puesto que los estimadores en un modelo con inclusión de variables irrelevantes no tienen sesgo, en este caso el ECM coincide con la varianza:

$$\mathrm{ECM}[\hat{\mathbf{B}}_1] = \mathrm{VAR}[\hat{\mathbf{B}}_1] = \sigma_u^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}. \qquad (1.29)$$

$\hat{\mathbf{B}}_1$ son las estimaciones MCO que se obtendrían a partir del modelo mal especificado.

Tomando límites en la expresión anterior, se tiene el resultado siguiente:

$$\lim_{N\to\infty} \mathrm{ECM}[\hat{\mathbf{B}}_1] = \lim_{N\to\infty} \mathrm{VAR}[\hat{\mathbf{B}}_1] = \qquad (1.30)$$

$$= \lim_{N\to\infty} \left[\frac{\sigma_u^2}{N}\left(\frac{\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1}{N}\right)^{-1}\right] = 0.$$

El resultado obtenido permite concluir que, ante un error por inclusión de variables irrelevantes, los estimadores MCO mantienen la propiedad de consistencia en ECM.

1.3.3. Propiedades del estimador de mínimos cuadrados ordinarios de la varianza del término de perturbación

A continuación estudiamos el comportamiento del estimador MCO de la varianza del término de perturbación en el caso de inclusión de variables irrelevantes. Partiendo de la expresión conocida del estimador mencionado:

$$\hat{\sigma}_{u0}^2 = \frac{\mathbf{e}_0'\mathbf{e}_0}{N - (k + s)}, \qquad (1.31)$$

Consultad la expresión del estimador MCO de la varianza del término de perturbación en el subapartado 2.4.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

se puede demostrar que el estimador MCO de la varianza del término de perturbación que se obtendrá no tiene sesgo.

1.3.4. Ejemplo para el caso de dos variables explicativas

Del mismo modo que hemos hecho con el caso de omisión de variables explicativas relevantes, planteamos un ejemplo teórico. Supongamos que el modelo correcto, donde las variables están expresadas en desviaciones respecto a sus medias muestrales, es el siguiente:

Consultad el ejemplo del caso de omisión de variables explicativas relevantes del modelo en el subapartado 1.2.4 de este módulo didáctico.

$$\tilde{Y}_i = \beta_2 \tilde{X}_{2i} + u_i \quad \forall i = 1, ..., N, \tag{1.32}$$

mientras que el especificado de forma errónea, donde se ha añadido la variable irrelevante $\mathbf{X}_3$ (también en desviaciones), es el siguiente:

$$\tilde{Y}_i = \beta_2 \tilde{X}_{2i} + \beta_3 \tilde{X}_{3i} + v_i \quad \forall i = 1, ..., N. \tag{1.33}$$

Los estimadores MCO de los parámetros del modelo 1.33 son los siguientes:

$$\hat{\mathbf{B}}_0 = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \tilde{X}_{2i}^2 & \sum_{i=1}^{N} \tilde{X}_{2i}\tilde{X}_{3i} \\ \sum_{i=1}^{N} \tilde{X}_{3i}\tilde{X}_{2i} & \sum_{i=1}^{N} \tilde{X}_{3i}^2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{N} \tilde{X}_{2i}\tilde{Y}_i \\ \sum_{i=1}^{N} \tilde{X}_{3i}\tilde{Y}_i \end{bmatrix} =$$

$$= \frac{1}{\sum_{i=1}^{N} \tilde{X}_{2i}^2 \sum_{i=1}^{N} \tilde{X}_{3i}^2 - \left(\sum_{i=1}^{N} \tilde{X}_{2i}\tilde{X}_{3i}\right)^2} \begin{bmatrix} \sum_{i=1}^{N} \tilde{X}_{3i}^2 & -\sum_{i=1}^{N} \tilde{X}_{2i}\tilde{X}_{3i} \\ -\sum_{i=1}^{N} \tilde{X}_{3i}\tilde{X}_{2i} & \sum_{i=1}^{N} \tilde{X}_{2i}^2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{N} \tilde{X}_{2i}\tilde{Y}_i \\ \sum_{i=1}^{N} \tilde{X}_{3i}\tilde{Y}_i \end{bmatrix}.$$

Para comprobar las propiedades de los estimadores MCO del modelo erróneamente especificado hacemos los cálculos siguientes:

a) Centrándonos en el estimador $\hat{\beta}_2$ asociado a la variable $\tilde{X}_2$, si sustituimos $\tilde{Y}_i$ por su verdadero valor tomado en 1.32, tenemos:

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{N} \tilde{X}_{3i}^2\right)\left[\sum_{i=1}^{N} \tilde{X}_{2i}(\beta_2 \tilde{X}_{2i} + u_i)\right] - \left(\sum_{i=1}^{N} \tilde{X}_{3i}\tilde{X}_{2i}\right)\left[\sum_{i=1}^{N} \tilde{X}_{3i}(\beta_2 \tilde{X}_{2i} + u_i)\right]}{\left(\sum_{i=1}^{N} \tilde{X}_{2i}^2\right)\left(\sum_{i=1}^{N} \tilde{X}_{3i}^2\right) - \left(\sum_{i=1}^{N} \tilde{X}_{2i}\tilde{X}_{3i}\right)^2}. \tag{1.34}$$

Se comprueba que la expresión de $\mathrm{VAR}[\hat{\beta}_2]$**...**

... equivale a la expresión

$$\frac{\sigma_u^2}{\sum_{i=1}^{N} \tilde{X}_{2i}^2 (1 - r_{23}^2)},$$

donde r_{23} es el coeficiente de correlación entre $\tilde{X}_2$ y $\tilde{X}_3$. Dado que $0 < r_{23}^2 < 1$, esta expresión es mayor, o como máximo igual, que la asociada al modelo correctamente especificado (comparad con 1.18)

- Tomando esperanzas matemáticas en la expresión 1.34, se llega a lo siguiente:

$$\mathrm{E}[\hat{\beta}_2] = \beta_2. \tag{1.35}$$

- Operando de la misma manera para el caso del estimador $\hat{\beta}_3$, se obtiene que su valor esperado es cero: $\mathrm{E}[\hat{\beta}_3] = 0$.

b) Por otra parte, la varianza del estimador $\hat{\beta}_2$ obtenida a partir del modelo erróneamente especificado es la siguiente:

$$\mathrm{VAR}[\hat{\beta}_2] = \sigma_u^2 \frac{\sum_{i=1}^{N} \tilde{X}_{3i}^2}{\left(\sum_{i=1}^{N} \tilde{X}_{2i}^2\right)\left(\sum_{i=1}^{N} \tilde{X}_{3i}^2\right) - \left(\sum_{i=1}^{N} \tilde{X}_{2i}\tilde{X}_{3i}\right)^2}.$$

Sin embargo, la varianza del estimador $\hat{\beta}_2$ que se obtendría si se trabajase con el modelo correctamente especificado sería $\dfrac{\sigma_u^2}{\sum_{i=1}^{N} \tilde{X}_{2i}^2}$.

1.3.5. Síntesis

Para acabar, a continuación presentamos un cuadro donde resumimos las principales conclusiones obtenidas en el análisis de los efectos de la omisión de variables explicativas relevantes y de la inclusión de variables explicativas irrelevantes realizada a lo largo de las páginas anteriores:

Propiedades de los estimadores MCO ante errores en la especificación de la parte determinista del modelo		
	Omisión de variables relevantes	**Inclusión de variables irrelevantes**
$\hat{\beta}_j$	Sesgados, excepto si las variables de X_1 y de X_2 son ortogonales.	No sesgados.
	Varianzas menores que en el modelo correctamente especificado, excepto si las variables de X_1 y de X_2 son ortogonales, porque entonces las varianzas son iguales.	Varianzas mayores que en el modelo correctamente especificado, excepto si las variables de X_1 y de X_2 son ortogonales, porque entonces las varianzas son iguales.
	Inconsistentes, excepto si las variables de X_1 y de X_2 son ortogonales.	Consistentes.
$\hat{\sigma}_u^2$	Sesgado.	No sesgado.

Se puede concluir que, en principio, viendo los resultados anteriores, en el modelo de regresión es más problemática la omisión de variables relevantes que la inclusión de variables irrelevantes.

1.4. Errores en la forma funcional

Se dice que se comete un error en la forma funcional cuando se especifica una relación (que puede ser lineal cuadrática, cúbica, exponencial, logarítmica, etc.) y la verdadera relación es diferente de la especificada.

Una especificación incorrecta en la forma funcional del modelo puede considerarse, en algunos casos, como un error de especificación asimilable a la omisión de variables relevantes. En estos casos, las consecuencias serán, por tanto, las mismas que las que provoca la omisión de variables relevantes, es decir, los estimadores serán sesgados e inconsistentes.

En general, especificar una relación equivocada puede conducir a obtener un término de perturbación no esférico (es decir, con hetero*sce*dasticidad y/o autocorrelación), así como al hecho de que la ley de probabilidad se aleje de la distribución del término de perturbación del modelo correctamente especificado (habitualmente, la distribución normal). En consecuencia, es importante disponer de algún método para detectar un posible error en la especificación de la forma funcional.

Consultad los conceptos de heteroscedasticidad, autocorrelación y término de perturbación no esférico en el subapartado 2.2.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

1.4.1. Contraste *reset*

Uno de los contrastes más utilizados para detectar si la forma funcional de un modelo es o no correcta es el **contraste *reset***, propuesto por Anscombe y Ramsey en la década de los sesenta. La hipótesis nula H_0 es que la forma funcional es la correcta. Para hacer este contraste se parte de un modelo de regresión especificado en forma lineal, como, por ejemplo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i \quad \forall i = 1, \ldots, N. \qquad (1.36)$$

Para comprobar si la forma funcional es la apropiada, seguiremos estas etapas:

1) Se estima por MCO el modelo 1.36, se obtiene el vector de valores ajustados de la variable endógena y se elevan al cuadrado.

2) A continuación se especifica la regresión auxiliar siguiente:

$$Y_i = \delta_1 + \delta_2 X_{2i} + \ldots + \delta_k X_{ki} + \gamma \hat{Y}_i^2 + \nu_i \quad \forall i = 1, \ldots, N, \qquad (1.37)$$

donde se ha añadido un regresor adicional al modelo original, la variable endógena ajustada elevada al cuadrado obtenida en la etapa anterior.

3) Acto seguido, se estima por MCO esta regresión auxiliar 1.37 y se contrasta de la manera habitual (mediante el test de la t de Student o el de la F de Snedecor) si el coeficiente asociado a la variable adicional, $\hat{Y}_i^2$, es significativamente distinto de cero, en cuyo caso se rechazará la linealidad de la relación.

Consultad el test de la t de Student o el de la F de Snedecor en el subapartado 2.7.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Consecuencias de una especificación incorrecta en la forma funcional de un modelo

Supongamos que el modelo correctamente especificado (en desviaciones) es el siguiente:

$$\tilde{Y}_i = \beta_2 \tilde{X}_{2i} + \beta_3 \tilde{X}_{2i}^2 + u_i \quad \forall i = 1, \ldots, N, \qquad (1.38)$$

donde $\tilde{X}_{2i} = (X_{2i} - \bar{X}_2)$. Pero se especifica lo siguiente:

$$\tilde{Y}_i = \beta_2 \tilde{X}_{2i} + v_i \quad \forall i = 1, ..., N. \tag{1.39}$$

Para obtener otras formas funcionales...

... distintas de la cuadrática, a veces en el modelo auxiliar pueden incluirse otros exponentes en el regresor Y ajustado. Entonces, el contraste de la F de Snedecor de significación del subconjunto de parámetros añadidos es el adecuado.

Para detectar el error cometido por el investigador a la hora de especificar la forma funcional del modelo, podemos usar el contraste reset. Para ello, seguimos los pasos siguientes:

1) En primer lugar, se estima por MCO el modelo 1.39. Los resultados obtenidos se presentan en el cuadro de resultados de la estimación MCO del modelo 1.39.

Resultados de la estimación MCO del modelo 1.39

Variable dependiente: Y

Número de observaciones: 50

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
X_2	2.4273341	0.5776526	4.2020654	0.0001

R-squared	0.150528	Mean of dependent var	0.493011
Adjusted R-squared	0.150528	S.D. of dependent var	1.262594
S.E. of regression	1.163692	Sum of squared resid	66.35478
Log likelihood	-78.02175	Durbin-Watson stat	1.855630

2) A continuación, se calcula el vector de valores ajustados de la variable endógena elevados al cuadrado (que en este ejemplo llamamos F_2).

3) Acto seguido, se estima la regresión auxiliar del test por MCO, que en este ejemplo es la siguiente:

$$\tilde{Y}_i = \delta \tilde{X}_{2i} + \gamma F_{2i} + w_i \quad \forall i = 1, ..., N. \tag{1.40}$$

Los resultados de la estimación del modelo 1.40 se reproducen en el cuadro siguiente:

Resultados de la estimación MCO del modelo 1.40

Variable dependiente: Y

Número de observaciones: 50

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
X_2	2.6815927	0.5065498	5.2938378	0.0000
F_2	**0.8796791**	**0.2152769**	**4.0862688**	**0.0002**

R-squared	0.369766	Mean of dependent var	0.493011
Adjusted R-squared	0.356636	S.D. of dependent var	1.262594
S.E. of regression	1.012726	Sum of squared resid	49.22949
Log likelihood	-70.55867	F-statistic	28.16217
Durbin-Watson stat	2.309458	Prob(F-statistic)	0.000003

El contraste consiste en estudiar si en el modelo 1.40 el parámetro γ es estadísticamente distinto de cero. Viendo los resultados obtenidos (el estadístico t y su significación), se rechaza la linealidad del modelo y, por tanto, se habría cometido un error si se hubiese especificado una forma lineal cuando en realidad la relación es cuadrática. Para hacer este contraste se puede emplear el estadístico de la t de Student, contrastando la hipótesis nula $H_0 : \gamma = 0$.

2. Errores en la muestra: multicolinealidad y observaciones atípicas

En este apartado explicamos dos situaciones que pueden plantear el conjunto de observaciones de las variables del modelo de regresión y que afectan a sus resultados. Nos referimos a la multicolinealidad y a las observaciones atípicas. Agrupamos a estos dos aspectos en un único apartado debido a que ambos problemas se plantean en la matriz **X** de datos.

A lo largo de los subapartados 2.1, 2.2 y 2.3 de este módulo didáctico se desarrolla la multicolinealidad, y en el subapartado 2.4 se tratan las observaciones atípicas.

2.1 Multicolinealidad: definición y consecuencias

Ya hemos visto que un MRLM estándar tenía que cumplir, entre otras, la hipótesis de ausencia de multicolinealidad perfecta.

Consultad la hipótesis de ausencia de multicolinealidad perfecta en el MRLM estándar en el subapartado 2.2.3 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Existe **ausencia de multicolinealidad perfecta** cuando no hay ninguna variable explicativa que se pueda obtener a partir de la combinación lineal de otras. Dicho de otro modo, no puede existir ninguna variable explicativa que presente una correlación perfecta respecto a una o diversas variables del resto de las variables explicativas.

Otras etapas del trabajo econométrico...

... que pueden resultar afectadas por la presencia de multicolinealidad son, por ejemplo, la contrastación y el análisis estructural.

La presencia de un grado elevado de multicolinealidad en un modelo tiene consecuencias negativas en la estimación del modelo y, por extensión, en las otras etapas del trabajo econométrico.

Así, en un modelo con k variables explicativas nos podemos encontrar ante tres tipos de situaciones, aunque las dos primeras no se dan de una manera habitual en la práctica. Veamos estas situaciones a continuación:

En un MRLS...

... no pueden existir problemas de multicolinealidad, por la misma definición de multicolinealidad, dado que sólo hay un regresor (aparte de que pueda haber un término independiente).

1) Presencia de multicolinealidad perfecta. En este caso, el rango de la matriz **X** será de orden menor que k ($\rho(\mathbf{X}) < k$), lo cual indica que existe alguna variable explicativa que se puede obtener a partir de la combinación lineal de otras variables explicativas. Es decir, hay multicolinealidad perfecta cuando las variables explicativas son linealmente dependientes entre sí.

2) Ausencia total de multicolinealidad. El rango de **X** será igual a k ($\rho(\mathbf{X}) = k$) y, además, no debe existir correlación entre las variables explicativas del modelo.

3) Presencia de un cierto nivel de multicolinealidad. Aunque $\rho(\mathbf{X}) = k$, existe una correlación distinta de cero entre algunas o todas las variables explicativas del modelo. Éste es el caso más habitual en la práctica, y, como

veremos más adelante, los inconvenientes del modelo derivados de la multicolinealidad van creciendo a medida que esta correlación es mayor.

En el caso de presencia...

... de un cierto nivel de multicolinealidad, el rango de la matriz **X** es el número de columnas linealmente independientes, que en este caso es *k*.

Para acabar, hay que remarcar que el concepto de multicolinealidad está directamente vinculado al de correlación entre los regresores del modelo, pero en ningún caso a la correlación entre éstos y la variable endógena.

2.1.1. Consecuencias de la multicolinealidad perfecta en la estimación

La existencia de multicolinealidad perfecta, como ya hemos comentado, conduce al incumplimiento de la hipótesis de rango máximo ($\rho(\mathbf{X}) < k$). Entonces, la matriz **X'X** no se puede invertir y las estimaciones MCO no se pueden obtener.

Ejemplo de existencia de multicolinealidad perfecta

Dado el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad i = 1, ..., N$, se conoce que existe la relación $X_{4i} = 3X_{2i} + 2X_{3i}$. Entonces, no será posible obtener una estimación MCO de cada uno de los parámetros*. Si introducimos la expresión de X_{4i} en el modelo, obtenemos lo siguiente:

*** La matriz X'X no se podría invertir en este caso.**

$$Y_i = \beta_1 + (\beta_2 + 3\beta_4)X_{2i} + (\beta_3 + 2\beta_4)X_{3i} + u_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i \quad i = 1, ..., N,$$

que permite obtener estimaciones MCO de los parámetros a (que son una combinación lineal de los parámetros β), pero no de los β.

2.1.2. Consecuencias de la ausencia de multicolinealidad en la estimación

La ausencia total de multicolinealidad es la situación deseable, aunque en la práctica no se presenta casi nunca. Supone que el coeficiente de correlación entre cada par de variables explicativas es cero.

La consecuencia más relevante es que la estimación de los parámetros del modelo, $\hat{\mathbf{B}}$, será la misma tanto si se obtiene a partir de un MRLM como si se obtiene a partir de tantos MRLS como variables explicativas tenga el modelo. Es decir, si la correlación entre las variables explicativas es nula, cada una de éstas explica una parte distinta de la variable endógena.

2.1.3. Consecuencias sobre la estimación en presencia de un cierto nivel de multicolinealidad

El último caso es el intermedio respecto a los dos casos anteriores. Veremos que los resultados que mostraremos aquí dependen del nivel de multicolinealidad* presente en el modelo. Por lo tanto, las consecuencias negativas serán mayores cuanto mayor sea la correlación entre las variables explicativas.

*** Nivel de correlación entre las variables explicativas.**

De todos modos, en presencia de un cierto nivel de multicolinealidad, independientemente del nivel, los estimadores MCO continúan conservando las propiedades óptimas deseables. Así, son estimadores lineales, no sesgados, consistentes y eficientes. Asimismo, la multicolinealidad no afecta a la estimación de la varianza del término de perturbación.

Las consecuencias principales de la presencia de un cierto nivel de multicolinealidad son las siguientes:

a) Cuanto mayor sea la correlación, el determinante de la matriz **X'X** será más próximo a cero y, en consecuencia, las varianzas y las covarianzas de los estimadores serán más elevadas*. Por lo tanto, la presencia de un grado alto de multicolinealidad llevará a una reducción de la precisión y de la estabilidad de la estimación, dado que las varianzas y las covarianzas –aunque sean mínimas– serán mayores que las que se tendrían en presencia de un nivel de multicolinealidad inferior. Esto es así porque la varianza de los estimadores continúa siendo la mínima dentro de la familia de los estimadores lineales y no sesgados.

*** Recordad que las varianzas y las covarianzas de los estimadores se calculan a partir de la expreión $\sigma_u^2(X'X)^{-1}$.**

Aumento de las varianzas y las covarianzas de los estimadores con la correlación

Se puede demostrar que en un MRLM con dos variables explicativas en desviaciones como el siguiente:

$$\tilde{Y}_i = \beta_2\tilde{X}_{2i} + \beta_3\tilde{X}_{3i} + u_i \quad i = 1, ..., N,$$

las varianzas y las covarianzas entre $\hat{\beta}_2$ y $\hat{\beta}_3$ son éstas:

$$\mathrm{VAR}[\hat{\beta}_2] = \frac{\sigma_u^2}{\sum_{i=1}^{N}\tilde{X}_{2i}^2(1 - r_{23}^2)}\ ;\ \mathrm{VAR}[\hat{\beta}_3] = \frac{\sigma_u^2}{\sum_{i=1}^{N}\tilde{X}_{3i}^2(1 - r_{23}^2)}\ ;$$

$$\mathrm{COV}[\hat{\beta}_2,\hat{\beta}_3] = \frac{-\sigma_u^2\sum_{i=1}^{N}\tilde{X}_{2i}\tilde{X}_{3i}}{\sum_{i=1}^{N}\tilde{X}_{2i}^2\tilde{X}_{3i}^2(1 - r_{23}^2)}\ .$$

En las expresiones anteriores queda explicitado con claridad que, cuando hay una correlación mayor entre los regresores (es decir, más cerca de 1, en valor absoluto), indicada por r_{23}, el valor de la varianza es mayor. Así, por ejemplo, si $(r_{23})^2 = 0{,}5$, la varianza es el doble con respecto al caso de ortogonalidad ($r_{23} = 0$), y, si $(r_{23})^2 = 0{,}95$, la varianza se multiplica por 20.

Si los regresores fuesen ortogonales...

... en el modelo con las dos variables explicativas en desviaciones considerado en el ejemplo, fijémonos que:

- $\mathrm{VAR}[\hat{\beta}_j] = \dfrac{\sigma_u^2}{\sum_{i=1}^{N}\tilde{X}_{ji}^2}$ $(j = 2, 3)$.
- $\mathrm{COV}[\hat{\beta}_2,\hat{\beta}_3] = 0$.

b) Dada la elevada varianza de los $\hat{\beta}_j$, las estimaciones de los parámetros resultan imprecisas y poco estables.

Una consecuencia inmediata de la anterior se ve en el contraste de la *t* de Student de significación individual de los parámetros. Al aumentar la varianza de los estimadores de forma artificial, tenderemos a no rechazar la H_0 con más frecuencia, a causa del hecho de que el denominador del estadístico de prueba (que es el error estándar del parámetro $\hat{\beta}_j$) será más elevado y, por tanto, el estadístico mencionado será más próximo a cero. Así pues, a partir del contraste de la *t* de Student podría concluirse que una variable explicativa es irrelevante, cuando en la realidad es significativa.

De esta manera, cuando hay multicolinealidad, una parte de la variabilidad de la variable endógena explicada por un regresor se comparte con otros regresores (a causa de la correlación entre éstos), y esta parte explicada en común no queda reflejada por el contraste de la *t* de Student. Si la correlación es muy elevada, puede considerarse que un parámetro no es estadísticamente significativo, cuando realmente sí que lo es.

Como volveremos a comentar más adelante, este problema no se da en el contraste de significación global, porque éste sí considera toda la explicación de la variabilidad de la variable endógena reunida por el conjunto de variables explicativas consideradas en el modelo, tanto la individual de cada una de éstas como la que comparte con otras.

c) Otra consecuencia derivada directamente de la elevada varianza de los $\hat{\beta}_j$ es que la varianza del predictor también será elevada y, por tanto, las predicciones por intervalo serán menos precisas. Recordad que en el cálculo del predictor interviene, entre otros términos, la varianza de los $\hat{\beta}_j$.

Consultad el predictor de las predicciones por intervalo en el subapartado 2.9.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

d) La imprecisión de las estimaciones del MRLM conduce al hecho de que, en presencia de multicolinealidad, no sea recomendable realizar un análisis estructural del modelo (es decir, basado en el valor obtenido de los parámetros) o que, por lo menos, sea recomendable tener en cuenta las consecuencias apuntadas antes con el fin de relativizar las conclusiones a las que se llegue.

2.2. Detección y valoración de la importancia de la multicolinealidad

Una vez que se han analizado las consecuencias negativas causadas por la presencia de multicolinealidad en el MRLM (cuanto mayor sea la correlación entre las variables explicativas más graves serán las consecuencias), es evidente que se debe analizar, en cada caso, si hay multicolinealidad y en qué grado se presenta.

No obstante, de acuerdo con lo que hemos comentado antes referente al hecho de que la situación más común es que haya un cierto nivel de multicolinealidad, más que la presencia o no de multicolinealidad en un modelo, nos interesa disponer de instrumentos que nos indiquen la intensidad con que ésta se presenta, para saber hasta qué punto afecta a los resultados del modelo.

La correlación de la población pequeña...

...acompañada de correlación muestral elevada se puede tener, por ejemplo, al utilizar un número reducido de observaciones, o una muestra con los individuos muy homogéneos, etc.

Hay que señalar, sin embargo, que no existen contrastes propiamente dichos para detectar la multicolinealidad ni su intensidad, dado que el problema no es tanto de la población como muestral. En este sentido, la correlación de la población entre las variables puede ser pequeña y, en cambio, ir acompañada de una correlación muestral elevada.

A pesar de la falta de contrastes, en la literatura se han propuesto muchos otros métodos para detectar la multicolinealidad*. Podemos destacar los ocho siguientes:

*** Detectar la multicolinealidad significa, en realidad, detectar la intensidad de multicolinealidad.**

1) Analizar de dos en dos los valores de los coeficientes de correlación simple entre los regresores. La razón de emplear este instrumento la encontramos en el concepto mismo de multicolinealidad. El nivel de multicolinealidad será más elevado cuanto mayor sea la correlación entre las variables explicativas. Así, coeficientes altos de correlación simple (correlación de dos en dos) entre las variables explicativas indicarán una multicolinealidad elevada.

2) El segundo método, relacionado también con el concepto de coeficiente de correlación, consiste en analizar la determinante de la matriz de correlaciones, $\mathbf{R}_X$, formada por los coeficientes de correlación entre las variables explicativas:

$$\mathbf{R}_X = \begin{bmatrix} r_{22} & r_{23} & \cdots & r_{2k} \\ & r_{33} & \cdots & r_{3k} \\ & & \ddots & \vdots \\ & & & r_{kk} \end{bmatrix}.$$

Los elementos de la diagonal...

... siempre valen 1, puesto que expresan la correlación de una variable consigo misma.

El valor de la determinante de $\mathbf{R}_X$ puede estar entre 0 y 1. Si hay multicolinealidad perfecta, el valor de la determinante será 0, mientras que en ausencia total de multicolinealidad valdrá 1. Así pues, cuanto más bajo sea el valor de la determinante, más nivel de multicolinealidad habrá en el modelo. Podemos expresar esto de la manera siguiente:

- $|\mathbf{R}_X| = 0 \Rightarrow$ multicolinealidad perfecta.
- $|\mathbf{R}_X| = 1 \Rightarrow$ ausencia total de multicolinealidad.

$|\mathbf{R}_X|$ simboliza la determinante de la matriz $\mathbf{R}_X$.

En el ámbito de un MRLM es preferible utilizar este instrumento que el anterior, dado que éste considera la correlación que se produce entre cualquier número de variables explicativas conjuntamente, mientras que el coeficiente de correlación solamente considera la correlación entre las variables explicativas de dos en dos.

Un análisis de multicolinealidad...

... de las variables de dos en dos no es suficiente. Si en un MRLM con $k = 4$ nos encontramos que $X_{4i} = a_1X_{2i} + a_2X_{3i}$, existirá multicolinealidad perfecta, pero puede ser que los coeficientes de correlación r_{42}, r_{43} y r_{23} no la reflejen. En este caso, la correlación estaría determinada por términos y no por pares.

3) El tercer método de detección de la multicolinealidad consiste en analizar los coeficientes de determinación de las regresiones en las cuales figure como variable endógena, sucesivamente, cada una de las variables explicativas del modelo:

$$X_{ji} = \delta_1 + \delta_2X_{2i} + \ldots + \delta_{j-1}X_{(j-1)i} + \delta_{j+1}X_{(j+1)i} + \ldots + \delta_kX_{ki} + v_i, \quad i = 1, \ldots, N. \qquad (2.1)$$

Si alguna de las estimaciones conduce a un coeficiente de determinación elevado*, indicará que la variable X_j que actúa como dependiente está altamente correlacionada con el resto de las variables explicativas y, por lo tanto, habrá multicolinealidad.

*** Por ejemplo, 0,8 o 0,9.**

4) El cuarto método consiste en analizar los parámetros estimados del modelo. La obtención de cambios significativos en la estimación de los parámetros del modelo al realizar pequeños cambios en los datos (por ejemplo, añadir o quitar al modelo una observación o un número reducido de observaciones) indica presencia de multicolinealidad.

La inestabilidad en las estimaciones...

... cuando se realizan pequeños cambios en los datos es consecuencia del aumento que la multicolinealidad provoca en las varianzas de los estimadores.

5) Adicionalmente, si obtenemos estimaciones de los parámetros muy alejadas de las previstas por la teoría económica (que incluso pueden llegar a cambiar de signo), puede ser a causa de la elevada varianza que presentan los estimadores cuando hay multicolinealidad en el modelo.

6) El sexto método de detección consiste en analizar los coeficientes de determinación de sucesivos MRLM en los cuales se elimina un regresor. Si al eliminar una variable explicativa del modelo la R^2 no se modifica de manera sustancial respecto a lo que se obtiene con todas las variables explicativas, esta variable es poco relevante para explicar la variable endógena (en caso de encontrarnos enfrente de una variable irrelevante) o lo que explica la variable eliminada ya queda explicado por otros regresores (a causa de la correlación entre las variables explicativas mencionadas), por lo cual en este último caso estaríamos ante un modelo con multicolinealidad intensa. En el caso hipotético de que la R^2 del modelo inicial coincidiese con la R^2 de la regresión en la cual hubiésemos eliminado una variable explicativa, estaríamos ante un modelo con multicolinealidad perfecta.

Consultad el coeficiente de determinación del ajuste de un modelo en el subapartado 2.6 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

7) La existencia de contradicción entre los resultados asociados al contraste de la *t* de Student de significación individual de parámetros y el contraste de la *F* de Snedecor de significación global del modelo es indicativa de una multicolinealidad intensa. Cuando los parámetros no sean estadísticamente significativos individualmente, pero sí que lo sean globalmente, todas las variables explicativas contribuirán equitativamente a la explicación de la variabilidad de la variable endógena. En consecuencia, la aportación individual de cada variable explicativa a la variación de la variable endógena (obtenida mediante el contraste de la *t* de Student) puede ser muy reducida y, en cambio, una gran parte de la explicación de la variable endógena será compartida por todas las variables explicativas al mismo tiempo.

Consultad los contrastes de significación individual y global en el subapartado 2.7.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Las aportaciones de cada regresor...

... a la explicación de la variable endógena se muestran en los diagramas del gráfico y se explican a continuación:

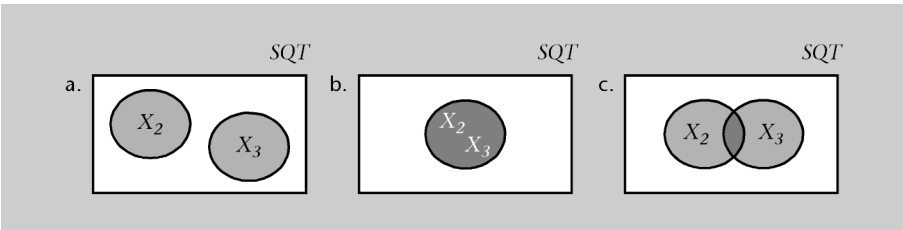
a. Esta hipótesis se corresponde con el caso de ausencia total de multicolinealidad.

b. Esta hipótesis se corresponde con el caso de multicolinealidad perfecta.

c. Esta hipótesis se corresponde con el caso intermedio de un cierto nivel de multicolinealidad.

Ejemplo de diferentes niveles de multicolinealidad

Sea el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, ..., N$. Según el nivel de multicolinealidad que haya, la aportación de cada regresor a la explicación de la variable endógena (obtenida esta última mediante la *SCT*) puede ser mayor o menor, tal como se explica a continuación:



a) En el caso de **ausencia total de multicolinealidad**, la correlación entre las variables explicativas es nula, $r_{23} = 0$, y toda la variación explicada por el modelo está determinada aisladamente por X_2 y X_3. Las respectivas t de Student asociadas a β_2 y β_3 no resultarían afectadas por una varianza mayor de los estimadores, es decir, no resultarían afectadas por una explicación de la variable endógena determinada conjuntamente por las variables explicativas y, por tanto, no habría contradicción entre los contrastes de la t de Student y de la F de Snedecor: ambos llevarían a rechazar las hipótesis nulas respectivas ($H_0 : \beta_2 = 0$ i $H_0 : \beta_3 = 0$, respectivamente).

b) En el caso de **multicolinealidad perfecta**, la correlación entre las variables explicativas es 1 (en valor absoluto), $|r_{23}| = 1$, lo cual significa que la aportación individual de cada una de las variables explicativas es nula, y que toda la variación explicada de la variable endógena está determinada conjuntamente por X_2 y X_3. Gráficamente, esta circunstancia se manifiesta en el hecho de que se superponen los dos círculos, tal como se ve en el diagrama **b** de este ejemplo, porque ambas variables explican lo mismo de Y. Entonces, los estadísticos t de significación individual llevarían a no rechazar la hipótesis nula, mientras que el estadístico F indicaría que el modelo es globalmente significativo* y, por tanto, existiría una contradicción entre los resultados que derivan de ambos contrastes.

*** El estadístico *F* considera tanto la aportación individual como la compartida de X_2 y X_3.**

c) Si hay un cierto **nivel de multicolinealidad**, una parte de la explicación de X_2 y X_3 deja de ser individual y pasa a ser compartida. Este hecho lo podemos observar en el gráfico anterior, donde la intersección entre los dos círculos del diagrama **c** del gráfico representa la parte de la explicación de Y compartida por los dos regresores, a causa de la correlación que existe entre ellos. Por lo tanto, si el nivel de correlación es suficientemente elevado, puede ser que el contraste t de significación individual (basado en la aportación individual de cada variable explicativa al comportamiento de la variable endógena) sea estadísticamente no significativo, mientras que el contraste F de significación global (que también incluye la aportación compartida) permita concluir la significación global del modelo.

Otra manera de explicar esta contradicción aparente entre los contrastes de significación de la t de Student y de la F de Snedecor consiste en recordar que, como hemos visto antes, la multicolinealidad provoca un aumento en la varianza del estimador $\hat{\beta}_j$ y, en consecuencia, el estadístico t de Student asociado a la significación estadística del parámetro β_j será más próximo a cero de lo que debería ser, lo cual lleva a no rechazar la hipótesis nula en más casos de los reales. Sin embargo, el valor del estadístico asociado al contraste de la F de Snedecor no resulta afectado por este hecho.

Consultad cómo la multicolinealidad provoca un aumento en la varianza del estimador $\hat{\beta}_j$ en el subapartado 2.1.3 de este módulo didáctico.

Esta contradicción aparente no nos debe llevar a concluir que hay multicolinealidad intensa en más casos de aquéllos en los que realmente se produce. No siempre que en un MRLM obtengamos un estadístico t asociado a algún regresor del modelo que nos indique no-rechazo de la H_0 y una F que indique significación global del modelo debemos concluir que hay multicolinealidad intensa. Existe una explicación mucho más sencilla: que estemos frente a una variable explicativa irrelevante. El resto de la información y de los instrumentos de detección de la multicolinealidad nos ayudará a saber en qué caso nos encontramos.

Asimismo, esta contradicción entre los contrastes de la t y la F no se detectará siempre que haya un nivel elevado de multicolinealidad. Una estimación muy elevada de los $\hat{\beta}_j$ respecto a su valor de la población o un valor de la población muy elevado de los parámetros β_j pueden ser dos razones que justifiquen la no-aparición de la contradicción entre los dos contrastes.

8) El último método de detección de multicolinealidad tratado consiste en calcular el **factor de incremento de la varianza*** de cada una de las variables explicativas. El *FIV* es un estadístico que nos permite saber si la varianza de un estimador está inflada por la presencia de multicolinealidad en el modelo respecto al caso de ortogonalidad, ya que relativiza la varianza que presenta

*** A partir de ahora nos referiremos al factor de incremento de la varianza con la notación *FIV*.**

el estimador mencionado respecto a la que presentaría en el supuesto de ortogonalidad entre los regresores (es decir, respecto a la varianza óptima). Recordemos que, en el caso de ausencia de correlación entre los regresores en un modelo de regresión donde las variables se han expresado en desviaciones respecto a la media, la varianza de los estimadores es la siguiente:

$$\mathrm{VAR}[\hat{\beta}_j^*] = \frac{\sigma_u^2}{\sum_{i=1}^{N} \tilde{X}_{ji}^2},$$

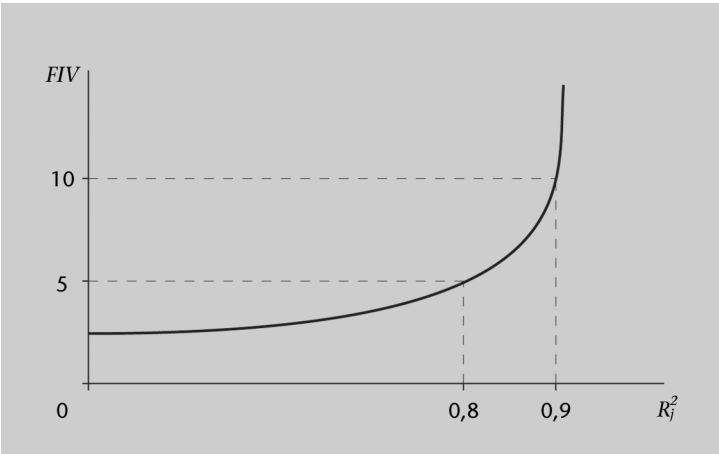
que es la varianza óptima* que puede alcanzar un estimador lineal y no sesgado en un MRLM estándar. Por otra parte, en general, la varianza de un estimador cualquiera es la siguiente:

$$\mathrm{VAR}[\hat{\beta}_j] = \frac{\sigma_u^2}{\sum_{i=1}^{N} \tilde{X}_{ji}^2 (1 - R_j^2)},$$

siendo R_j^2 el coeficiente de determinación de la regresión entre X_{ji} y el resto de las variables explicativas del modelo inicial (la regresión 2.1). Entonces, el FIV_j, es decir, el *FIV* del estimador *j*-ésimo será el siguiente:

$$FIV_j = \frac{\mathrm{VAR}[\hat{\beta}_j]}{\mathrm{VAR}[\hat{\beta}_j^*]} = \frac{1}{1 - R_j^2}. \qquad (2.2)$$

Por lo tanto, hay una relación inversa entre los FIV_j y los R_j^2, lo cual produce grandes crecimientos de los *FIV* ante pequeños cambios en los coeficientes de determinación a partir de valores de éstos superiores a 0,9, tal como podemos comprobar en el gráfico siguiente:



No existe un valor umbral de los *FIV* a partir del cual se tenga que afirmar que hay problemas graves de multicolinealidad. Como ya hemos comentado antes, es un problema que crece a medida que crece la correlación entre los regresores. De todas maneras, valores de $FIV_j > 5$ están asociados a $R_j^2 > 0{,}8$, y puede considerarse que las consecuencias sobre el MRLM ya pueden ser relevantes. En todo caso, esto se puede afirmar con más rotundidad en el caso de $FIV_j > 10$, asociado a $R_j^2 > 0{,}9$.

Lectura complementaria

Podéis encontrar la demostración de la expresión para la varianza de los estimadores, entre otras, en la obra siguiente:

A. Novales (1993). *Econometría* (2.ª ed., cap. 10). Madrid: McGraw-Hill.

*** La varianza óptima es la más pequeña.**

Consultad los problemas originados por un cierto nivel de multicolinealidad en el subapartado 2.1.3 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Existe una forma alternativa a 2.2. para calcular los FIV_j. Consiste en calcular la inversa de la matriz de correlaciones entre los regresores, $[\mathbf{R}_X]^{-1}$. Los elementos de la diagonal principal de $[\mathbf{R}_X]^{-1}$ son los FIV asociados a los diferentes regresores.

Otras consideraciones sobre el *FIV*

El *FIV* está acotado inferiormente por 1, hipótesis que se corresponde con el caso de ortogonalidad de los regresores.
En cambio, el *FIV* no está acotado superiormente: si hubiese multicolinealidad perfecta, la $R_j^2 = 1$ y el *FIV* sería infinito.

Ejemplo de análisis de un problema de multicolinealidad

Dado el MRLM $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t \quad t = 1, ..., N$, donde Y_t es el número de artículos vendidos, X_1, el precio del producto de una empresa competidora, y X_2, el gasto en publicidad, se dispone de ocho observaciones, que aparecen en la tabla siguiente:

Y	X_1	X_2
15	3,0	7
18	4,0	13
23	6,5	19
27	8,2	25
37	9,7	31
32	12,0	37
38	14,7	43
43	16,0	49

Se pide que se analice si existe un problema de multicolinealidad. Debe quedar claro que el ejemplo de análisis de un problema de multicolinealidad sirve para estudiar los contenidos teóricos del presente apartado. En realidad no sería recomendable estimar un modelo como éste con $k = 3$ y sólo ocho observaciones, debido a los pocos grados de libertad disponibles. La estimación MCO del modelo proporcionaría, entre otros, los resultados siguientes:

$$Y_t = 11{,}2146 - 2{,}5123X_{1t} + 1{,}4707X_{2t},$$

amb $R^2 = 0{,}9439$, $\bar{R}^2 = 0{,}9215$, y unos errores estándar asociados a los parámetros del modelo de 2,6384 y 0,8538, respectivamente.

Tenemos diferentes maneras de analizar la presencia de multicolinealidad en este modelo. Presentamos algunas a continuación:

- Puesto que el modelo tiene dos regresores, el primer conjunto de instrumentos consiste en analizar el coeficiente de correlación entre aquéllos. Se puede calcular que r_{12} es 0,9964, por lo cual queda claro que estamos frente a un modelo con una correlación elevada entre los dos únicos regresores y, por tanto, con una multicolinealidad elevada. Recordad lo siguiente:

$$r_{12} = \frac{\sum_{t=1}^{8} \tilde{X}_{1t}\tilde{X}_{2t}}{\sqrt{\sum_{t=1}^{8} \tilde{X}_{1t}^2}\sqrt{\sum_{t=1}^{8} \tilde{X}_{2t}^2}},$$

  donde X_1 y X_2 están expresadas en desviaciones respecto a la media.

- El segundo instrumento, equivalente al anterior en este ejemplo por el hecho de que $k = 3$, consiste en calcular la determinante de la matriz $\mathbf{R}_X$. Puede comprobarse que vale 0,0072, muy próximo a cero, por lo cual también indica multicolinealidad elevada.

- Otro instrumento consiste en analizar la constancia de las estimaciones modificando ligeramente la muestra. Así, si eliminamos, por ejemplo, la primera observación, se puede comprobar que las estimaciones MCO de los parámetros que acompañan a las variables explicativas son, respectivamente, –3,7987 y 1,9255. Estos valores difieren de manera significativa de las estimaciones iniciales: otro indicio de multicolinealidad*.

*** Recordad que los indicios de multicolinealidad se pueden deber tanto a la eliminación de algún individuo de la muestra como al cambio de uno por otro.**

- Otro indicio de multicolinealidad viene determinado por las mismas estimaciones obtenidas en el modelo inicial. Por ejemplo, es contradictorio respecto a la teoría económica que la relación entre "número de artículos vendidos" y "precio de la competencia" sea negativa, en el sentido de que, cuando crece una, decrece la otra (que es lo que indica un parámetro estimado negativo). Esta estimación anómala se confirma si se observa que el coeficiente de correlación entre ambas variables es positivo (0,9543) y que, estimando un MRLS entre ambas, el valor del parámetro estimado es 2,0.

- Otro indicador de multicolinealidad consiste en comparar la R^2 del modelo inicial con la R^2 asociada al modelo de regresión con una variable explicativa menos. Recordemos que, si la diferencia entre ambas era pequeña, indicaba multicolinealidad. En el ejemplo presente, si eliminamos la variable X_1, obtenemos una R^2 de 0,9337 y una de 0,9227 (incluso ligeramente más alta que la del modelo inicial). Por lo tanto, se obtiene otro indicio de multicolinealidad.

- Un nuevo instrumento que podría indicar problemas de multicolinealidad es la contradicción de resultados entre los contrastes *t* y *F* de significación individual y global del modelo. Si se calculan los estadísticos por el contraste de significación individual de los parámetros que acompañan a las variables explicativas (β_1 y β_2) se obtienen unas *t* de Student de –0,952 y 1,723, respectivamente, por lo cual no se rechaza la hipótesis nula de no-significación. En cambio, el estadístico *F* asociado a la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ vale 42,09, que comparado con el valor crítico en tablas de la *F* de Snedecor con 2 y 5 grados de libertad (5,79), lleva a rechazar la hipótesis nula. Por lo tanto, hay indicios de multicolinealidad.

- Para acabar, si calculamos los *FIV*, por ejemplo calculando la matriz inversa de $\mathbf{R}_X$, se obtiene que ambos valen 138,89. Estos *FIV*, mayores que 5, indican claramente una multicolinealidad elevada.

En el sexto caso,...

... si estimásemos un MRLS entre *Y* y cada una de las dos variables explicativas por separado, veríamos que cada uno de los dos parámetros estimados asociados a los regresores sería estadísticamente significativo, lo cual confirma que las *t* de Student de nuestro MRLM inicial son no significativas debido a un grado alto de multicolinealidad.

Actividad

2.1. Comprobad los resultados de la estimación MCO del modelo presentado en el ejemplo anterior.

2.3. Posibles soluciones a la presencia de multicolinealidad

Antes de entrar a estudiar las posibles soluciones a la multicolinealidad, se debe señalar que intentar eliminarla no tiene por qué ser un objetivo primordial. Dependerá de la finalidad con la que se haya construido el modelo. Si esta finalidad consiste en hacer predicciones, dado que la multicolinealidad no impide alcanzar un buen ajuste, y, en principio, hay que esperar que la correlación que existe entre las variables explicativas en el periodo muestral también se dé en el de predicción, nada impide que las predicciones surgidas de un modelo con multicolinealidad sean adecuadas. En cambio, si el objetivo es realizar un análisis estructural, sí que es necesario considerar el problema de la multicolinealidad.

Realizar un análisis estructural...

... implica conocer con precisión el impacto de las variables explicativas sobre la variable endógena, ver si las primeras son relevantes, etc.

Se debe advertir, no obstante, que aunque existen distintas maneras de solucionar el problema de la multicolinealidad, no siempre son satisfactorias y que, a veces, es preferible "convivir" con el problema y actuar teniéndolo en cuenta, antes que aplicarle soluciones que puedan suponer el incumplimiento de alguna de las hipótesis básicas del MRLM, lo cual, sin duda, generaría problemas más graves.

Recordad que el problema...

... de la multicolinealidad es una deficiencia de la información muestral, de los datos, y no del método de estimación, lo cual dificulta su solución.

2.3.1. Incorporación de nueva información

Una primera solución posible a la presencia de multicolinealidad consiste en incorporar más información al modelo. Esta información puede ser de muchos tipos, entre los cuales podemos citar dos: aumentar el tamaño muestral y utilizar información extramuestral (*a priori*).

Veámoslos con más detalle:

Aumentar el tamaño de la muestra

Dado que el problema de la multicolinealidad es un problema de la muestra, una primera vía de solución consiste en aumentar el tamaño muestral y, de esta manera, procurar reducir el problema de la correlación entre las variables explicativas. A pesar de ello, no siempre es posible aplicar esta solución, básicamente a causa de los tres hechos siguientes:

a) Si se trabaja con datos de corte temporal, normalmente no será posible disponer de más observaciones (si se hubiesen tenido inicialmente, ya se habrían incorporado desde el principio al proceso de estimación del modelo). De todos modos, una alternativa para aumentar el numero de observaciones podría consistir en cambiar la frecuencia de los datos (por ejemplo, pasar de anuales a trimestrales), pero esto no siempre es posible porque no siempre se dispondrá de esta información desagregada y, aunque se disponga de ella, puede suceder que la relación de la población cambie al variar la frecuencia de los datos, con lo cual se tendría que modificar el modelo especificado. Además, otro problema que puede aparecer al trabajar con datos de frecuencia mayor es la autocorrelación en el término de perturbación.

b) La introducción de nueva información en el modelo puede conducir a problemas de cambio estructural, si la nueva información se asocia a un subperiodo o a una realidad diferentes de lo que inicialmente se quería analizar.

c) La nueva información debe ser cualitativamente mejor que la inicial, en el sentido de que no reproduzca la misma correlación que los datos iniciales, sino que debe aportar nueva información a la estructura de relaciones entre las variables.

Utilizar información extramuestral

En cuanto a la segunda solución a la presencia de multicolinealidad, utilizar información extramuestral, consiste en restringir el valor de alguno de los parámetros del modelo inicial, de modo que se reduzca el número de parámetros iniciales y se mejore la eficiencia global de estimación. Esta información se obtendría de estudios previos realizados, de la incorporación de estimaciones realizadas de manera complementaria, etc.

Veamos un par de casos:

El primer caso sería aquel en que se conociese el cumplimiento de una restricción sobre los parámetros (por ejemplo, $\beta_2 = \beta_3 + 2\beta_1$). Incorporando la restricción al modelo, se reduciría el número de parámetros que hay que estimar y, probablemente, también se reduciría la multicolinealidad entre los regresores.

b) El segundo caso consistiría en que se incluyese en el modelo información sobre el valor de algún parámetro a partir de estimaciones complementarias. Estudiamos sus efectos a partir de un ejemplo.

El problema de la multicolinealidad...

... no resuelto se daría en el caso de que la restricción fuese estocástica y no determinista, por ejemplo $\beta_2 = \beta_3 + 2\beta_1 + v$ con $v \sim N(0,\sigma_v^2)$.

En este caso, la solución a la multicolinealidad es compleja por el hecho de haberse abordado la estimación *mixt*, que queda fuera del alcance de este material.

Supongamos que tenemos el modelo (con datos expresados en desviaciones) $\tilde{Y}_i = \beta_1\tilde{X}_{1i} + + \beta_2\tilde{X}_{2i} + u_i$, con $r_{12} \cong 1$. Se propone como solución la utilización de información extramuestral. Se obtiene una estimación auxiliar de β_1, que llamaremos $\overline{\beta}_1$. Se podría demostrar que $\overline{\beta}_2$ (la estimación de β_2 surgida del nuevo modelo en el que se ha incorporado la información extramuestral) es no sesgado si, y sólo si, $E[\overline{\beta}_1] = \beta_1$, pero que la varianza de $\overline{\beta}_2$ será mayor que la asociada al parámetro $\hat{\beta}_2$ del modelo inicial (suponiendo que el estimador $\overline{\beta}_1$, como mínimo, fuese no sesgado).

Este caso de incorporar información extramuestral o *a priori* podría consistir, por ejemplo, en incorporar una estimación obtenida a partir de un modelo con datos de corte transversal en un modelo inicial con datos temporales. No es necesario detallar que, muchas veces, el significado atribuible a un estimador obtenido con datos de corte transversal puede ser muy diferente del obtenido con datos temporales. Este caso sirve como ejemplo de las dificultades para considerar válido este tipo de soluciones propuestas.

2.3.2. Reespecificación del modelo

Un segundo conjunto posible de soluciones a la presencia de multicolinealidad gira en torno a reespecificar el modelo. Así, dentro de este conjunto de soluciones básicamente podemos distinguir entre la eliminación de alguna variable explicativa y la transformación de las variables del modelo.

A continuación enunciaremos las ventajas e inconvenientes de cada una de las dos soluciones mencionadas en el párrafo anterior:

1) La solución que consiste en eliminar alguna variable explicativa puede reducir (o incluso eliminar completamente) el problema de la multicolinealidad, aunque también es una solución que presenta limitaciones, entre las cuales podemos citar las dos siguientes:

a) La dificultad de elegir la variable que hay que excluir.

b) El hecho de caer en un problema de especificación incorrecta por omisión de variables relevantes.

Así, podemos sustituir un problema de estimación imprecisa, con una varianza elevada, causado por la multicolinealidad, por un problema de sesgo e inconsistencia de los estimadores, causado por la omisión de variables. Un criterio para decidir el mal menor puede consistir en analizar el ECM de los estimadores $\hat{\beta}_j$ en el modelo inicial (con un nivel elevado de multicolinealidad) y en el modelo reespecificado (sin una variable explicativa).

En este sentido, puede demostrarse que, si el estadístico *t* asociado a la variable que sería susceptible de eliminación en el modelo inicial (es decir, el modelo con todos los regresores) es menor que la unidad, entonces el ECM asociado al resto de los estimadores será más bajo en el modelo reespecificado (el modelo con menos regresores) y viceversa.

Lecturas complementarias

Podéis encontrar la demostración de la condición de eliminación de la variable del modelo en las obras siguientes:

J. Johnston (1987). *Métodos de econometría* (trad. J. Sánchez Fernández). Barcelona: Vicens-Vives.

G.G. Judge; R.C. Hill; W.E. Griffiths; H. Lütkepohl; T.C. Lee (1988). *Introduction to the Theory and Practice of Econometrics* (2.ª ed.). Nueva York: John Willey & Sons.

A. Novales (1993). *Econometría* (2.ª ed.). Madrid: McGraw-Hill.

Por lo tanto, el criterio para decidir cuál es el problema que más nos conviene plantear es el siguiente:

- Si el contraste de significación individual de los estimadores de β_j es inferior a 1, es preferible, en términos de ECM, eliminar la variable explicativa X_j del modelo inicial.

- Si el contraste de significación individual de los estimadores de β_j es superior a 1, es preferible, en términos de ECM, no eliminar la variable explicativa X_j del modelo inicial.

2) En cuanto a la otra solución, consistente en transformar las variables del modelo, tiene el inconveniente de que podemos violar las hipótesis básicas del modelo de regresión lineal múltiple (homoscedasticidad, no autocorrelación, etc.).

2.3.2. Estimación Ridge

Un tercer tipo de soluciones a la presencia de multicolinealidad consiste en cambiar de método de estimación. La solución habitual es estimar los parámetros por el método Ridge. Con este método se intentan evitar los problemas asociados a la estimación MCO causados por el valor reducido de la determinante de la matriz **X'X** cuando hay multicolinealidad.

La solución de la **estimación Ridge** consiste en sumar una determinada cantidad a los elementos de la diagonal principal de **X'X**. El estimador Ridge es $\tilde{\beta}_{\text{Ridge}} = [\mathbf{X'X} + c\mathbf{I}_k]^{-1}\mathbf{X'Y}$, siendo *c* una constante arbitraria positiva.

Este estimador es sesgado, y el sesgo aumenta con el valor de la constante *c*, aunque puede presentar unas varianzas para los estimadores Ridge menores que las del estimador tradicional MCO, si el valor de *c* se elige adecuadamente.

A partir de las propiedades analizadas, será preferible el estimador Ridge si su ECM es inferior al del estimador MCO. En todo caso, el problema reside en determinar el valor del escalar *c*.

2.4. Presencia de valores extraños: detección y tratamiento

En todo lo que hemos estudiado hasta ahora, los datos se han considerado como algo dado y no se han sometido a ningún tipo de análisis particularizado. No obstante, a veces, el análisis de las observaciones puede resultar fundamental, dado que puede evidenciar la existencia de comportamientos atípicos, en el sentido de que se pueden detectar observaciones por las cuales puede no resultar creíble que se hayan generado por el mismo proceso que el resto. Así pues, este análisis sirve para conocer mejor la muestra y sus efectos sobre los resultados del modelo.

Ejemplos de tratamiento de observaciones atípicas

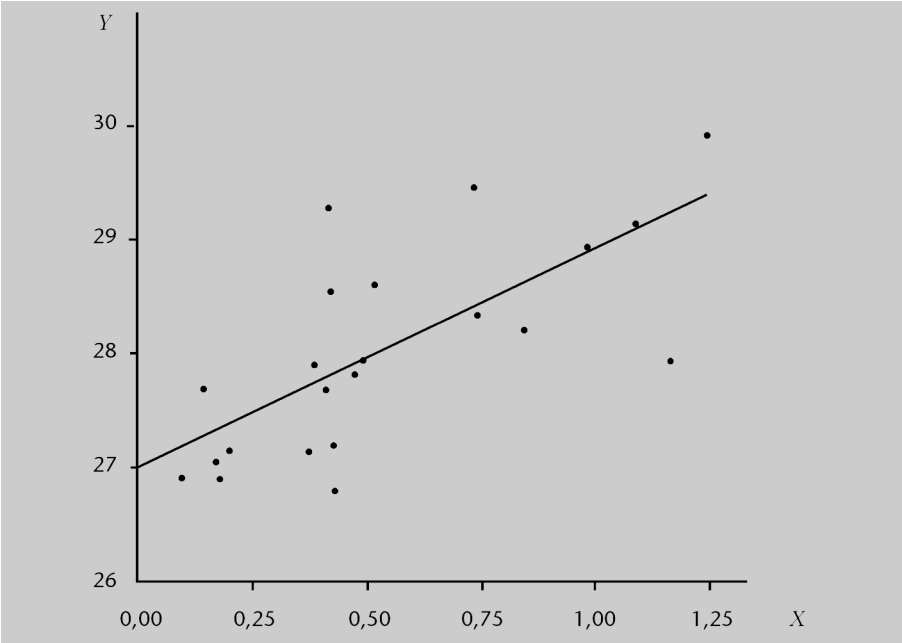
i	Y	X	Y_2	X_2	Y_3	X_3	Y_4	X_4	Y_5	X_5
1	26,9248	0,109324	26,9248	0,109324	26,9248	0,109324	26,9248	0,109324	26,9248	0,109324
2	27,7243	0,133186	27,7243	0,133186	27,7243	0,133186	27,7243	0,133186	27,7243	0,133186
3	27,0601	0,164674	27,0601	0,164674	27,0601	0,164674	27,0601	0,164674	27,0601	0,164674
4	26,9248	0,170578	26,9248	0,170578	26,9248	0,170578	26,9248	0,170578	26,9248	0,170578
5	27,343	0,176482	27,343	0,176482	27,343	0,176482	27,343	0,176482	27,343	0,176482
6	27,1585	0,192226	27,1585	0,192226	27,1585	0,192226	27,1585	0,192226	27,1585	0,192226
7	27,1585	0,36787	27,1585	0,36787	27,1585	0,36787	27,1585	0,36787	27,1585	0,36787
8	27,958	0,379309	27,958	0,379309	27,958	0,379309	27,958	0,379309	27,958	0,379309
9	27,712	0,405877	27,712	0,405877	27,712	0,405877	27,712	0,405877	27,712	0,405877
10	29,3233	0,411289	29,3233	0,411289	29,3233	0,411289	29,3233	0,411289	29,3233	0,411289
11	28,573	0,418546	28,573	0,418546	28,573	0,418546	28,573	0,418546	28,573	0,418546
12	27,22	0,42076	27,22	0,42076	27,22	0,42076	27,22	0,42076	27,22	0,42076
13	26,8264	0,424573	26,8264	0,424573	26,8264	0,424573	26,8264	0,424573	26,8264	0,424573
14	28,081	0,430846	28,081	0,430846	28,081	0,430846	28,081	0,430846	28,081	0,430846
15	27,835	0,465286	**33,7**	**0,443212**	27,835	0,465286	27,835	0,465286	27,835	0,465286
16	27,958	0,484966	27,835	0,465286	27,958	0,484966	27,958	0,484966	27,958	0,484966
17	28,6468	0,511042	27,958	0,484966	28,6468	0,511042	28,6468	0,511042	28,6468	0,511042
18	27,6505	0,537487	28,6468	0,511042	27,6505	0,537487	27,6505	0,537487	27,6505	0,537487
19	29,4832	0,730228	27,6505	0,537487	29,4832	0,730228	29,4832	0,730228	29,4832	0,730228
20	28,3516	0,73687	29,4832	0,730228	28,3516	0,73687	28,3516	0,73687	28,3516	0,73687
21	28,204	0,840313	28,3516	0,73687	28,204	0,840313	28,204	0,840313	28,204	0,840313
22	28,942	0,985699	28,204	0,840313	28,942	0,985699	28,942	0,985699	28,942	0,985699
23	29,1511	1,089265	28,942	0,985699	29,1511	1,089265	29,1511	1,089265	29,1511	1,089265
24	27,958	1,16245	29,1511	1,089265	27,958	1,16245	27,958	1,16245	27,958	1,16245
25	29,926	1,240309	27,958	1,16245	29,926	1,240309	29,926	1,240309	29,926	1,240309
26	—	—	29,926	1,240309	**30,2646**	**2,922869**	**24,3457**	**2,06547**	**30,5621**	**2,123778**

Para mostrar diferentes situaciones en las cuales se debería llevar a cabo un análisis más exhaustivo y particularizado de alguna de las observaciones de la muestra, dado su comportamiento distinto respecto al resto, planteamos un ejemplo. Supongamos que para estimar el modelo siguiente:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \forall i = 1, ..., N, \tag{2.3}$$

disponemos de un conjunto de veinticinco observaciones (observad las dos primeras columnas de la tabla anterior). A continuación añadimos a la muestra una observación adicional de características diferentes de las veinticinco observaciones originales. Este ejercicio lo hemos repetido cuatro veces (observad las ocho últimas columnas de la tabla anterior, donde, en negrita, señalamos la observación añadida en cada caso).

Estas observaciones que añadimos son extrañas, atípicas. Sin embargo, no todas tienen los mismos efectos, como podremos comprobar a continuación.



En el gráfico anterior mostramos la nube de puntos formada por las veinticinco observaciones originales, así como la recta de regresión que deriva de ella. En el cuadro que mostramos a continuación reunimos los resultados de la estimación del modelo obtenidos con estas veinticinco observaciones.

Entre otras cosas, podemos observar lo siguiente:

- El ajuste obtenido es bastante bueno.
- El término independiente y el parámetro que acompaña a la variable X son estadísticamente significativos.
- El modelo globalmente considerado también es significativo.

Resultados de la estimación del modelo 2.3 inicial

Variable dependiente: Y

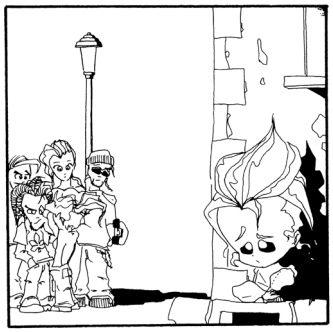
Número de observaciones: 25

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
Constant	**27.030013**	**0.2398103**	**112.71416**	0.0000
X	**1.8741038**	**0.3928246**	**4.7708407**	0.0001

R-squared	**0.497388**	Mean of dependent var	28.00376
Adjusted R-squared	0.475535	**S.D. of dependent var**	**0.869237**
S.E. of regression	**0.629501**	**Sum of squared resid**	**9.114254**
Log likelihood	-22.86051	**F-statistic**	**22.76092**
Durbin-Watson stat	2.120382	**Prob(F-statistic)**	**0.000082**

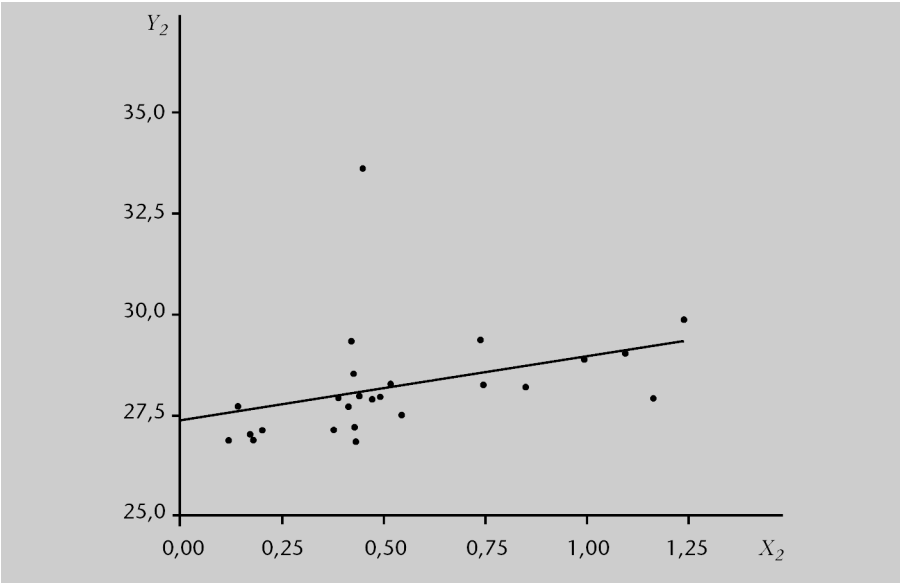
A continuación estudiamos los cuatro supuestos planteados:

1) En el primer supuesto (columnas correspondientes a Y_2 y a X_2 de la tabla de la página 33), la observación añadida es la decimoquinta, de coordenadas (0,4432, 33,7). Como podemos ver en el gráfico siguiente, donde presentamos la nube de puntos y la nueva recta de regresión, se trata de una observación atípica respecto al eje de ordenadas, pero no respecto al de abscisas, dado



La observación atípica queda sensiblemente apartada de la nube de puntos.

que aproximadamente se encuentra sobre la recta imaginaria que pasa por la media de las observaciones de la variable explicativa ($\bar{X}_2$ = 0,5166).



Notación del gráfico

El eje de coordenadas se corresponde con la variable endógena, y el eje de abscisas se corresponde con la variable explicativa.

Los resultados obtenidos al estimar los parámetros del modelo con estas veintiséis observaciones (observad el cuadro de resultados de la página siguiente) no cambian demasiado respecto a los obtenidos con las veinticinco observaciones originales (observad el cuadro de resultados anterior).

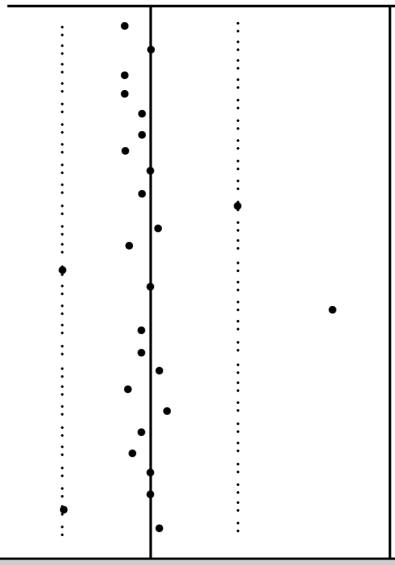
Resultados de la estimación del modelo 2.3 bajo la primera hipótesis

Variable dependiente: Y_2

Número de observaciones: 26

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
Constant	**27.340679**	**0.4977912**	**54.923993**	0.0000
X_2	**1.7074980**	**0.8229324**	**2.0748945**	0.0489

R-squared	**0.152099**	Mean of dependent var	28.22284
Adjusted R-squared	0.116770	**S.D. of dependent var**	**1.404749**
S.E. of regression	**1.320188**	**Sum of squared resid**	**41.82951**
Log likelihood	-43.07397	**F-statistic**	**4.305187**
Durbin-Watson stat	2.044177	**Prob(F-statistic)**	**0.048885**



obs	RESIDUAL	ACTUAL	FITTED
1	-0.60255	26.9248	27.5273
2	0.15621	27.7243	27.5681
3	-0.56176	27.0601	27.6219
4	-0.70714	26.9248	27.6319
5	-0.29902	27.3430	27.6420
6	-0.51040	27.1585	27.6689
7	-0.81032	27.1585	27.9688
8	-0.03035	27.9580	27.9883
9	-0.32171	27.7120	28.0337
10	1.28035	29.3233	28.0430
11	0.51765	28.5730	28.0553
12	-0.83913	27.2200	28.0591
13	-1.23924	26.8264	28.0656
14	0.00465	28.0810	28.0763
15	**5.60254**	**33.7000**	**28.0975**
16	-0.30015	27.8350	28.1352
17	-0.21076	27.9580	28.1688
18	0.43352	28.6468	28.2133
19	-0.60794	27.6505	28.2584
20	0.89566	29.4832	28.5875
21	-0.24728	28.3516	28.5989
22	-0.57151	28.2040	28.7755
23	-0.08176	28.9420	29.0238
24	-0.04950	29.1511	29.2006
25	-1.36756	27.9580	29.3256
26	0.46750	29.9260	29.4585

De todos modos, la consideración de esta observación adicional tiene efectos significativos sobre los factores que mencionamos a continuación:

a) Los errores estándar de los estimadores (son mayores), con lo cual se reduce la significación de las variables y se está a punto de no rechazar que β_2 valga cero.

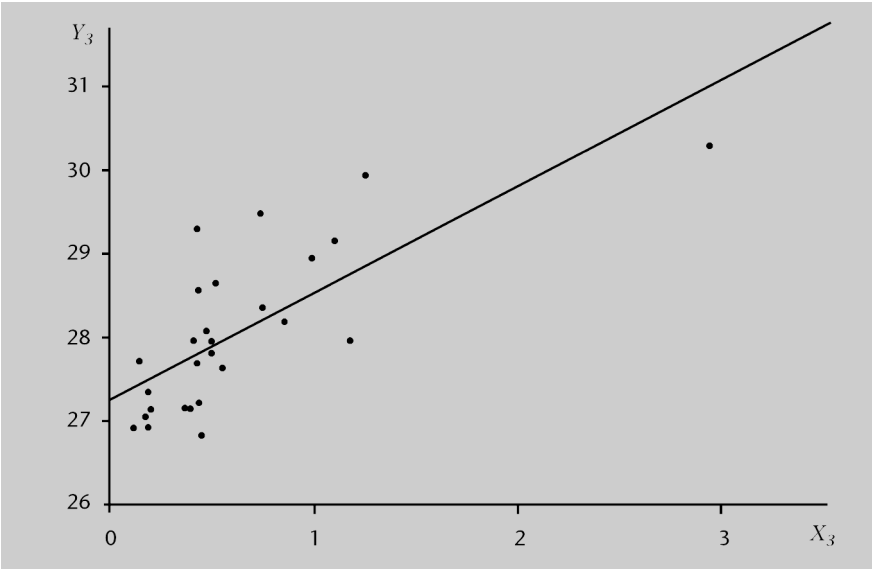
b) El ajuste global del modelo (observad la variación experimentada por la R^2 y por el estadístico de la F de Snedecor).

c) La *SCE*.

d) La estimación de la varianza del término de perturbación.

Además, en el gráfico de los residuos (*Residual plot*) podemos observar que el error cometido en el ajuste para esta observación es, respecto al resto, mucho mayor.

2) En el supuesto siguiente (Y_3 y X_3), la observación añadida a las veinticinco originales es la vigesimosexta. Sus coordenadas son las siguientes: (2,922869, 30,2646). Como podemos observar en el gráfico de la derecha, se trata de una observación extraña respecto al eje de abscisas: está bastante alejada de la recta imaginaria que pasa por $\bar{X}_3 = 0{,}61201$.



A pesar de todo lo que hemos visto con anterioridad, el hecho de considerar esta observación no hace que cambie de manera significativa ninguno de los resultados obtenidos al estimar el modelo con las veinticinco observaciones originales: ni los parámetros estimados, ni los errores estándar, ni la R^2, ni el estadístico de la F de Snedecor del contraste de significación global del modelo han sufrido grandes variaciones (observad el cuadro anterior y comparad los resultados con los obtenidos en el cuadro inicial de veinticinco datos). Además, fijaos en que los residuos asociados al ajuste de cada una de las veinticinco observaciones originales son muy semejantes en el primer supuesto y en éste, y que, a diferencia del supuesto anterior, el residuo asociado a la observación añadida no es grande si se compara con los de las veinticinco observaciones restantes. Fijaos también en que una única observación ha hecho disminuir el valor del parámetro estimado correspondiente a la pendiente, que pasa de 1,87 a 1,24.

Resultados de la estimación del modelo 2.3 bajo la segunda hipótesis

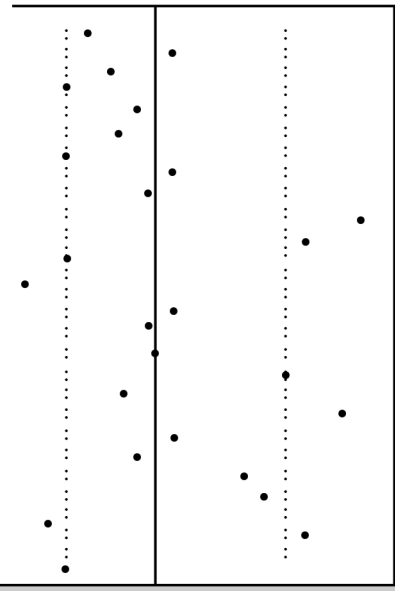
Variable dependiente: Y_3

Número de observaciones: 26

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
Constant	27.334353	0.1936790	141.13226	0.0000
X_3	1.2358544	0.2336824	5.2886078	0.0000

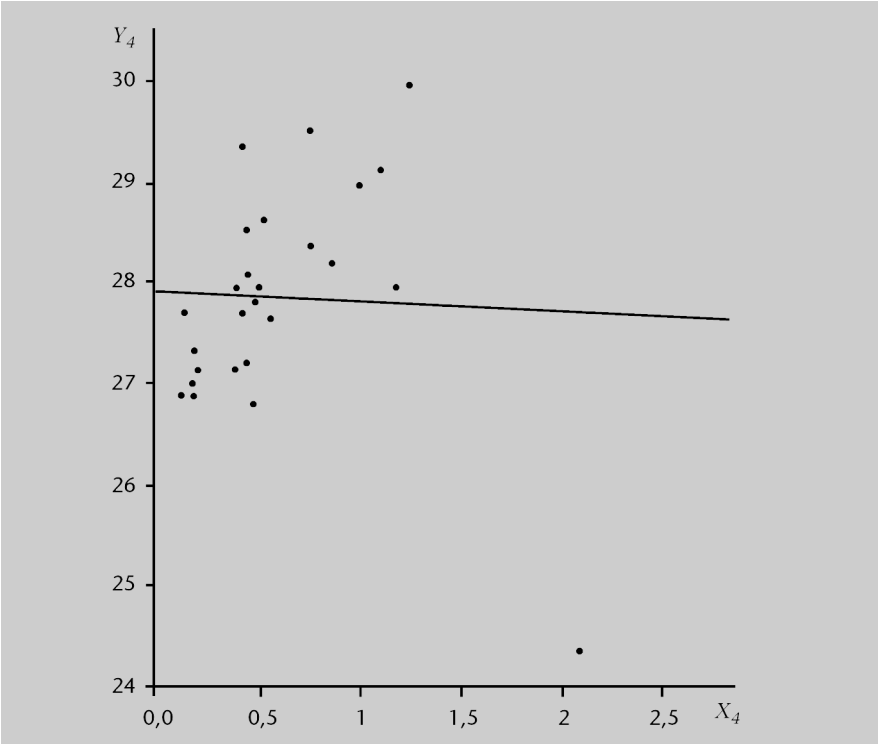
R-squared	**0.538190**	Mean of dependent var	28.09071
Adjusted R-squared	0.518947	**S.D. of dependent ver**	**0.960179**
S.E. of regression	**0.665961**	**Sum of squared resid**	**10.64408**
Log likelihood	-25.28220	**F-statistic**	**27.96937**
Durbin-Watson stat	2.138500	**Prob(F-statistic)**	**0.000020**

Residual Plot



obs	RESIDUAL	ACTUAL	FITTED
1	-0.54466	26.9248	27.4695
2	0.22535	27.7243	27.4990
3	-0.47777	27.0601	27.5379
4	-0.62036	26.9248	27.5452
5	-0.20946	27.3430	27.5525
6	-0.41342	27.1585	27.5719
7	-0.63049	27.1585	27.7890
8	0.15488	27.9580	27.8031
9	-0.12396	27.7120	27.8360
10	1.48065	29.3233	27.8426
11	0.72138	28.5730	27.8516
12	-0.63435	27.2200	27.8544
13	-1.03266	26.8264	27.8591
14	0.21418	28.0810	27.8668
15	-0.07438	27.8350	27.9094
16	0.02430	27.9580	27.9337
17	0.68087	28.6468	27.9659
18	-0.34811	27.6505	27.9986
19	1.24639	29.4832	28.2368
20	0.10658	28.3516	28.2450
21	-0.16886	28.2040	28.3729
22	0.38947	28.9420	28.5525
23	0.47057	29.1511	28.6805
24	-0.81297	27.9580	28.7710
25	1.05881	29.9260	28.8672
26	**-0.68199**	**30.2646**	**30.9466**

3) En el tercer supuesto (Y_4 y X_4), la observación añadida, la vigesimosexta, de coordenadas (2,0655, 24,3457), es extraña respecto al resto, tanto en lo referente al eje de ordenadas como al de abscisas (la media de X_4 vale 0,5790). Podéis observar este hecho en el gráfico siguiente, donde presentamos la nube de puntos y la recta de regresión asociada a ella.



Los resultados de la estimación (observad el cuadro siguiente) cambian muy drásticamente al compararlos con los obtenidos con las veinticinco observa-

ciones originales. De hecho, incluso la pendiente de la recta es ahora negativa, cuando inicialmente era positiva. Además, en este supuesto no se puede rechazar la hipótesis nula del contraste de significación individual del parámetro asociado a la variable X_4, ni la del contraste de significación del modelo, y la R^2 toma un valor muy próximo a cero.

Fijaos en cómo cambian radicalmente...

... la interpretación y las conclusiones que derivan del ajuste realizado por el solo hecho de considerar en la muestra una observación con las características del tercer supuesto.

En cuanto a los residuos del ajuste, podemos comprobar que son mucho mayores para todas las observaciones que los obtenidos con las veinticinco observaciones originales y, en particular el residuo asociado a la observación añadida es mucho mayor que el del resto de las observaciones.

Resultados de la estimación del modelo 2.3 bajo la tercera hipótesis

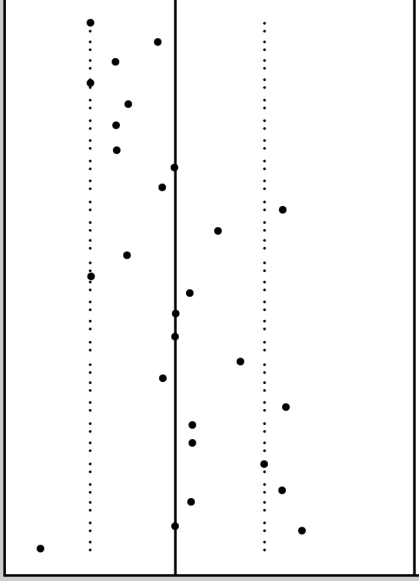
Variable dependiente: Y_4

Número de observaciones: 26

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
Constant	27.937407	0.3719205	75.116601	0.0000
X_4	-0.1283950	0.5145596	-0.2495241	0.8051

R-squared	0.002588	Mean of dependent var	27.86306
Adjusted R-squared	-0.038971	**S.D. of dependent var**	**1.113561**
S.E. of regression	1.135053	**Sum of squared resid**	**30.92026**
Log likelihood	-39.14550	**F-statistic**	**0.062262**
Durbin-Watson stat	1.646759	**Prob(F-statistic)**	**0.805079**

Residual Plot

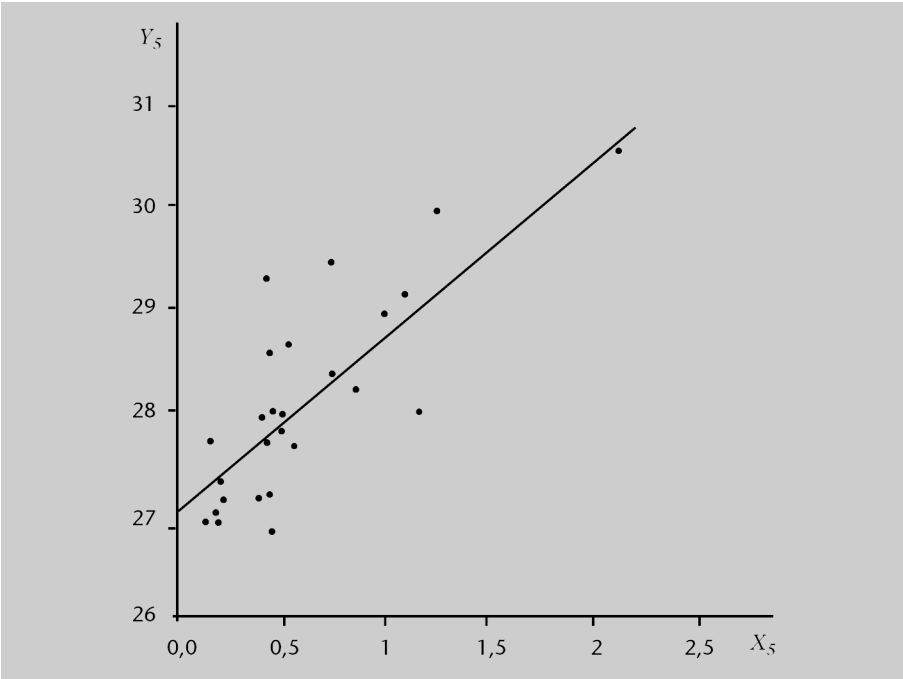


obs	RESIDUAL	ACTUAL	FITTED
1	-0.99857	26.9248	27.9234
2	-0.19601	27.7243	27.9203
3	-0.85616	27.0601	27.9163
4	-0.99070	26.9248	27.9155
5	-0.57175	27.3430	27.9147
6	-0.75423	27.1585	27.9127
7	-0.73167	27.1585	27.8902
8	0.06929	27.9580	27.8887
9	-0.17329	27.7120	27.8853
10	1.43870	29.3233	27.8846
11	0.68933	28.5730	27.8837
12	-0.66338	27.2200	27.8834
13	-1.05649	26.8264	27.8829
14	0.19891	28.0810	27.8821
15	-0.04267	27.8350	27.8777
16	0.08286	27.9580	27.8751
17	0.77501	28.6468	27.8718
18	-0.21790	27.6505	27.8684
19	1.63955	29.4832	27.8436
20	0.50880	28.3516	27.8428
21	0.37449	28.2040	27.8295
22	1.13115	28.9420	27.8108
23	1.35355	29.1511	27.7976
24	0.16985	27.9580	27.7882
25	2.14784	29.9260	27.7782
26	**-3.32651**	**24.3457**	**27.6722**

4) Para terminar, en el cuarto supuesto que se plantea (Y_5 y X_5), la observación añadida (observad el gráfico siguiente) es la vigesimosexta, cuyas coordenadas son (2,1238, 30,5621). Se trata de un caso parecido al que hemos considerado antes con las variables Y_3 y X_3, aunque ahora la observación añadida no es tan extraña respecto al eje de abscisas como lo era en aquel supuesto.

No obstante, los resultados obtenidos al estimar el modelo, que aparecen en el cuadro que veremos más adelante, no cambian demasiado respecto a los

obtenidos con las veinticinco observaciones originales (de la misma manera que ocurría en el supuesto de Y_3 y X_3): los errores estándar, la *SCE* y la estimación de la varianza del término de perturbación son muy parecidos; la diferencia principal reside en la estimación de la pendiente de la recta de regresión.



Los ejemplos anteriores ponen de manifiesto que, efectivamente, ante la representación gráfica de la nube de puntos, hay que someter las observaciones que visualmente parezcan extrañas a un análisis particularizado, para averiguar cuáles son (o pueden ser) las consecuencias que puede tener su consideración en la estimación como si se tratase de una observación más y, de esta manera, evitar caer en el error de extraer conclusiones equivocadas por la presencia de alguna observación extraña en la muestra.

De todos modos, lo que hemos visto antes no debe entenderse en el sentido de que la situación ideal es aquella en que los individuos que integran la muestra son muy homogéneos entre sí. Muy al contrario, en los casos en los cuales las observaciones responden a una realidad económica relevante, la riqueza del análisis reside en la heterogeneidad, pero si esta heterogeneidad es una consecuencia de la existencia de errores de medida, de un hecho atípico no relevante para el conjunto del análisis, etc., es necesario detectar y tratar las observaciones adecuadamente para que no distorsionen el análisis.

En cualquier caso, hay que tener en cuenta que la detección de las observaciones atípicas se hace más difícil en los modelos con más de una variable explicativa, ya que no es posible hacer los gráficos anteriores para todas las variables del

modelo a la vez. Además, no nos interesa solamente ver si hay observaciones extrañas, sino también evaluar su influencia sobre los resultados.

Con esta finalidad, en la literatura sobre la materia se han propuesto medidas y estadísticos distintos, entre los cuales podemos destacar los siguientes: el *lever*, los residuos del ajuste, los residuos con pequeñas transformaciones (residuos estandarizados, residuos estudentizados y residuos estudentizados con omisión) y la distancia de Cook. En los subapartados siguientes analizaremos cada uno de estos aspectos y los aplicaremos a los cuatro supuestos planteados a lo largo de las páginas anteriores.

Resultados de la estimación del modelo 2.3 bajo la cuarta hipótesis

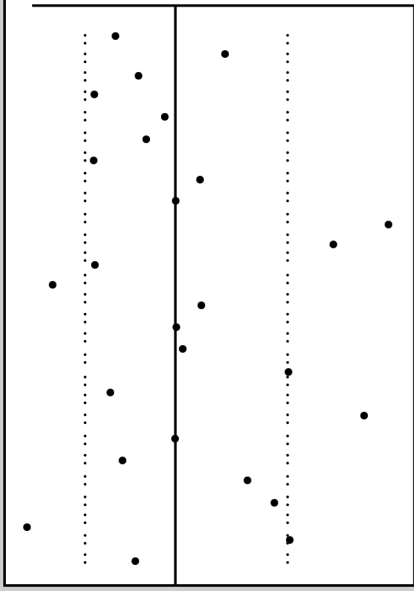
Variable dependiente: Y_5

Número de observaciones: 26

VARIABLES	COEFICIENTE	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
Constant	27.092455	0.2012090	134.64829	0.0000
X_5	1.7370315	0.2759070	6.2957137	0.0000

R-squared	**0.622855**	Mean of dependent var	28.10215
Adjusted R-squared	0.607141	**S.D. of dependent var**	**0.988477**
S.E. of regression	**0.619562**	**Sum of squared resid**	**9.212578**
Log likelihood	-23.40455	**F-statistic**	**39.63601**
Durbin-Watson stat	2.193903	**Prob(F-statistic)**	**0.000002**

Residual Plot



obs	RESIDUAL	ACTUAL	FITTED
1	-0.35755	26.9248	27.2824
2	0.40050	27.7243	27.3238
3	-0.31840	27.0601	27.3785
4	-0.46395	26.9248	27.3888
5	-0.05601	27.3430	27.3990
6	-0.26786	27.1585	27.4264
7	-0.57296	27.1585	27.7315
8	0.20667	27.9580	27.7513
9	-0.08548	27.7120	27.7975
10	1.51642	29.3233	27.8069
11	0.75352	28.5730	27.8195
12	-0.60333	27.2200	27.3382
13	-1.00355	26.8264	27.8300
14	0.24015	28.0810	27.8408
15	-0.06567	27.8350	27.9007
16	0.02314	27.9580	27.9349
17	0.66665	28.6468	27.9802
18	-0.37559	27.6505	28.0261
19	1.12232	29.4832	28.3609
20	-0.02082	28.3516	28.3724
21	-0.34811	28.2040	28.5521
22	0.13735	28.9420	28.8046
23	0.16656	29.1511	28.9845
24	-1.15367	27.9580	29.1117
25	0.67909	29.9260	29.2469
26	**-0.21943**	**30.5621**	**30.7815**

2.4.1. Apalancamiento de una observación: *el leverage*

Una observación presenta apalancamiento si está muy alejada del resto de las observaciones en cuanto a las coordenadas de las variables explicativas.

Para analizar el apalancamiento de las observaciones, en la literatura sobre la materia se han propuesto diferentes estadísticos, pero quizá el más empleado es el llamado *leverage*.

A partir de ahora abreviaremos *leverage* por *lever*.

Hay un *lever* asociado a cada observación; se acostumbra a simbolizar mediante h_{ii}, y no es más que el elemento *i*-ésimo de la diagonal principal de la matriz **H** (llamada en inglés *hat matrix*). Esta matriz **H** se obtiene a partir de la expresión siguiente:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'.$$

La matriz es simétrica, idempotente y su traza* es *k*.

*** Es decir, la suma de los elementos de la diagonal.**

El *lever* presenta otras propiedades, como las siguientes:

a) El *lever* de una observación es mayor cuanto más se diferencie, en términos de las variables explicativas, del resto de las observaciones.

b) Los *lever* están acotados inferiormente por $1/N$ y superiormente por 1:

$$\frac{1}{N} \leq h_{ii} \leq 1.$$

Observad que la cota inferior de los *lever*...

... no es cero, pese a que, a medida que aumenta la muestra, la cota inferior efectivamente tiende a cero.

En el supuesto de que el *lever* de una observación alcance la cota inferior, esta observación no presentará nada de apalancamiento. En este supuesto, la observación se encuentra sobre la media de las variables explicativas, y, en consecuencia, la distancia desde la observación hasta la media de las *X* es cero. Por el contrario, si el *lever* vale 1, la observación alcanza el máximo apalancamiento posible.

De todos modos, el supuesto más habitual que se suele dar en la práctica es aquel en que el *lever* se encuentra en el interior del intervalo $[1/N,1]$. Pues bien, puesto que los *lever* (del mismo modo que los regresores) se consideran fijos, no siguen ninguna distribución de probabilidad y, en consecuencia, la determinación del apalancamiento de las observaciones se lleva a término a partir de un criterio *ad hoc*.

¿Cómo se puede averiguar si una observación presenta apalancamiento?

En concreto, se considera que una observación presenta apalancamiento si su *lever* es mayor que dos veces la media de los *lever* ($\bar{h}$). Lo concretamos en la regla práctica siguiente*:

- Si $h_{ii} \geq 2\bar{h} \Rightarrow$ la observación *i*-ésima presenta apalancamiento.
- Si $h_{ii} < 2\bar{h} \Rightarrow$ la observación *i*-ésima no presenta apalancamiento.

*** Algunos autores señalan como cota para la existencia de apalancamiento tres veces la media de los lever, en lugar de dos.**

Ejemplo de determinación del apalancamiento

Una vez que hemos visto qué es y cómo hay que determinar el apalancamiento de las observaciones de la muestra, a continuación aplicamos este concepto a los cuatro supuestos planteados anteriormente. A partir de los datos de la tabla que nos muestra los supuestos, hemos calculado los lever asociados a cada una de las veintiséis observaciones en cada uno de los cuatro casos considerados. Los resultados obtenidos los presentamos en un cuadro que está disponible en el material asociado de la asignatura.

Consultad los cuatro supuestos planteados en el inicio del apartado 2.4 de este módulo didáctico.

Como sabemos, se dice que una observación presenta apalancamiento si su *lever* es mayor que dos veces la media de los *lever*. Dado que en los cuatro supuestos planteados k es 2 y N vale 26, y teniendo en cuenta que la suma de los *lever* coincide con la traza de **H** y ésta con el número de parámetros del modelo, la media de los *lever* en todos los supuestos considerados es la siguiente:

$$\bar{h} = \frac{2}{26} = 0{,}0769.$$

Dos veces la media de los *lever* es 0,1538. Teniendo en cuenta este resultado y los *lever* calculados, podemos concluir lo siguiente:

Supuesto	Observación añadida	¿Apalancada?	Otras observaciones que presentan apalancamiento
Y_2, X_2	15	No	24, 25 i 26*
Y_3, X_3	26	Sí	Ninguna
Y_4, X_4	26	Sí	Ninguna
Y_5, X_5	26	Sí	Ninguna

* Estas tres observaciones salen apalancadas porque, al añadir la observación número quince, el lever medio se reduce mucho.

2.4.2. Residuo, residuo estandarizado, residuo estudentizado y residuo estudentizado con omisión

Los *outliers* son aquellas observaciones (individuos) atípicas. Son extrañas o raras en el sentido de que son observaciones que tienen un proceso generador diferente del que tiene el resto de las observaciones.

Para detectar las observaciones atípicas existen diferentes métodos. Os presentamos los métodos de detección de estas observaciones más básicos (los más clásicos) a continuación:

a) Estudiar el **residuo** asociado a cada individuo. En principio parece evidente que aquellos individuos que tengan asociado un proceso generador diferente del que tiene el resto tendrán asociado un residuo en el ajuste mayor que el resto de los individuos. Sin embargo, el problema de emplear los residuos para detectar los *outliers* deriva del hecho de que su magnitud depende de las unidades de medida en que estén expresadas las variables.

En cualquier caso, como regla práctica, se suele emplear la siguiente:

- Si $|e_i| \geq 2\dfrac{\sum_{i=1}^{N}|e_i|}{N}$ $\Rightarrow$ la observación *i*-ésima puede considerarse un *outlier*.
- Si $|e_i| < 2\dfrac{\sum_{i=1}^{N}|e_i|}{N}$ $\Rightarrow$ la observación *i*-ésima no puede considerarse un *outlier*.

En esta notación $|e_i|$ es el residuo de la estimación MCO del modelo asociado a la observación *i*-ésima.

b) Analizar el **residuo estandarizado**, que es el siguiente:

$$\frac{e_i}{\sqrt{\dfrac{\mathbf{e'e}}{N-k}}} = \frac{e_i}{\hat{\sigma}_u}.$$

Nota: recordad que la estandarización elimina las unidades de medida.

En tal caso, se emplea la regla práctica siguiente:

- Si $\dfrac{e_i}{\hat{\sigma}_u} \geq 2 \Rightarrow$ la observación *i*-ésima puede considerarse un *outlier*.
- Si $\dfrac{e_i}{\hat{\sigma}_u} < 2 \Rightarrow$ la observación *i*-ésima no puede considerarse un *outlier*.

c) Para evitar interferencias entre observaciones, se suele utilizar el llamado **residuo estudentizado**. El residuo estudentizado, que simbolizaremos por r_i, a diferencia del residuo estandarizado, pondera el error de ajuste MCO asociado a la observación *i*-ésima por su desviación estándar, en lugar de hacerlo por la desviación estándar del término de perturbación. Es decir:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}_u^2(1-h_{ii})}} = \frac{e_i}{\sqrt{\dfrac{\mathbf{e'e}}{N-k}(1-h_{ii})}},$$

donde $\sqrt{\dfrac{\mathbf{e'e}}{N-k}(1-h_{ii})}$ es la desviación estándar del error asociado a la observación *i*-ésima.

Matriz de varianzas y covarianzas

Recordad que la matriz de varianzas y covarianzas de los errores del ajuste MCO se expresa de la manera siguiente:

$$\mathrm{VAR}[\mathbf{e}] = \sigma_u^2\mathbf{M},$$

Consultad la matriz de varianzas y covarianzas en el subapartado 2.3.2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

donde **M** es una matriz que se define como la diferencia entre la matriz identidad (de dimensión N x N) y la matriz **H** : **M** = $\mathbf{I}_N$ - **H**. Así pues, la varianza del error i-ésimo es la siguiente:

$$\mathrm{VAR}[e_i] = \sigma_u^2(1 - h_{ii}).$$

El residuo estudentizado se distribuye asintóticamente según una t de Student con $N - k$ grados de libertad.

En tal caso, la regla que utilizamos es la siguiente:

- Si $r_i \geq t_{N-k;\alpha/2} \Rightarrow$ la observación i-ésima puede considerarse un *outlier*.
- Si $r_i < t_{N-k;\alpha/2} \Rightarrow$ la observación i-ésima no puede considerarse un *outlier*.

d) Hay autores que defienden la utilización de lo que llamamos **residuo estudentizado con omisión** como estadístico para averiguar si una observación es un *outlier* o no lo es. De hecho, este estadístico, que se simboliza por r_i^*, no es más que una corrección del anterior, donde en lugar de emplear la *SCE* asociada a la estimación con todas las observaciones, se usa la asociada a la estimación sin considerar la observación analizada. Una vez que se ha calculado el residuo estudentizado con omisión, se utiliza el mismo criterio de antes.

Uso del residuo estudentizado con omisión

Fijaos en que, si utilizamos el residuo estudentizado con omisión, eliminaremos la observación analizada (la i-ésima) y calcularemos la *SCE*. Después, la utilizaremos en el cálculo del residuo dividiéndolo por $N - k - 1$.

Ejemplo de análisis de la posible presencia de outliers en la muestra

A continuación, igual que hemos hecho cuando hemos estudiado el apalancamiento, presentamos en la web de la asignatura los resultados obtenidos al calcular el residuo estudentizado para cada una de las veintiséis observaciones en los cuatro supuestos planteados anteriormente, para averiguar la posible presencia de *outliers* en la muestra.

Consultad los cuatro supuestos planteados en el inicio del apartado 2.4 de este módulo didáctico.

De acuerdo con lo que hemos expuesto en las páginas anteriores, tenemos lo siguiente:

- El residuo estudentizado se distribuye asintóticamente según una t de Student con $N - k$ grados de libertad, es decir, en nuestro caso con 26 − 2 = 24 grados de libertad.
- Por otra parte, como sabemos, si el residuo estudentizado (en valor absoluto) asociado a una observación es mayor que el valor crítico proporcionado por la tabla de la t de Student, podemos afirmar que la observación en cuestión es un *outlier.*

Por lo tanto, teniendo en cuenta, por un lado, los resultados obtenidos y, por el otro, que el valor crítico para un nivel de significación del 5% y con 24 grados de libertad es 2,064, llegamos a las conclusiones siguientes:

Supuesto	Observación añadida	¿r_i elevado?	Otras observaciones con r_i elevado
Y_2, X_2	15	Sí	Ninguna
Y_3, X_3	26	No	10 (y tal vez la 19)
Y_4, X_4	26	Sí	25
Y_5, X_5	26	No	10 y 24

2.4.3. Distancia de Cook

La distancia de Cook es una medida que permite detectar la extrañeza de una observación valorando tanto el apalancamiento que presenta como la concordancia del proceso que la ha generado con el proceso generador del resto de las observaciones.

La distancia de Cook permite detectar, pues, aquellas observaciones que tienen un efecto mayor en el ajuste que el resto, y que por sí solas pueden hacer cambiar los valores estimados por los parámetros del modelo de una manera sustancial. Además, la existencia de una o de más observaciones de este tipo en la muestra puede hacer que se incumplan algunas de las hipótesis básicas del modelo de regresión, como, por ejemplo, la linealidad de la forma funcional y la normalidad del término de perturbación.

¿Cómo se puede detectar la presencia de observaciones de este tipo en la muestra?

En la literatura sobre la materia se pueden encontrar diferentes propuestas para detectar observaciones de este tipo en la muestra. Una de estas propuestas es la llamada **distancia de Cook**, que, para la observación *i*-ésima, se define del modo siguiente:

$$DC_i = \frac{\dfrac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^i)'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^i)}{k}}{\dfrac{\mathbf{e}'\mathbf{e}}{N-k}},$$

$\hat{\mathbf{Y}}$ y e'e responden a las definiciones habituales.

donde $\hat{\mathbf{Y}}^i$ representa la estimación obtenida sin tener en consideración la observación *i*-ésima, de la cual se desea averiguar la influencia en el ajuste; $\hat{\mathbf{Y}}$ es la estimación obtenida empleando todas las observaciones, y $\mathbf{e}'\mathbf{e}$ denota la *SCE* asociada al modelo estimado con todas las observaciones.

Cuando se elimina la observación *i*-ésima quedan $N - 1$ observaciones. Con estas observaciones se estima el vector de parámetros **B**. Después se calcula el vector $\hat{\mathbf{Y}}^i$ tomando **X** y la estimación de **B** y agrupando a todas las observaciones en **X**, incluida la observación *i*-ésima.

Precisiones sobre el número de parámetros y las dimensiones

Observad que:

$$\hat{\mathbf{Y}} = \mathbf{X}^i\hat{\mathbf{B}}^i, \text{ on } \hat{\mathbf{B}}^i = (\mathbf{X}^{i\prime}\mathbf{X}^i)^{-1}\,\mathbf{X}^{i\prime}\mathbf{Y}^i,$$

es decir, que en la matriz **X** y el vector **Y** se ha eliminado la fila i-ésima, y, por tanto, $\mathbf{X}^i$ e $\mathbf{Y}^i$ son de dimensión $(N-1)\times k$ y $(N-1)$, respectivamente. Aún así, observad que el número de parámetros que estimamos continúa siendo *k*; lo que ocurre es que los estimamos a partir de $N-1$ observaciones y no empleando *N* observaciones, como es habitual. Por otra parte, de acuerdo con lo que hemos visto antes, observad también que $\hat{\mathbf{Y}}^i$ e $\hat{\mathbf{Y}}$ son dos vectores columna de dimensión N. Por lo tanto, $(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^i)$ es un vector de dimensión *N*, y $(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^i)'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^i)$ es un escalar.

Existe una expresión que relaciona los conceptos que se han abordado a lo largo de las páginas anteriores (*lever*, residuo estudentizado y distancia de Cook), de modo que, si conocemos dos de ellos, siempre se puede encontrar el otro directamente. Esta expresión es la siguiente:

$$DC_i = \frac{r_i^2}{k}\frac{h_{ii}}{1 - h_{ii}}.$$

A diferencia de los *lever*, que no siguen ninguna distribución de probabilidad, la distancia de Cook se distribuye según una F de Snedecor con k grados de libertad en el numerador y $N - k$ en el denominador. Así pues, una vez que hemos calculado la distancia de Cook y conocemos los grados de libertad, diremos que la observación i-ésima tiene una influencia mayor en el ajuste del modelo de regresión que el resto si el valor obtenido por el estadístico de pruebas es mayor, o igual, que el valor de las tablas. En caso contrario, podremos considerar que la observación en cuestión no tiene más influencia en el ajuste del modelo que el resto.

Esquematizamos así la regla para realizar el contraste con la distancia de Cook:

- Si $DC_i \geq F_{k,N-k;\alpha} \Rightarrow$ la observación i-ésima tiene más influencia en el ajuste del modelo que el resto.
- Si $DC_i < F_{k,N-k;\alpha} \Rightarrow$ la observación i-ésima no tiene más influencia en el ajuste del modelo que el resto.

Idea intuitiva del contraste con la distancia de Cook

Una idea intuitiva que nos puede ayudar a entender el contraste con la distancia de Cook es la siguiente: en el supuesto de que la observación i-ésima tuviese más influencia en el ajuste del modelo de regresión que el resto, la diferencia entre el ajuste que se obtendría al considerarla y al no hacerlo, $(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{\boldsymbol{i}})$, sería "grande", y, al elevar esta diferencia al cuadrado, incluso lo sería más, por lo cual el estadístico DC_i tomaría un valor "alto". ¿Hasta qué punto "alto"? Pues más alto que el valor de la F de Snedecor de las tablas.

Ejemplo de aplicación de la distancia de Cook

A continuación, aplicaremos el concepto que hemos presentado en este subapartado a cada uno de los cuatro casos planteados anteriormente. Así, a partir de los datos aparecidos en el cuadro inicial que muestra los cuatro supuestos, hemos calculado la distancia de Cook asociada a cada una de las veintiséis observaciones de los cuatro supuestos. Los resultados que hemos obtenido los podemos encontrar en el material multimedia asociado.

Consultad los cuatro casos planteados al principio del subapartado 2.4 de este módulo didáctico.

Como sabemos, la distancia de Cook se distribuye de acuerdo con una F de Snedecor con k grados de libertad en el numerador y $N - k$ grados de libertad en el denominador, es decir, en nuestro caso con 2 y 26 – 2 = 24 grados de libertad, respectivamente. El valor de las tablas para un nivel de significación del 5% es, para una F de estas características, 3,40. Así pues, teniendo en cuenta este valor crítico y los resultados, llegamos a las conclusiones que se presentan en la tabla.

Supuesto	Observación añadida	¿DC_i elevada?
Y_2, X_2	15	No
Y_3, X_3	26	Sí
Y_4, X_4	26	Sí
Y_5, X_5	26	No

En cuanto al resto de las observaciones, podemos comprobar que ninguna de ellas (en ninguno de los cuatro supuestos planteados) tiene más influencia en el ajuste del modelo de regresión que el resto, dado que en todos los casos la distancia de Cook que tienen asociada es inferior al valor crítico 3,37.

Para acabar, únicamente hay que destacar que, tal como podemos ver en la tabla siguiente, las características de las observaciones de la muestra que hemos estudiado a lo largo de este subapartado no son excluyentes entre sí: una observación puede presentar apalancamiento, tener una influencia mayor en el ajuste del modelo de regresión que el resto y ser al mismo tiempo un *outlier*, pero también puede presentar dos de estas tres características o bien sólo una.

i	¿Apalancamiento?	¿r_i elevado?	¿DC_i elevada?
15 (caso Y_2 i X_2)	No	Sí	No
26 (caso Y_3 i X_3)	Sí	No	Sí
26 (caso Y_4 i X_4)	Sí	Sí	Sí
26 (caso Y_5 i X_5)	SI	No	No

Glosario

apalancamiento de una observación
Observación extraña respecto a las variables explicativas (eje de abscisas).

ausencia total de multicolinealidad
Caso en el que las variables explicativas están incorrelacionadas entre sí.

contraste *reset*
Contraste que sirve para detectar si la forma funcional del modelo especificado es la correcta o no lo es.

distancia de Cook
Estadístico que permite estudiar la extrañeza de una observación valorando tanto el apalancamiento que presenta como la concordancia del proceso que la ha generado con el proceso generador del resto de las observaciones.

error de especificación
Error asociado a la especificación del MRLM, aunque lo habitual es asociar este concepto a la mala especificación de la parte sistemática del modelo y, en concreto, a uno de los tres casos siguientes: omisión de variables relevantes, inclusión de variables irrelevantes o errores en la forma funcional del modelo.

factor de incremento de la varianza
Estadístico acotado inferiormente por 1 y asociado a cada $\hat{\beta}$ que indica en qué cantidad está inflada su varianza respecto al caso de ortogonalidad entre los regresores. El factor de incremento de la varianza es un indicador de multicolinealidad elevada si es mayor que 5.
sigla: FIV

FIV
Ved *factor de incremento de la varianza.*

forma funcional incorrecta
Caso en el que la relación especificada entre la variable endógena y las variables explicativas no es la correcta.

inclusión de variables irrelevantes
Caso en el que en el MRLM hay una o más variables que no son realmente relevantes para explicar la variable endógena.

influencia en el ajuste
Efecto de una observación en la estimación del modelo de regresión.

***lever*(*age*)**
Estadístico que permite medir el apalancamiento de una observación.

multicolinealidad elevada
Caso en el que la correlación entre las variables explicativas del modelo es lo suficientemente importante como para que afecte de manera relevante al proceso de estimación y contrastación.

multicolinealidad perfecta
Caso en el que dos o más variables explicativas son linealmente dependientes entre sí.

omisión de variables relevantes
Caso en el que en un MRLM no figuran una o más variables que son explicativas de la variable endógena.

outlier
Observación extraña respecto al resto de las observaciones en el sentido de que se ha generado mediante un proceso distinto del que ha generado al resto de las observaciones de la muestra.

R_X
Matriz formada por los coeficientes de correlación entre los regresores. El valor de su determinante es útil para detectar la presencia de multicolinealidad. Está acotada entre 0 y 1. Si vale 0, indica multicolinealidad perfecta y, si vale 1, indica la ausencia total de multicolinealidad.

residuo estudentizado
Estadístico que permite averiguar si una observación es un *outlier* o no lo es.

residuo estudentizado con omisión
Estadístico que permite averiguar si una observación es un *outlier* o no lo es.

Bibliografía

Gujarati, D. (1990). *Econometría* (2.ª ed.). Bogotá: McGraw-Hill.
En el momento de estudiar el apartado 1 de este módulo, es recomendable consultar el capítulo 11 de este libro. Para el estudio del apartado 2, consultad el capítulo 8.

Johnston, J. (1987). *Métodos de econometría* (trad. J. Sánchez Fernández). Barcelona: Vicens-Vives.
Consultad el capítulo 6 para los conceptos trabajados en este módulo. Además, el capítulo 10 ofrece información sobre contenidos del apartado 2, y el capítulo 12, sobre contenidos del apartado 1.

Judge, G. G.; Hill, R. C.; Griffiths, W. E.; Lütkepohl, H.; Lee, T. C. (1988). *Introduction to the Theory and Practice of Econometrics* (2.ª ed., cap. 22). Nueva York: John Willey & Sons.

Maddala, G. S. (1992). *Econometría*. México: McGraw-Hill.
Los capítulos 9 y 13 de este libro son útiles para la materia trabajada en el apartado 1, y el capítulo 10, para los contenidos del apartado 2 de este módulo.

Novales, A. (1993). *Econometría* (2.ª ed.). Madrid: McGraw-Hill.
Con relación al apartado 1 de este módulo, es interesante consultar los capítulos 3 y 11 de este libro. El capítulo 10 es adecuado como bibliografía del apartado 2 de este módulo.

Peña, D. (1989). "Modelos lineales y series temporales". *Estadística, modelos y métodos* (2.ª ed., vol. 2, cap. 9). Madrid: Alianza Editorial (Alianza Universidad Textos).

Pulido, A. (1983). *Modelos econométricos*. Madrid: Pirámide.
Podéis consultar el capítulo 9 para el estudio de los contenidos de este módulo, y el capítulo 10, para algunos contenidos del apartado 1.

Ramsey, J. B. (1969). "Test for Specification Errors in Classical Linear Squares Regression Analysis". *Journal of the Royal Statistical Society* (B, núm. 31, pág. 350-371).

Ramsey, J. B. (1970). "Models Specification Error and Inference: A Discussion of Some Problems in Econometric Methodology". *Bulletin of the Oxford Institute of Economics and Statistic* (núm. 32, pág. 301-318).

Uriel, E. y otros (1990). *Econometría. El modelo lineal*. Madrid: AC.
Podéis consultar el capítulo 9 de este libro a la hora de estudiar los contenidos del apartado 1 de este módulo didáctico. Los contenidos del apartado 2 de este módulo pueden consultarse en el capítulo 7 de este libro.