DELTA
PUBLICACIONES

Sara M. González
(Coordinadora)

Eduardo Acosta
Carmen Delia Dávila
Santiago Rodríguez
Yolanda Santana

EJERCICIOS RESUELTOS DE ECONOMETRÍA

EL MODELO DE REGRESIÓN MULTIPLE

EJERCICIOS RESUELTOS DE ECONOMETRÍA

El Modelo de Regresión Múltiple

EJERCICIOS RESUELTOS DE ECONOMETRÍA
El Modelo de Regresión Múltiple

SARA M. GONZÁLEZ BETANCOR
(Coordinadora)
EDUARDO ACOSTA GONZÁLEZ
CARMEN DELIA DÁVILA QUINTANA
SANTIAGO RODRÍGUEZ FEIJÓO
YOLANDA SANTANA JIMÉNEZ
Facultad de Ciencias Económicas y Empresariales
Departamento de Métodos Cuantitativos para Economía y Gestión
UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

DELTA PUBLICACIONES

DELTA PUBLICACIONES

EJERCICIOS RESUELTOS DE ECONOMETRÍA
EL MODELO DE REGRESIÓN MÚLTIPLE

por SARA M. GONZÁLEZ BETANCOR (Coordinadora)
EDUARDO ACOSTA GONZÁLEZ
CARMEN DELIA DÁVILA QUINTANA
SANTIAGO RODRÍGUEZ FEIJÓO
YOLANDA SANTANA JIMÉNEZ

Editor gerente Fernando M. García Tomé
Diseño de cubierta Mizar Publicidad, S.L.
Preimpresión Delta Publicaciones
Impresión Jacaryan, S.A.
Avda. Pedro Díez, 3. 28019 Madrid (España)

ISBN 84-96477-55-X
Depósito Legal M-46242-2006

(1106-60)

A nuestros alumnos

PRESENTACIÓN

El libro que tiene entre sus manos es un manual de ejercicios de Econometría resueltos paso a paso, centrado en el modelo de regresión lineal múltiple para datos de corte transversal. La tipología de ejercicios varía entre ejercicios teóricos, de demostraciones y de cálculo numérico, partiendo de datos primarios o de salidas de ordenador. Se trata, por tanto, de un manual exclusivamente de ejercicios resueltos que no incorpora una introducción teórica aparte, pues consideramos que actualmente existe una literatura amplia de manuales teóricos enfocados al nivel de la asignatura propuesta, por lo que asumimos que no era necesario incluirla.

La idea de publicar un manual como éste surge tras haber comprobado que se trata de una herramienta fundamental para el alumno que se inicia y quiere profundizar en el ámbito de la Econometría, puesto que el contenido teórico de dicha asignatura suele ser tan denso, que apenas da tiempo a resolver problemas en el aula. De esta forma el manual se convierte en la herramienta básica necesaria para los alumnos de cara a la compresión y a la práctica de la asignatura.

Hemos distribuido el contenido en cinco grandes capítulos, en los que incorporamos aspectos a menudo olvidados en otros manuales al uso, como son el estudio de las formas funcionales, las variables ficticias, el tratamiento de *outliers* y la no-normalidad de los residuos. Además, otro aspecto diferenciador del presente libro radica en la forma de presentar los datos en los ejercicios, pues estos son planteados tanto con datos respecto al origen como con datos centrados respecto a la media. De esta forma, se adquiere suficiente soltura a la hora de trabajar tanto con una tipología de datos como con la otra. Por último, hemos tratado de plantear los enunciados simplificando —en la

medida de lo posible— el cálculo numérico en la resolución, con el objetivo de evitar la pérdida de tiempo en cálculos y fomentar aspectos como la interpretación y el razonamiento del alumno.

La distribución de los ejercicios dentro de cada capítulo se ha realizado en orden creciente de dificultad, de forma que en ocasiones son necesarios los conocimientos asimilados en los primeros ejercicios como base para la resolución de los siguientes. Por su parte, la distribución de los capítulos se estructura en función del estudio clásico de los modelos de regresión lineal múltiple. De esta forma, dedicamos los dos primeros capítulos al análisis de esta tipología de modelos bajo el cumplimiento de sus hipótesis básicas y dejamos los tres últimos al estudio de la problemática en caso de incumplimiento de algunas de dichas hipótesis.

Así, el Capítulo Primero incorpora ejercicios centrados en las fases de especificación, estimación y bondad del ajuste, mientras que el Segundo se centra en las fases de contrastación y predicción, así como en la estimación por mínimos cuadrados restringidos. Por su parte, el Tercer Capítulo aborda la problemática de la multicolinealidad, los *outliers* y la no-normalidad de los residuos; el Capítulo Cuarto se ocupa de la especificación de la forma funcional, los aspectos cualitativos y el cambio estructural, así como de la selección de regresores; y, por último, el Capítulo Quinto se centra en la problemática de la existencia de perturbaciones no esféricas, proponiendo la estimación por mínimos cuadrados generalizados para el tratamiento de la heterocedasticidad.

Estamos convencidos de que la realización de los ejercicios que proponemos facilitan la comprensión de una asignatura a menudo demasiado abstracta para los alumnos, acercándolos un poco más a la aplicabilidad de la misma. Esperamos no estar equivocados en este convencimiento y que este manual les sea de utilidad.

LOS AUTORES

CONTENIDO

1

El modelo básico de regresión lineal múltiple: Especificación, estimación y bondad del ajuste

EJERCICIO 1.1

Demuestre que se cumple la relación matricial: $\hat{Y}'\hat{Y} = \hat{\beta}' X'Y$

Solución

$$\hat{Y}'\hat{Y} = \left(X\hat{\beta}\right)'\left(X\hat{\beta}\right) = \hat{\beta}' X'X\hat{\beta} = \hat{\beta}' X'X\left(X'X\right)^{-1} X'Y = \hat{\beta}' X'Y$$

EJERCICIO 1.2

En el contexto del modelo de regresión lineal múltiple (MRLM),

(a) demuestre que los residuos de la estimación por Mínimos Cuadrados Ordinarios (MCO) pueden expresarse como $e = MU$, siendo M la siguiente matriz idempotente:

$$M = \left[I - X\left(X'X\right)^{-1} X'\right],$$

(b) pruebe que la suma de cuadrados de los residuos de la estimación MCO puede escribirse como:

$$Y'Y - \hat{\beta}'X'Y$$

Solución

(a) $e = Y - X\hat{\beta} = X\beta + U - X\left[\beta + (X'X)^{-1}X'U\right] =$

$$= \left[I - X(X'X)^{-1}X'\right]U = MU$$

(b) $e'e = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) = Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} =$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X(X'X)^{-1}X'Y =$$

$$= Y'Y - \hat{\beta}'X'Y$$

EJERCICIO 1.3

Demuestre que, si denotamos por y a la matriz de datos centrados de Y, entonces $y'y$ coincide con la suma de los cuadrados totales (*SCT*) de la matriz Y.

Solución

La suma de los cuadrados totales de Y viene definida por $SCT = \sum\left(Y_i - \overline{Y}\right)^2$, por tanto, es directo comprobar que:

$$y'y = \begin{pmatrix} Y_1 - \overline{Y} & Y_2 - \overline{Y} & \dots & Y_N - \overline{Y} \end{pmatrix} \begin{pmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \dots \\ Y_N - \overline{Y} \end{pmatrix} = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

EJERCICIO 1.4

Demuestre que en un modelo de regresión lineal múltiple con ordenada en el origen, el vector de valores estimados $\hat{Y}$ está incorrelacionado con el vector de errores mínimo cuadrático ordinarios e.

Solución

Se trata de demostrar que se cumple la siguiente relación:

$$\hat{Y}'e = 0$$

Teniendo en cuenta que $\hat{Y} = X\hat{\beta}$, se obtiene que:

$$\hat{Y}'e = \hat{\beta}'X'e = \hat{\beta}'X'\left(Y - \hat{Y}\right) = \hat{\beta}'X'Y - \hat{\beta}'X'X\hat{\beta} =$$

$$= \hat{\beta}'X'Y - \hat{\beta}'X'X\left(X'X\right)^{-1}X'Y = \hat{\beta}'X'Y - \hat{\beta}'X'Y = 0$$

EJERCICIO 1.5

Demuestre que el estimador de la varianza de la perturbación aleatoria, definido como $\tilde{\hat{\sigma}}_u^2 = SCE/N$ —donde SCE es la suma de cuadrados de los errores—, es un estimador sesgado. Calcule su sesgo.

Solución

Dado el siguiente desarrollo, comprobamos que $\tilde{\hat{\sigma}}_u^2$ es un estimador sesgado, ya que su esperanza no coincide con el valor del parámetro que se pretende estimar.

$$E\left(\tilde{\hat{\sigma}}_u^2\right) = E\left(\frac{SCE}{N}\right) = \frac{E\left(SCE\right)}{N} = \frac{\sigma_u^2\left(N-k\right)}{N} \neq \sigma_u^2$$

El sesgo de un estimador se define como la diferencia entre el valor del parámetro que estima y la media del estimador. Realizando los cálculos oportunos concluimos que el sesgo es igual a $\sigma_u^2 k/N$. Como se puede observar en esta última expresión, al incrementar el tamaño muestral el sesgo se reduce. Por ello decimos que $\tilde{\hat{\sigma}}_u^2 = SCE/N$ es un estimador asintóticamente insesgado de la varianza de la perturbación aleatoria.

EJERCICIO 1.6

Demuestre que el coeficiente de determinación con datos centrados se puede calcular como:

$$R^2 = \frac{\hat{\beta}'x'y}{y'y}$$

Solución

El coeficiente de determinación del modelo con datos centrados y no centrados es el mismo.

Por definición, los errores centrados se definen como $e = y - \hat{y}$. Por tanto, sabiendo que con datos centrados se cumple que $SCT = y'y$, es inmediato demostrar:

$$e'e = y'y - \hat{\beta}'x'y = SCT - \hat{\beta}'x'y$$

Podemos expresar R^2 de la siguiente manera:

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{SCT - \hat{\beta}x'y}{SCT} = \frac{\hat{\beta}x'y}{y'y}$$

Esta misma conclusión se obtiene si razonamos de la siguiente manera:

Teniendo en cuenta que

$$R^2 = \frac{\hat{\beta}'X'Y - N\overline{Y}^2}{Y'Y - N\overline{Y}^2},$$

cuando las variables están centradas, dado que la media vale cero, se tendrá que

$$R^2 = \frac{\hat{\beta}'x'y}{y'y}.$$

EJERCICIO 1.7

Calcule la matriz de covarianzas entre el vector de estimadores de los coeficientes y la perturbación aleatoria del modelo de RLM.

Solución

El ejercicio nos pide calcular la expresión

$$V\left(\hat{\beta},U\right) = E\left[\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)\left(U - E(U)\right)'\right].$$

Esta expresión la podemos escribir de la siguiente manera:

$$E\left[\left(\hat{\beta} - \beta\right)U'\right] = E\left[\left(X'X\right)^{-1}X'UU'\right] = \left(X'X\right)^{-1}X'\sigma_u^2 I = \sigma_u^2\left(X'X\right)^{-1}X'$$

EJERCICIO 1.8

Demuestre que la matriz de covarianzas entre los coeficientes estimados y los errores mínimo cuadrático ordinarios del modelo de RLM es una matriz de ceros.

Solución

Tenemos que demostrar que

$$V\left(\hat{\beta},e\right)=E\left[\left(\hat{\beta}-E\left(\hat{\beta}\right)\right)\left(e-E\left(e\right)\right)'\right]=E\left[\left(\hat{\beta}-\beta\right)e'\right]=0\,.$$

Para ello únicamente tenemos que recordar que $e=MU$. Por tanto, podemos calcular $V\left(\hat{\beta},e\right)$ como se indica a continuación:

$$\begin{aligned}V\left(\hat{\beta},e\right)&=E\left[\left(\hat{\beta}-\beta\right)e'\right]=E\left[\left(X'X\right)^{-1}X'UU'M'\right]=\left(X'X\right)^{-1}X'E\left[UU'\right]M'=\\&=\sigma_u^2\left(X'X\right)^{-1}X'M=\sigma_u^2\left(X'X\right)^{-1}X'\left[I-X\left(X'X\right)^{-1}X'\right]=\\&=\sigma_u^2\left(X'X\right)^{-1}X'-\sigma_u^2\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X'=\\&=\sigma_u^2\left(X'X\right)^{-1}X'-\sigma_u^2\left(X'X\right)^{-1}X'=0\end{aligned}$$

EJERCICIO 1.9

Demuestre analíticamente que, en el modelo de regresión lineal simple $Y_i=\beta_1+\beta_2X_i+u_i$, el cuadrado del coeficiente de correlación lineal coincide con el coeficiente de determinación.

Solución

El coeficiente de correlación lineal simple para valores centrados es:

$$r_{x,y}=\frac{S_{x,y}}{S_xS_y}=\frac{x'y}{\sqrt{x'x}\sqrt{y'y}}$$

Por tanto, su cuadrado será:

$$r_{x,y}^2=\underbrace{\frac{\left(x'y\right)}{\left(x'x\right)}}_{\hat{\beta}_2}\cdot\frac{\left(x'y\right)}{\left(y'y\right)}=\frac{\hat{\beta}_2x'y}{y'y}=R^2$$

EJERCICIO 1.10

La variabilidad de los beneficios de las empresas se quiere explicar en función de la productividad de sus empleados y de su gasto en promoción. Para ello se toma una muestra de 100 empresas y se obtienen los sumatorios para cada una de las variables y para todos sus cruces dos a dos. En la primera parte de la Tabla 1.1 se muestran los sumatorios para cada variable y en la segunda parte el sumatorio de sus cruces.

Tabla 1.1

		Beneficios (€)	Productividad (€ de producción por trabajador)	Promoción (€)
$\sum_{i=1}^{100} z_{ji}$		4881	263	991
		Beneficios	**Productividad**	**Promoción**
$\sum_{i=1}^{100} z_{ji}z_{si}$	Beneficios	311894	13512	47766
	Productividad		930	2662
	Promoción			13771

A partir de estos datos, obtenga las matrices $X'X$ y $X'Y$.

Solución

Dado que X es una matriz formada, por columnas, por los datos de cada una de las variables del modelo, es inmediato demostrar que

$$X'X = \begin{pmatrix} N & \sum_{i=1}^{N} X_{2i} & \sum_{i=1}^{N} X_{3i} \\ \sum_{i=1}^{N} X_{2i} & \sum_{i=1}^{N} X_{2i}^2 & \sum_{i=1}^{N} X_{2i}X_{3i} \\ \sum_{i=1}^{N} X_{3i} & \sum_{i=1}^{N} X_{2i}X_{3i} & \sum_{i=1}^{N} X_{3i}^2 \end{pmatrix}; \; X'Y = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} Y_i X_{2i} \\ \sum_{i=1}^{N} Y_i X_{3i} \end{pmatrix}$$

Por tanto, también es inmediato obtener las siguientes matrices a partir de la información que nos da la tabla del enunciado:

$$X'X = \begin{pmatrix} 100 & 263 & 991 \\ 263 & 930 & 2662 \\ 991 & 2662 & 13771 \end{pmatrix}; \; X'Y = \begin{pmatrix} 4881 \\ 13512 \\ 47766 \end{pmatrix}$$

EJERCICIO 1.11

Sea el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{1.1}$$

y los siguientes datos muestrales:

Tabla 1.2

Y	X_2	X_3
1	7	10
3	3	3
1	6	8
2	3	1
3	8	6
4	6	3
1	9	13

Sabiendo que para la estimación por Mínimos Cuadrados Ordinarios (MCO) del modelo (1.1) $e'e = 2.195756$, conteste a las siguientes preguntas:

(a) Estime el modelo (1.1) por MCO.

(b) Estime la varianza de las perturbaciones.

(c) Calcule el coeficiente de determinación y el coeficiente de determinación corregido.

(d) Estime la matriz de varianzas y covarianzas de los coeficientes estimados.

(e) Indique qué propiedades debe cumplir la perturbación aleatoria del modelo (1.1) para que no viole ninguna de las hipótesis básicas del modelo de regresión lineal múltiple.

Solución

(a) A partir de los datos de la Tabla 1.2 construimos las matrices correspondientes y calculamos el vector de estimadores $\hat{\beta}$:

$$X'X = \begin{pmatrix} 7 & 42 & 44 \\ & 284 & 313 \\ & & 388 \end{pmatrix} \quad X'Y = \begin{pmatrix} 15 \\ 85 \\ 72 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1.4992 & -0.3096 & 0.0797 \\ & 0.0957 & -0.0420 \\ & & 0.0275 \end{pmatrix}$$

Aplicando la expresión matricial para el cálculo de $\hat{\beta}$ obtenemos el siguiente resultado:

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 1.9140 \\ 0.4592 \\ -0.4019 \end{pmatrix}$$

(b) La varianza estimada de u_i es:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k}$$

Para su cálculo, necesitamos obtener la suma de cuadrados de los errores (*SCE*). Una expresión matricial que permite su cálculo es la siguiente:

$$e'e = Y'Y - \hat{\beta}'X'Y = 41 - 38.8042 = 2.1958$$

Por tanto, la varianza estimada de la perturbación es:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{2.1958}{7-3} = 0.5489$$

(c) El coeficiente de determinación representa la proporción de la varianza de Y explicada por la regresión. Puede expresarse como:

$$R^2 = 1 - \frac{e'e}{Y'Y - N\bar{Y}^2} = 1 - \frac{2.1958}{41 - 7 \cdot \left(\frac{15}{7}\right)^2} = 0.7521$$

En este caso, un 75% de la varianza de Y queda explicada por la regresión.

El coeficiente de determinación ajustado corrige el coeficiente de determinación por los grados de libertad de la *SCE* y de la suma cuadrática de la regresión (*SCR*), permitiendo comparar la bondad del ajuste entre modelos con idéntica variable endógena. Su cálculo es el siguiente:

$$\bar{R}^2 = 1 - \frac{e'e}{Y'Y - N\bar{Y}^2} \cdot \frac{N-1}{N-k} = 1 - \frac{2.1958}{41 - 7 \cdot \left(\frac{15}{7}\right)^2} \cdot \frac{7-1}{7-3} = 0.6281$$

(d) La matriz de la estimación de las varianzas y covarianzas de los estimadores se obtiene a partir de la siguiente expresión:

$$\hat{V}\left(\hat{\beta}\right) = \hat{\sigma}_u^2 \left(X'X\right)^{-1} = 0.5489 \begin{pmatrix} 1.4992 & -0.3096 & 0.0797 \\ & 0.0957 & -0.0420 \\ & & 0.0275 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.8230 & -0.1699 & 0.0437 \\ & 0.0525 & -0.0231 \\ & & 0.0151 \end{pmatrix}$$

(e) La perturbación aleatoria debe cumplir las siguientes propiedades para que no viole ninguna de las hipótesis básicas:

- La covarianza entre la perturbación aleatoria y las variables explicativas tiene que ser cero.
- La media de la perturbación aleatoria tiene que ser igual a cero.
- La varianza de la perturbación aleatoria tiene que ser constante (homocedasticidad).
- La covarianza entre las diferentes perturbaciones aleatorias tiene que ser cero.
- La perturbación aleatoria tiene que distribuirse como una normal.

EJERCICIO 1.12

Sea el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{1.2}$$

y los siguientes datos muestrales:

Tabla 1.3

Y	X_2	X_3
0.6	0.8	0.4
0.8	0.6	0.1
0.6	0.4	−0.1
0.5	0.4	0.3
0.6	0.5	0.0
0.9	0.5	−0.2
−0.4	0.0	0.5

Sabiendo que para la estimación por Mínimos Cuadrados Ordinarios (MCO) del modelo (1.2) $e'e = 0.068135$,

(a) Estime el modelo (1.2) por MCO.

(b) Estime la varianza de las perturbaciones.

(c) Calcule el coeficiente de determinación y el coeficiente de determinación corregido.

(d) Estime la matriz de varianzas y covarianzas de los coeficientes estimados.

(e) Demuestre que los estimadores MCO son insesgados cuando se cumplen las hipótesis básicas del modelo.

Solución

(a) A partir de los datos de la Tabla 1.3 construimos las matrices correspondientes y calculamos el vector de estimadores $\hat{\beta}$:

$$X'X = \begin{pmatrix} 7 & 3.20 & 1.00 \\ & 1.82 & 0.36 \\ & & 0.56 \end{pmatrix} \quad X'Y = \begin{pmatrix} 3.60 \\ 2.15 \\ 0.03 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 0.910729 & -1.466010 & -0.683870 \\ & 2.989353 & 0.696151 \\ & & 2.559378 \end{pmatrix}$$

Aplicando la expresión matricial para el cálculo de $\hat{\beta}$ obtenemos el siguiente resultado:

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 0.10618 \\ 1.17035 \\ -0.88840 \end{pmatrix}$$

(b) Calculamos previamente la *SCE*:

$$e'e = Y'Y - \hat{\beta}X'Y = 2.94 - 2.871865 = 0.068135$$

Sustituyendo en la expresión de la varianza estimada de la perturbación, obtenemos

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{0.068135}{7-3} = 0.01703$$

(c) El coeficiente de determinación es:

$$R^2 = 1 - \frac{e'e}{Y'Y - N\bar{Y}^2} = 1 - \frac{0.068135}{2.94 - 7 \cdot \left(\frac{3.6}{7}\right)^2} = 0.93741$$

El coeficiente de determinación ajustado es:

$$\bar{R}^2 = 1 - \frac{e'e}{Y'Y - N\bar{Y}^2} \cdot \frac{N-1}{N-k} = 1 - \frac{0.068135}{2.94 - 7 \cdot \left(\frac{3.6}{7}\right)^2} \cdot \frac{7-1}{7-3} = 0.90611$$

(d) La matriz de varianzas y covarianzas de $\hat{\beta}$ es:

$$\hat{V}\left(\hat{\beta}\right) = \hat{\sigma}_u^2 \left(X'X\right)^{-1} = 0.01703 \begin{pmatrix} 0.910729 & -1.466010 & -0.683870 \\ & 2.989353 & 0.696151 \\ & & 2.559378 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.01551 & -0.02497 & -0.01165 \\ & 0.05092 & 0.01186 \\ & & 0.04360 \end{pmatrix}$$

(e) Un estimador es insesgado cuando su valor esperado coincide con el parámetro poblacional. Calculando el valor esperado de $\hat{\beta}$, obtenemos:

$$E\left(\hat{\beta}\right) = E\left[\left(X'X\right)^{-1} X'Y\right] = E\left[\left(X'X\right)^{-1} X'\left(X\beta + U\right)\right] =$$

$$= E\left[\beta + \left(X'X\right)^{-1} X'U\right] = \beta + \left(X'X\right)^{-1} X'E\left(U\right) = \beta$$

Por tanto, $\hat{\beta}$ es un estimador insesgado de β.

EJERCICIO 1.13

Se quiere estimar el modelo de regresión lineal múltiple $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ con $N = 11$. La información muestral disponible para valores centrados es:

$$x'x = \begin{pmatrix} 607.97 & 117.66 \\ 117.66 & 28.55 \end{pmatrix} \quad x'y = \begin{pmatrix} 325.08 \\ 62.22 \end{pmatrix} \quad y'y = 288.92$$

$$\overline{X}_1 = 9.42 \quad \overline{X}_2 = 3.36 \quad \overline{Y} = 4.09$$

(a) Estime el modelo y escríbalo como $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$.

(b) Obtenga la matriz de varianzas y covarianzas de los estimadores MCO de β_2 y β_3.

Solución

(a) Los valores estimados de los coeficientes β_2 y β_3 vienen dados por

$$\hat{\beta} = (x'x)^{-1} x'y = \begin{pmatrix} 0.56 \\ -0.12 \end{pmatrix}$$

La estimación del término independiente la realizamos a partir de la expresión $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 = 4.09 - 0.56 \cdot 9.42 + 0.12 \cdot 3.36 = -0.76$, por lo que la estimación solicitada vendrá dada por:

$$\hat{Y}_i = -0.76 + 0.56 X_{2i} - 0.12 X_{3i}$$

(b) La matriz de varianzas y covarianzas de los estimadores de los parámetros es:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_u^2 (x'x)^{-1} = \begin{pmatrix} S^2(\hat{\beta}_2) & \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_3) \\ \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_3) & S^2(\hat{\beta}_3) \end{pmatrix} \tag{1.3}$$

Previamente calculamos el estimador de σ_u^2 a partir de:

$$\hat{\sigma}_u^2 = \frac{y'y - \hat{\beta}'x'y}{N-k} =$$

$$= \frac{288.92 - \begin{pmatrix} 0.56 & -0.12 \end{pmatrix} \begin{pmatrix} 325.08 \\ 62.22 \end{pmatrix}}{11-3} = \frac{288.92 - 173.89}{11-3} = 14.38$$

de donde, sin más que sustituir en (1.3) obtenemos la estimación de la matriz de varianzas y covarianzas de los estimadores MCO:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_u^2 (x'x)^{-1} = 14.38 \begin{pmatrix} 0.01 & -0.03 \\ -0.03 & 0.17 \end{pmatrix} = \begin{pmatrix} 0.12 & -0.48 \\ -0.48 & 2.49 \end{pmatrix}$$

EJERCICIO 1.14

Dados los datos de la Tabla 1.4, siendo el modelo a estimar:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

donde la variable endógena es el precio de los coches (en miles de €), la primera exógena (X_2) mide la potencia en caballos del coche y la segunda exógena (X_3) mide la longitud en metros del coche.

Tabla 1.4

Y	X_2	X_3
15	90	3
20	100	3.2
15	80	3.2
15	70	3
20	100	4

Calcule:

(a) Los estimadores de posición puntuales MCO y su interpretación, sabiendo que

$$(XX)^{-1} = \begin{pmatrix} 17.480 & -0.06100 & -3.6300 \\ -0.061 & 0.00226 & -0.0421 \\ -3.630 & -0.04210 & 2.2370 \end{pmatrix}$$

(b) La dispersión de los estimadores de posición. Comente sus valores.

(c) La bondad del ajuste mediante el coeficiente de determinación.

Solución

(a) El modelo propuesto en forma matricial se puede escribir como $Y = X\beta + U$, siendo:

$$X = \begin{pmatrix} 1 & 90 & 3.0 \\ 1 & 100 & 3.2 \\ 1 & 80 & 3.2 \\ 1 & 70 & 3.0 \\ 1 & 100 & 4.0 \end{pmatrix} \quad Y = \begin{pmatrix} 15 \\ 20 \\ 15 \\ 15 \\ 20 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}$$

El estimador mínimo cuadrático ordinario del vector β se obtiene como $(X'X)^{-1}X'Y$. Donde $X'Y$ es la matriz columna formada por los siguientes elementos:

$$X'Y = \begin{pmatrix} 85 \\ 7600 \\ 282 \end{pmatrix}$$

Por tanto, el vector de parámetros estimados es igual a

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} -1.95 \\ 0.14 \\ 2.11 \end{pmatrix}$$

En consecuencia, el precio de un coche con potencia cero y longitud cero sería de −1950 euros. Como se puede deducir, dado que no existen coches con potencia y longitud cero, el estimador de la ordenada en el origen no tiene interpretación económica. La estimación del parámetro de la variable potencia indica que por cada caballo adicional de potencia, por término medio se incrementa el precio del coche en 140 euros. En lo que se refiere a la variable longitud, por cada metro de más que tenga el vehículo su precio, en términos medios, se incrementa en 2110 euros.

(b) El estimador de la dispersión de los estimadores se obtiene mediante la expresión

$$\hat{V}(\hat{\beta}) = \frac{e'e}{N-k}(X'X)^{-1}$$

Para su cálculo se necesita obtener la suma de los cuadrados de los errores. Existen distintas alternativas. Una de ellas es la que se indica a continuación:

$$e'e = Y'Y - \hat{\beta}'X'Y = \sum_{i=1}^{5} Y_i^2 - \begin{pmatrix} -1.95 & 0.14 & 2.11 \end{pmatrix} \begin{pmatrix} 85 \\ 7600 \\ 282 \end{pmatrix} = 6.84$$

Con este resultado es inmediato obtener la matriz estimada de varianzas y covarianzas de los coeficientes estimados.

$$\hat{V}(\hat{\beta}) = \frac{6.84}{5-3} \begin{pmatrix} 17.480 & -0.06100 & -3.6300 \\ -0.061 & 0.00226 & -0.0421 \\ -3.630 & -0.04210 & 2.2370 \end{pmatrix}$$

En consecuencia, las desviaciones típicas estimadas de los estimadores coinciden con la raíz cuadrada de los elementos de la diagonal principal de la matriz resultante. Esto es,

$$S\left(\hat{\beta}_1\right) = 7.73 \quad S\left(\hat{\beta}_2\right) = 0.088 \quad S\left(\hat{\beta}_3\right) = 2.77$$

Comparando cada uno de estos resultados con el parámetro estimado, es decir, con

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} -1.95 \\ 0.14 \\ 2.11 \end{pmatrix}$$

se puede concluir que especialmente el estimador de β_1 es poco preciso.

(c) Para calcular la bondad del ajuste obtenemos el coeficiente de determinación, cuya expresión es la siguiente:

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

La *SCE* ya la hemos calculado, y vale 6.84, y la *SCT* la calculamos como

$$SCT = \sum_{i=1}^{5} Y_i^2 - \frac{\left(\sum_{i=1}^{5} Y_i\right)^2}{5} = 1475 - \frac{(85)^2}{5} = 30$$

Por tanto,

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{6.84}{30} = 0.772$$

Es decir, el 77.2% de la variación del precio de los coches está explicado de forma conjunta por la longitud del vehículo y su potencia.

EJERCICIO 1.15

¿Qué mide el coeficiente de determinación corregido en grados de libertad? Calcúlelo para los datos del ejercicio 1.14 e interprételo.

Solución

El coeficiente de determinación corregido mide la bondad del ajuste, pero teniendo en cuenta la relación entre la pérdida de grados de libertad que supone

incrementar una nueva variable en el modelo y lo que ella aporta para la explicación de la variable endógena. Éste oscila entre $-\infty$ y 1.

Su cálculo se muestra a continuación:

$$\overline{R}^2 = 1 - \frac{\frac{SCE}{N-k}}{\frac{SCT}{N-1}} = 1 - \frac{SCE \cdot (N-1)}{SCT \cdot (N-k)} = \frac{6.84 \cdot 4}{30 \cdot 2} = 0.456$$

La gran diferencia que existe entre ambos coeficientes de determinación muestra que en el modelo hay variables que aportan poco o nada a la explicación del precio de los coches. En cualquier caso, en este ejemplo, al trabajar con tan pocos datos, los grados de libertad con los que se trabaja son muy bajos.

EJERCICIO 1.16

Las siguientes salidas de regresión recogen los resultados de dos especificaciones distintas para estimar la posición en el ranking de los jugadores de la liga ACB en la temporada 2002/03.

Cuadro 1.1

Dependent Variable: *RANKING*
Method: Least Squares
Sample: 1 195
Included observations: 195

Variable	Coefficient	Std. Error	t-Statistic	Prob.
%_TIROS_DE_2	0.980772	0.260425	3.766038	0.0002
%_TIROS_DE_3	3.888446	0.356306	10.913220	0.0000
%_TIROS_LIBRES	0.641341	0.278114	2.306035	0.0222
REBOTES_DEFENSIVOS	1.067425	0.277576	3.845530	0.0002
FALTAS_COMETIDAS	2.478362	0.325485	7.614369	0.0000
BALONES_RECUPERADOS	2.329961	0.516692	4.509377	0.0000
BALONES_PERDIDOS	2.092656	0.441097	4.744204	0.0000
C	11.143260	3.748271	2.972908	0.0033

R-squared		Mean dependent var	141.91790
Adjusted R-squared		S.D. dependent var	79.32274
S.E. of regression	23.015440	Akaike info criterion	
Sum squared resid	99055.870000	Schwarz criterion	
Log likelihood	–884.160900	F-statistic	
Durbin-Watson stat	1.969352	Prob(F-statistic)	

Cuadro 1.2

Dependent Variable: *RANKING*
Method: Least Squares
Sample: 1 195
Included observations: 195

Variable	Coefficient	Std. Error	t-Statistic	Prob.
%_TIROS_DE_2	1.548346	0.313809	4.934036	0.0000
%_TIROS_DE_3	4.617886	0.424302	10.883490	0.0000
%_TIROS_LIBRES	0.872375	0.334474	2.608199	0.0098
REBOTES_DEFENSIVOS	2.028618	0.321359	6.312631	0.0000
BALONES_RECUPERADOS	4.276269	0.590222	7.245181	0.0000
C	31.47325	3.757417	8.376301	0.0000

R-squared		Mean dependent var	141.91790
Adjusted R-squared		S.D. dependent var	79.32274
S.E. of regression	28.433600	Akaike info criterion	
Sum squared resid	152800.700000	Schwarz criterion	
Log likelihood	–926.422300	F-statistic	
Durbin-Watson stat	1.960884	Prob(F-statistic)	

(a) Calcule los 2 valores ($R^2; \overline{R}^2$) que se han borrado en cada una de las estimaciones.

(b) ¿Qué puede decir de la bondad del ajuste de estas estimaciones?

Solución

Para realizar el cálculo de los valores borrados, tanto en el Cuadro 1.1 como en el Cuadro 1.2, se procederá de la siguiente forma:

(a) Con respecto a los datos del Cuadro 1.1, es necesario realizar los siguientes cálculos:

- Coeficiente de determinación:

La *SCT* la obtenemos despejándola a partir de la cuasidesviación típica de Y:

$$79.32274 = \sqrt{\frac{SCT}{N-1}} \Rightarrow$$

$$\Rightarrow \text{SCT} = 79.32274^2 \cdot (N-1) = 79.32274^2 \cdot (195-1) = 1220666.834$$

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{e'e}{y'y} = 1 - \frac{99055.87}{1220666.834} = 0.91885$$

- Coeficiente de determinación ajustado:

$$\overline{R}^2 = 1-(1-0.91885)\frac{N-1}{N-k} = 1-(1-0.91885)\frac{195-1}{195-8} = 0.9158$$

Con respecto a los datos del Cuadro 1.2, es necesario realizar los siguientes cálculos:

- Coeficiente de determinación:

Para el cálculo del coeficiente de determinación no es necesario volver a obtener el valor de la *SCT*, puesto que la endógena de ambos modelos es la misma y, por tanto, la *SCT* también lo es.

$$R^2 = 1-\frac{SCE}{SCT} = 1-\frac{e'e}{y'y} = 1-\frac{152800.7}{1220666.834} = 0.8748$$

- Coeficiente de determinación ajustado:

$$\overline{R}^2 = 1-(1-0.8748)\frac{N-1}{N-k} = 1-(1-0.8748)\frac{195-1}{195-6} = 0.8715$$

(b) Aunque ambos modelos presentan un coeficiente de determinación bastante alto, de forma que podemos hablar de una buena bondad del ajuste, el modelo que recoge el Cuadro 1.1 presenta una mejor bondad del ajuste que el del Cuadro 1.2, puesto que su coeficiente de determinación ajustado es más alto.

EJERCICIO 1.17

De un modelo de regresión sabemos que la matriz de datos y el vector de errores MCO son lOs siguientes:

$$X = \begin{pmatrix} 1 & 4 & \square \\ 1 & -2 & 1 \\ 1 & 1 & 4 \\ 1 & -2 & 8 \\ 1 & \square & 2 \end{pmatrix} \quad e = \begin{pmatrix} \square \\ 4 \\ 5 \\ -7 \\ -1 \end{pmatrix}$$

Complete los valores que faltan en dichas matrices.

Solución

La solución es inmediata si recordamos que las variables explicativas están incorrelacionadas con los errores mínimo cuadrático ordinarios. Es decir, si recordamos que se cumple que $X'e = 0$. La solución que se obtiene es la siguiente:

$$X = \begin{pmatrix} 1 & 4 & -34 \\ 1 & -2 & 1 \\ 1 & 1 & 4 \\ 1 & -2 & 8 \\ 1 & 7 & 2 \end{pmatrix} \quad e = \begin{pmatrix} -1 \\ 4 \\ 5 \\ -7 \\ -1 \end{pmatrix}$$

EJERCICIO 1.18

Si en un modelo que se estima con 100 datos y tiene 4 regresores, incluyendo la constante, se obtiene un R^2 corregido igual a 0.7, determine qué porcentaje de variación de la variable endógena queda explicado por la regresión.

Solución

El estadístico que mide el porcentaje de variabilidad de la endógena que viene explicado por la regresión es el coeficiente de determinación. La fórmula del coeficiente de determinación corregido en función del coeficiente de determinación es la siguiente:

$$\overline{R}^2 = 1 - \left(1 - R^2\right)\frac{N-1}{N-k}$$

Partiendo de esta expresión, es inmediato demostrar que se cumple lo siguiente:

$$R^2 = 1 + \frac{\left(\overline{R}^2 - 1\right)\left(N-k\right)}{\left(N-1\right)} = 1 + \frac{\left(0.7-1\right)\left(100-4\right)}{\left(100-1\right)} = 0.71$$

Por tanto, la regresión explica el 71% de las variaciones de la variable endógena.

EJERCICIO 1.19

En un modelo de regresión lineal se dispone de tres valores de Y para realizar su estimación. Estos valores son 2, 4 y 8. Después de estimar el modelo se obtiene que

$$\sum_{i=1}^{3} \hat{Y}_i^2 = 80$$

¿Cuánto vale la suma de los cuadrados de los errores? Calcule e interprete el coeficiente de determinación.

Solución

El cálculo de la *SCE* es inmediato si recordamos que

$$SCE = e'e = Y'Y - \hat{Y}'\hat{Y} = \sum_{i=1}^{3} Y_i^2 - \sum_{i=1}^{3} \hat{Y}_i^2$$

Por tanto, $SCE = (4+16+64) - 80 = 4$.

El coeficiente de determinación se puede calcular como

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{SCE}{Y'Y - N\overline{Y}^2} = 1 - \frac{4}{84 - 3 \cdot 4.67^2} = 0.7857$$

Es decir, aproximadamente el 78% de la variabildidad de la variable endógena viene explicado por la regresión.

EJERCICIO 1.20

Sea y la variable endógena centrada de un modelo con tres coeficientes (incluida la constante), en el cual se ha obtenido un vector de errores mínimo cuadrático ordinarios igual a

$$e = \begin{pmatrix} 2 \\ -4 \\ 2 \\ 5 \\ 3 \\ -8 \end{pmatrix}$$

Sabiendo que $\sum_{i=1}^{N} (\hat{y}_i)^2 = 100$, calcule e interprete el coeficiente de determinación y el coeficiente de determinación corregido.

Solución

Dado que $\sum_{i=1}^{N} \left(\hat{y}_i\right)^2 = \sum_{i=1}^{N} \left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 = SCR$, sabemos que $SCR = 100$. Por otra parte, puesto que tenemos el vector de errores, al sumar sus cuadrados obtenemos un valor para la $SCE = 122$. Teniendo en cuenta que $SCT = SCR + SCE$, obtenemos que $SCT = 222$. Con esta información es inmediato calcular el coeficiente de determinación.

$$R^2 = \frac{SCR}{SCT} = \frac{100}{222} = 0.45$$

Es decir, el 45% de los cambios de la variable que se desea explicar es explicado por la variabilidad de los regresores. El resto, un 55%, viene explicado por las variables que forman los errores.

Por último, teniendo en cuenta que $k = 3$, podemos calcular el coeficiente de determinación corregido como

$$\overline{R}^2 = 1 - \frac{\frac{SCE}{N-k}}{\frac{SCT}{N-1}} = 1 - \frac{SCE}{SCT} \cdot \frac{N-1}{N-k} = 1 - \frac{122}{222} \cdot \frac{6-1}{6-3} = 0.084$$

Este valor no se puede interpretar como un porcentaje, sólo podemos decir que su valor se aleja mucho de uno, con lo cual el modelo no tiene una buena bondad de ajuste. Además, la diferencia que existe entre el coeficiente de determinación y el coeficiente de determinación corregido nos está indicando que hay variables explicativas que explican muy poco y no compensan la pérdida de grados de libertad que produce su inclusión en el modelo. No obstante, en este caso, la diferencia también está afectada por el reducido tamaño muestral con el que se trabaja.

EJERCICIO 1.21

Siendo C la variable consumo e Y la variable renta, estamos interesados en estimar qué parte de esta última se destina al consumo mediante un modelo de regresión lineal simple. Para ello se dispone de los datos que se muestran en la Tabla 1.5, todos ellos medidos en euros mensuales y correspondientes a 20 individuos extraídos de forma aleatoria de la población objeto de estudio.

Tabla 1.5

C	Y
327	470
169	269
24	89
283	415
493	676
37	101
511	694
586	784

(continúa en la página siguiente)

Tabla 1.5 (continuación)

C	Y
356	505
570	771
18	81
258	384
472	651
360	512
239	351
521	710
219	327
296	423
532	721
34	104

(a) Estime por MCO la propensión marginal al consumo.

(b) Estime la precisión del estimador obtenido en el apartado anterior.

Solución

(a) Si especificamos el modelo como

$$C_i = \beta_1 + \beta_2 Y_i + u_i$$

siendo C el consumo e Y la renta, β_2 (propensión marginal al consumo) mide en cuánto se incrementa el consumo por cada unidad de incremento de la renta. Así, si β_2 fuese 0.5, eso significaría que, en promedio, el 50% de los incrementos en renta estarían destinados al consumo. Por tanto, el parámetro β_2 mide la parte de la renta que, en promedio, un individuo dedica al consumo cuando se incrementa su renta. En consecuencia, lo único que tenemos que hacer para contestar a la pregunta es estimar el valor del parámetro β_2.

Con los datos de la tabla del enunciado obtenemos las siguientes matrices:

$$X'X = \begin{pmatrix} 20 & 9038 \\ 9038 & 5170120 \end{pmatrix} \quad X'Y = \begin{pmatrix} 6305 \\ 3718895 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 0.238000 & -0.00041600 \\ -0.000416 & 0.00000092 \end{pmatrix}$$

A partir de estas matrices, y utilizando la expresión, $\hat{\beta} = (X'X)^{-1} X'Y$ es inmediato obtener que el estimador de β_2 es igual a 0.8. Por tanto, en promedio, el 80% de los incrementos de renta se destina al consumo.

(b) La precisión de un estimador es inversamente proporcional a la desviación típica del mismo. La expresión a utilizar para estimar la matriz de varianzas y covarianzas es

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_u^2 (X'X)^{-1}$$

siendo la raíz cuadrada del elemento (2,2) de la matriz resultante la desviación típica de $\hat{\beta}_2$. Para calcular $\hat{\sigma}_u^2$ tenemos en cuenta que

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{(Y-\hat{Y})'e}{20-2} = \frac{Y'e - \hat{\beta}'X'e}{18} = \frac{Y'(Y-\hat{Y})}{18} =$$

$$= \frac{Y'Y - Y'X\hat{\beta}}{18} = \frac{2684317 - 2684174.36}{18} = 7.9$$

Con este último dato y el elemento (2,2) de la matriz $(X'X)^{-1}$, obtenemos que la varianza del estimador de β_2 se calcula como se indica a continuación:

$$S^2(\hat{\beta}_2) = 7.9 \cdot 0.00000092 = 0.0000073$$

La raíz cuadrada de este valor es la desviación típica del estimador, que es igual a 0.0027.

EJERCICIO 1.22

Se ha estimado el siguiente modelo de regresión lineal múltiple:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

obteniéndose los siguientes estimadores:

$$\hat{\beta}_2 = 4 \quad \hat{\beta}_3 = 1 \quad \hat{\beta}_4 = 2$$

Si las medias de las variables Y, X_2, X_3 y X_4 son 10, 15, 20 y 100 respectivamente, ¿cuál es el valor estimado de Y cuando el resto de variables toman el valor cero?

Solución

Por término medio, cuando todas las variables explicativas toman el valor cero, el valor esperado de Y es –270 unidades. Ello se debe a que el modelo estimado es

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i}$$

Por tanto, $\hat{Y}$ para el caso en el cual todas las variables explicativas tomen el valor cero, se obtiene directamente sustituyendo las X_i por el valor cero, con lo cual se concluye que el valor pedido es igual al parámetro estimado de la constante. Es decir,

$$\hat{Y}\Big|_{X_2=0;\, X_3=0;\, X_4=0} = \hat{\beta}_1$$

Por otra parte, el parámetro de la constante lo podemos estimar por MCO como

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 - \hat{\beta}_4 \overline{X}_4$$

Sustituyendo en esta última expresión los valores de las medias de las variables y los coeficientes estimados de las variables X_j, se obtiene directamente el valor estimado de β_1. Ese valor es –270 que, dependiendo del modelo considerado, tendrá o no interpretación económica.

EJERCICIO 1.23

Dado el siguiente modelo de la función de consumo de un producto expresado en desviaciones respecto a la media:

$$y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

donde: x_2 = renta disponible; x_3 = precio relativo del bien respecto al índice general de precios al consumo, disponemos de los siguientes productos cruzados:

	x_2	x_3	y
x_2	100 000	9 000	8 000
x_3		850	700
y			670

Sabiendo además que $\overline{Y} = 54$ $\overline{X}_2 = 300$ $\overline{X}_3 = 30$ y $N = 5$,

(a) Estime los coeficientes del modelo.

(b) Obtenga la varianza de los coeficientes estimados.

(c) Obtenga el coeficiente de determinación.

Solución

(a) Construimos las matrices con datos centrados:

$$x'x = \begin{pmatrix} 100\,000 & 9\,000 \\ 9\,000 & 850 \end{pmatrix} \quad x'y = \begin{pmatrix} 8\,000 \\ 700 \end{pmatrix}$$

$$(x'x)^{-1} = \begin{pmatrix} 2.125 \cdot 10^{-4} & -2.23 \cdot 10^{-3} \\ -2.230 \cdot 10^{-3} & 0.025 \end{pmatrix}$$

El vector de coeficientes estimados será:

$$\hat{\beta} = (x'x)^{-1} x'y = \begin{pmatrix} 0.125 \\ -0.500 \end{pmatrix}$$

(b) La matriz de varianzas y covarianzas estimada de los estimadores:

$$\hat{V}(\hat{\beta}_i) = \hat{\sigma}_u^2 (x'x)^{-1} = \frac{e'e}{N-k}(x'x)^{-1} = \frac{y'y - \hat{\beta}'x'y}{N-k}(x'x)^{-1} =$$

$$= \frac{670 - (0.125 \quad -0.5)\begin{pmatrix} 8000 \\ 700 \end{pmatrix}}{5-3}(x'x)^{-1} =$$

$$= \frac{20}{2}(x'x)^{-1} = \begin{pmatrix} 2.125 \cdot 10^{-3} & -2.23 \cdot 10^{-2} \\ -2.230 \cdot 10^{-2} & 0.25 \end{pmatrix}$$

(c) El coeficiente de determinación se obtiene como:

$$R^2 = \frac{\hat{\beta}'x'y}{y'y} = \frac{(0.125 \quad -0.5)\begin{pmatrix} 8000 \\ 700 \end{pmatrix}}{670} = \frac{650}{670} = 0.97$$

EJERCICIO 1.24

La Consejería de Turismo del Cabildo Insular de Gran Canaria desea estimar la relación que hay entre los ingresos por turismo en la Comunidad Canaria

(Y), los precios medios de los paquetes turísticos (X_2) y la renta de los turistas que visitan el archipiélago (X_3). El modelo a estimar es

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

La muestra está formada por datos anuales desde 1981 a 1990. A partir de dichas series, se han obtenido los siguientes datos:

$$\begin{array}{lll} \overline{Y} = 64.5 & \overline{X}_2 = 7.4 & \overline{X}_3 = 146.4 \\ S_Y^2 = 0.7 & S_{X_2}^2 = 1.6 & S_{X_3}^2 = 465.2 \\ Cov(X_2, Y) = -0.04 & Cov(X_3, Y) = 16.8 & Cov(X_2, X_3) = 8.2 \end{array}$$

Además se sabe que, para una estimación MCO, $\hat{\beta}_2 = -0.2309$ y $\hat{\beta}_3 = 0.0402$.

(a) Obtenga el coeficiente estimado para la constante.

(b) Obtenga el coeficiente de determinación.

Solución

(a) Sabiendo que la coordenada $(\overline{Y}, \overline{X}_2, \overline{X}_3)$ es un punto de la recta de regresión del modelo respecto al origen, se verifica: $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3$. Por tanto, despejando se obtiene:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 = 64.5 - (-0.2309) \cdot 7.4 - 0.0402 \cdot 146.4 = 60.32$$

(b) El coeficiente de determinación con datos centrados se calcula mediante la siguiente expresión:

$$R^2 = \frac{\hat{\beta}' x' x \hat{\beta}}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

La matriz $x'x$ la obtenemos despejando los datos a partir de las varianzas y covarianzas de las variables:

$$S_{x_2}^2 = \frac{\sum_{i=1}^{N} x_{2i}^2}{N} \Rightarrow \sum_{i=1}^{N} x_{2i}^2 = N \cdot S_{x_2}^2 = 10 \cdot 1.6 = 16$$

Análogamente:

$$\sum_{i=1}^{N} x_{3i}^2 = NS_{x_3}^2 = 10S_{x_3}^2 = 10 \cdot 465.2 = 4652$$

$$Cov(x_2, x_3) = \frac{\sum_{i=1}^{N} x_{2i}x_{3i}}{N} \Rightarrow \sum_{i=1}^{N} x_{2i}x_{3i} = 10 \cdot 8.2 = 82$$

$$S_y^2 = \frac{\sum_{i=1}^{N} y_i^2}{N} \Rightarrow \sum_{i=1}^{N} y_i^2 = 10 \cdot 0.7 = 70$$

Por tanto, el valor pedido es:

$$R^2 = \frac{\hat{\beta}' x' x \hat{\beta}}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} =$$

$$= \frac{\begin{pmatrix} -0.2309 & 0.0402 \end{pmatrix} \begin{pmatrix} 16 & 82 \\ 82 & 4652 \end{pmatrix} \begin{pmatrix} -0.2309 \\ 0.0402 \end{pmatrix}}{70} = \frac{6.848}{70} = 0.097$$

EJERCICIO 1.25

Considere el siguiente modelo de regresión expresado en desviaciones respecto a las medias:

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

para el que se han calculado, con una muestra de tamaño 100, los siguientes valores resumen:

$$\sum_{i=1}^{N} y_i^2 = \frac{493}{3} \qquad \sum_{i=1}^{N} x_{2i}^2 = 30 \qquad \sum_{i=1}^{N} x_{3i}^2 = 3$$

$$\sum_{i=1}^{N} x_{2i}y_i = 30 \qquad \sum_{i=1}^{N} x_{3i}y_i = 20 \qquad \sum_{i=1}^{N} x_{2i}x_{3i} = 0$$

Calcule:

(a) Las estimaciones MCO de los parámetros β_2 y β_3.

(b) Las desviaciones típicas de los estimadores obtenidos.

(c) El valor del coeficiente de determinación.

Solución

(a) Sabemos que

$$\hat{\beta} = (x'x)^{-1} x'y$$

por lo que necesitamos calcular el valor de las matrices $x'x$ y $x'y$.

Conociendo la expresión de cada uno de los elementos que conforman estas matrices, resulta sencillo deducir sus valores.

$$x'x = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^2 & \sum_{i=1}^{N} x_{2i}x_{3i} \\ \sum_{i=1}^{N} x_{2i}x_{3i} & \sum_{i=1}^{N} x_{3i}^2 \end{pmatrix} = \begin{pmatrix} 30 & 0 \\ 0 & 3 \end{pmatrix} \quad x'y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}y_i \\ \sum_{i=1}^{N} x_{3i}y_i \end{pmatrix} = \begin{pmatrix} 30 \\ 20 \end{pmatrix}$$

Calculando la inversa de la matriz $x'x$ obtenemos

$$(x'x)^{-1} = \frac{1}{90}\begin{pmatrix} 3 & 0 \\ 0 & 30 \end{pmatrix} = \begin{pmatrix} \frac{1}{30} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

Por tanto, los estimadores mínimo cuadráticos valdrán

$$\hat{\beta} = \begin{pmatrix} \frac{1}{30} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} 30 \\ 20 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{20}{3} \end{pmatrix}$$

quedando el modelo estimado en los siguientes términos:

$$\hat{y}_i = x_{2i} + \frac{20}{3} x_{3i}$$

(b) La matriz de varianzas y covarianzas de los estimadores toma la siguiente expresión:

$$V(\hat{\beta}) = \sigma_u^2 (x'x)^{-1}$$

En donde σ_u^2 es desconocida y hay que estimarla mediante la expresión $\hat{\sigma}_u^2 = e'e/(N-k)$. De esta forma obtenemos la matriz estimada de varianzas y covarianzas de los estimadores, $\hat{V}(\hat{\beta})$. Ésta recoge en su diagonal

principal las varianzas de los estimadores, por lo que calculando su raíz obtenemos su desviación típica.

Calculamos en primer lugar el valor de $\hat{\sigma}_u^2$:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{y'y - \hat{\beta}'x'y}{N-k} = \frac{\frac{493}{3} - \begin{pmatrix} 1 & \frac{20}{3} \end{pmatrix}\begin{pmatrix} 30 \\ 20 \end{pmatrix}}{100-3} = \frac{\frac{493}{3} - \frac{490}{3}}{100-3} = \frac{1}{97}$$

A continuación ya se puede obtener la matriz de varianzas y covarianzas estimada

$$\hat{V}\left(\hat{\beta}\right) = \frac{1}{97}\begin{pmatrix} \frac{1}{30} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{2910} & 0 \\ 0 & \frac{1}{291} \end{pmatrix}$$

Por lo que la desviación típica de los estimadores para $\hat{\beta}_2$ y $\hat{\beta}_3$ es, respectivamente,

$$S\left(\hat{\beta}_2\right) = \sqrt{\frac{1}{2910}} = 0.0185, \quad S\left(\hat{\beta}_3\right) = \sqrt{\frac{1}{291}} = 0.0586$$

(c) El coeficiente de determinación se puede obtener a través de la siguiente expresión:

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{e'e}{y'y} = 1 - \frac{1}{\frac{493}{3}} = 0.9939$$

EJERCICIO 1.26

Con una muestra de 100 datos para las variables Y, X_2 y X_3 se han obtenido unas medias respectivas de 10, 30 y 5 unidades. También se ha calculado la matriz de varianzas y covarianzas muestrales obteniéndose el siguiente resultado:

$$\begin{array}{c} \\ Y \\ X_2 \\ X_3 \end{array}\begin{array}{c} \begin{array}{ccc} Y & X_2 & X_3 \end{array} \\ \begin{pmatrix} 200 & 12 & 15 \\ & 60 & 25 \\ & & 15 \end{pmatrix} \end{array}$$

(a) Estime el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

(b) Analice la bondad del ajuste.

(c) Interprete los coeficientes estimados.

Solución

(a) La forma más rápida de resolver el ejercicio es mediante la notación del modelo centrado. Las matrices se pueden escribir de la siguiente forma:

$$x'x = \begin{pmatrix} N \cdot S_{X_2}^2 & N \cdot S_{X_2X_3} \\ N \cdot S_{X_3X_2} & N \cdot S_{X_3}^2 \end{pmatrix} \quad x'y = \begin{pmatrix} N \cdot S_{YX_2} \\ N \cdot S_{YX_3} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$$

Por tanto, como $\hat{\beta} = (x'x)^{-1} x'y$ se obtiene:

$$\hat{\beta} = (x'x)^{-1} x'y = \begin{pmatrix} 6000 & 2500 \\ 2500 & 1500 \end{pmatrix}^{-1} \begin{pmatrix} 1200 \\ 1500 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.0005454 & -0.00090910 \\ -0.0009091 & 0.00218182 \end{pmatrix} \begin{pmatrix} 1200 \\ 1500 \end{pmatrix} = \begin{pmatrix} -0.71 \\ 2.18 \end{pmatrix}$$

Estos dos parámetros estimados coinciden con los del modelo en niveles para las variables X_2 y X_3 respectivamente. El estimador de β_1 se obtiene como

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 = 10 - (-0.71) \cdot 30 - 2.19 \cdot 5 = 20.35$$

(b) Teniendo en cuenta que se cumple

$$R^2 = \frac{\hat{\beta}' x'y}{y'y}$$

y que $y'y = N \cdot S_Y^2$ es inmediato obtener el coeficiente de determinación de la siguiente forma:

$$R^2 = \frac{\hat{\beta}' x'y}{y'y} = \frac{(-0.71 \quad 2.18) \cdot \begin{pmatrix} 1200 \\ 1500 \end{pmatrix}}{100 \cdot 200} = 0.12$$

Esto significa que únicamente el 12% de los cambios de Y está explicado por la regresión estimada.

(c) El valor estimado para un individuo con valor igual a cero en las dos variables explicativas es igual a 20.35. Además, por cada unidad que se incremente la variable X_2, la variable Y se reduce, en promedio, en 0.71 unidades, mientras que por cada unidad que se incremente la variable X_3, la variable Y también se ve incrementada en un promedio de 2.18 unidades.

EJERCICIO 1.27

Dado el modelo (1.4):

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1.4}$$

se obtiene el siguiente resultado tras una estimación MCO, tomando una muestra de tamaño 6:

$$\hat{\sigma}_u = 0.42008 \quad R^2 = 0.8967$$

(a) Calcule la suma de cuadrados totales (SCT).

(b) Dado el modelo (1.4), se toma una muestra de tamaño 20. La varianza de Y es 250; la covarianza entre X_2 e Y es 69.57, y la varianza de X_2 es 121. ¿Cuál es la cifra de la varianza de Y no explicada por la regresión?

Solución

(a) La *SCT* la obtenemos despejándola a partir de la expresión del coeficiente de determinación:

$$R^2 = 1 - \frac{e'e}{SCT} \Rightarrow SCT \cdot R^2 = SCT - e'e \Rightarrow$$

$$\Rightarrow SCT\left(R^2 - 1\right) = -e'e \Rightarrow SCT = \frac{e'e}{1 - R^2}$$

Necesitamos calcular $e'e$:

$$\hat{\sigma}_u^2 = \frac{e'e}{(N-k)} \Rightarrow e'e = (N-k)\hat{\sigma}_u^2 \Rightarrow e'e = (6-2)\cdot 0.42008^2 = 0.7058$$

Sustituyendo:

$$SCT = \frac{e'e}{1-R^2} = \frac{0.7058}{1-0.8967} = 6.833$$

(b) La varianza de Y no explicada por la regresión coincide con la varianza de los errores SCE/N. En regresión simple (regresión con una única variable explicativa), el cuadrado del coeficiente de correlación entre X_2 e Y coincide con el coeficiente de determinación. De esta manera:

$$r = \frac{Cov(x,y)}{S_x S_y} = \frac{69.57}{\sqrt{121}\cdot\sqrt{250}} = 0.4$$

Por tanto, el coeficiente de determinación valdrá

$$R^2 = r^2 = 0.16$$

con lo que el 84% (100%–16%) de la varianza de Y no es explicada por la regresión, concretamente $250 \cdot 0.84 = 210$.

EJERCICIO 1.28

Se sabe que la variable número de piezas fabricadas depende linealmente del número de trabajadores que tiene la empresa. Se toma una muestra de 100 empresas y se sabe que el número medio de piezas fabricadas es 12000, el número medio de trabajadores es 20, la correlación entre las piezas fabricadas y los trabajadores empleados es 0.8 y las varianzas de los trabajadores y de las piezas fabricadas son, respectivamente, 20 y 6000.

(a) Estime el modelo e interprete los coeficientes estimados.

(b) Calcule la bondad del ajuste e interprétela.

(c) Estime la varianza del coeficiente estimado del número medio de trabajadores.

Solución

(a) El ejercicio se resolverá utilizando el modelo centrado respecto a la media, ya que los cálculos son más sencillos que empleando el modelo con datos respecto al origen.

Sabiendo que la varianza de X es:

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N} = S_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} = \frac{x'x}{N} = 20$$

Despejando, obtenemos que $x'x = 2000$.

Por otro lado, para calcular $x'y$ calculamos la covarianza entre X e Y a partir de su correlación:

$$r_{xy} = \frac{Cov(x,y)}{S_x S_y} \Rightarrow Cov(x,y) = r_{xy} S_x S_y = 0.8 \cdot \sqrt{20}\sqrt{6000} = 277.128$$

Análogamente, a partir de la covarianza entre X e Y, despejando, podemos obtener $x'y$:

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N} = Cov(x,y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} = \frac{x'y}{N} \Rightarrow$$

$$\Rightarrow x'y = NCov(x,y) = 100 \cdot 277.128 = 27712.8$$

Por tanto, los coeficientes MCO son:

$$x'x = 2000 \quad x'y = 27712.8$$

$$\Rightarrow \hat{\beta} = (x'x)^{-1} x'y = \frac{1}{2000} \cdot 27712.8 = 13.85$$

$$\Rightarrow \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} = 12000 - 13.85 \cdot 20 = 11723$$

El modelo estimado resultante es el siguiente:

$$\hat{Y}_i = 11723 + 13.85 X_i$$

Estimamos que, en términos medios, el incremento de un trabajador supone un incremento de 13.85 piezas fabricadas.

(b) El coeficiente de determinación en el modelo centrado es:

$$R^2 = \frac{\hat{\beta} x'y}{y'y} = \frac{13.25 \cdot 27712.8}{100 \cdot 6000} = 0.639$$

Lo que implica que casi el 64% de las variaciones en el número de piezas fabricadas viene explicado por nuestro modelo.

(c) Dado que,

$$\hat{V}(\hat{\beta}_2) = \hat{\sigma}_u^2 (x'x)^{-1}$$

y teniendo en cuenta que

$$\hat{\sigma}_u^2 = \frac{y'y - \hat{\beta}' x'y}{N-k} = 2205.89$$

la varianza estimada de $\hat{\beta}_2$ es $\hat{V}(\hat{\beta}_2) = 1.1$.

EJERCICIO 1.29

Repita el ejercicio 1.28 pero usando los siguientes datos:

$$N=300 \quad \overline{Y}=800 \quad \overline{X}=15 \quad r_{XY}=-0.7 \quad S_Y^2=300 \quad S_X^2=50$$

Solución

(a) La resolución del presente ejercicio es análoga al del ejercicio 1.28. A partir de los datos se obtiene la siguiente matriz correspondiente al modelo centrado:

$$x'x=50\cdot 300=15000$$

Para obtener $x'y$ despejamos $Cov(x,y)$ a partir de la correlación entre x e y:

$$r_{xy}=\frac{Cov(x,y)}{S_xS_y}; \quad Cov(x,y)=r_{xy}S_xS_y=(-0.7)\sqrt{50}\sqrt{300}=-85.73$$

Sabiendo que $Cov(x,y)=\dfrac{\sum_{i=1}^{N}x_{2i}y_i}{N}$, despejando, obtenemos:

$$\sum_{i=1}^{N}x_{2i}y_i=x'y=N\cdot(-85.73)=300(-85.73)=-25719$$

Los coeficientes MCO, entonces, son:

$$\hat{\beta}_2=(x'x)^{-1}x'y=\frac{1}{50\cdot 300}\cdot(-25719)=-1.7146$$

$$\hat{\beta}_1=\overline{Y}-\hat{\beta}_2\overline{X}_2=800+1.7146\cdot 15=825.72$$

El modelo estimado es $\hat{Y}_i=825.72-1.715X_i$. Por tanto, al incrementar en una unidad X, Y se reduce, en promedio, en 1.715 unidades.

(b) Para el cálculo el coeficiente de determinación trabajamos también con datos centrados:

$$R^2=\frac{\hat{\beta}'x'y}{y'y}=\frac{1.7146\cdot 25719}{300\cdot 300}=0.49$$

De esta forma, el coeficiente de determinación es igual a 0.49, lo que significa que el 49% de las variaciones de Y es explicado por las variaciones en X.

(c) Sabiendo que:

$$\hat{\sigma}_u^2 = \frac{y'y - \hat{\beta}'x'y}{N-k} = \frac{300 \cdot 300 - 1.7146 \cdot 25719}{300-2} = 154.03$$

La varianza estimada de $\hat{\beta}_2$ es igual a 0.01, tal y como se muestra a continuación:

$$\Rightarrow \hat{V}\left(\hat{\beta}_2\right) = \hat{\sigma}_u^2 \left(x'x\right)^{-1} = 154.03 \cdot \frac{1}{50 \cdot 30} = 0.01$$

EJERCICIO 1.30

Dado el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

del cual se conocen los siguientes datos no centrados:

$$N = 30 \qquad \sum_{i=1}^{N} Y_i = 294.97429 \qquad \sum_{i=1}^{N} X_{2i} = 324.837911$$

$$\sum_{i=1}^{N} X_{3i} = 1097.4243 \qquad \sum_{i=1}^{N} Y_i^2 = 3057.52282 \qquad \sum_{i=1}^{N} X_{2i}^2 = 3697.6731$$

$$\sum_{i=1}^{N} X_{3i}^2 = 40330.6001 \qquad Cov\left(Y, X_2\right) = 0.555511 \qquad Cov\left(Y, X_3\right) = 0.541541$$

y la siguiente matriz para datos centrados:

$$\left(x'x\right)^{-1} = \begin{pmatrix} 0.387352 & -0.353140 \\ -0.353140 & 0.375728 \end{pmatrix}$$

(a) Obtenga el valor de los estimadores mínimo cuadrático ordinarios.

(b) ¿Cuál de las dos variables tiene un mayor peso en el modelo estimado? ¿Por qué?

Solución

(a) Al tener la matriz $\left(x'x\right)^{-1}$ con datos centrados, realizaremos todos los cálculos con datos centrados. De esta forma obtendremos los valores de $\hat{\beta}_2$ y $\hat{\beta}_3$ y, por último, sólo tendremos que estimar el valor de $\hat{\beta}_1$.

Sabemos que

$$\hat{\beta} = (x'x)^{-1} x'y$$

Por tanto, necesitamos calcular el valor de la matriz $x'y$, cuyos elementos son los que se muestran a continuación:

$$x'y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i} y_i \\ \sum_{i=1}^{N} x_{3i} y_i \end{pmatrix}$$

Sabemos que $Cov(Y, X_2) = \frac{1}{N} \sum_{i=1}^{N} x_{2i} y_i = 0.555511$ y despejando obtenemos

$$\sum_{i=1}^{N} x_{2i} y_i = 16.66533.$$

Igualmente $Cov(Y, X_3) = \frac{1}{N} \sum_{i=1}^{N} x_{3i} y_i = 0.541541$ y despejando obtenemos

$$\sum_{i=1}^{N} x_{3i} y_i = 16.24623.$$

De esta forma, ya tenemos todos los elementos de la matriz $x'y$ y podemos calcular los valores de $\hat{\beta}$ como mostramos a continuación:

$$\hat{\beta} = (x'x)^{-1} x'y = \begin{pmatrix} 0.387352 & -0.353140 \\ -0.353140 & 0.375728 \end{pmatrix} \begin{pmatrix} 16.66533 \\ 16.24623 \end{pmatrix} = \begin{pmatrix} 0.718155 \\ 0.218968 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$$

El término constante, que no se obtiene directamente cuando trabajamos con datos centrados, se puede obtener a partir de la expresión de la recta de regresión expresada en términos de valores medios:

$$\rightarrow \quad \overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3$$

Sustituyendo los valores conocidos, obtenemos el valor de $\hat{\beta}_1$:

$$\frac{294.97429}{30} = \hat{\beta}_1 + 0.718155 \cdot \frac{324.837911}{30} + 0.218968 \cdot \frac{1097.4243}{30} \Rightarrow$$

$$\Rightarrow \hat{\beta}_1 = -5.953717$$

(b) Para responder a esta pregunta habrá que calcular el modelo estandarizado y ver qué coeficiente toma el valor más alto.

Los coeficientes estandarizados se pueden obtener a partir de los coeficientes originales mediante la siguiente relación:

$$\hat{\beta}_j^s = \hat{\beta}_j \frac{S_{X_j}}{S_Y} \quad j = 1,2,3$$

Necesitamos calcular, en primer lugar, los valores de las desviaciones típicas de las variables explicativas y de la variable endógena.

$$S_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2} = \sqrt{\frac{1}{30} \cdot 3057.52282 - \left(\frac{294.97429}{30}\right)^2} = 2.28906891$$

$$S_{X_2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_{2i}^2 - \bar{X}_2^2} = \sqrt{\frac{1}{30} \cdot 3697.6731 - \left(\frac{324.837911}{30}\right)^2} = 2.3263049$$

$$S_{X_3} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_3^2 - \bar{X}_3^2} = \sqrt{\frac{1}{30} \cdot 40330.6001 - \left(\frac{1097.4243}{30}\right)^2} = 2.48951329$$

De esta forma ya podemos obtener los valores de los dos coeficientes estandarizados (recordemos que el modelo estandarizado pierde el término constante):

$$\hat{\beta}_2^s = \hat{\beta}_2 \frac{S_{X_2}}{S_Y} = 0.718155 \cdot \frac{2.3263049}{2.28906891} = 0.7298$$

$$\hat{\beta}_3^s = \hat{\beta}_3 \frac{S_{X_3}}{S_Y} = 0.218968 \cdot \frac{2.48951329}{2.28906891} = 0.2381$$

Por tanto, basándonos en los resultados obtenidos, podemos afirmar que la variable X_2 es la que tiene un mayor peso en el modelo estimado, puesto que es la que tiene un coeficiente estandarizado mayor en valor absoluto.

EJERCICIO 1.31

Disponemos del siguiente modelo de regresión:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

donde conocemos los siguientes resultados del modelo expresado con datos respecto al origen:

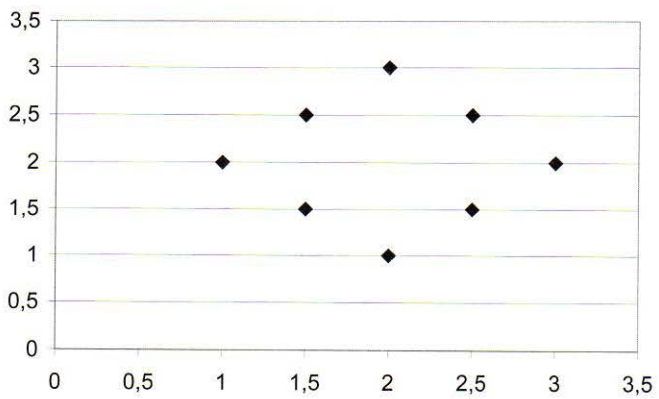
$$X'X = \begin{pmatrix} 8 & 16 & 16 \\ 16 & 35 & 34 \\ 16 & 34 & 36 \end{pmatrix} \quad X'Y = \begin{pmatrix} 16 \\ 32 \\ 32 \end{pmatrix}$$

(a) Calcule con datos centrados los coeficientes estimados del modelo (incluida la constante) e interprete los resultados.

(b) Calcule el coeficiente de determinación trabajando con datos centrados e interprete su resultado.

(c) ¿Cuál será el valor del coeficiente de correlación de un modelo de regresión simple, si su diagrama de dispersión es el siguiente? Explique su respuesta.

Gráfico 1.1



Solución

(a) Sabiendo que

$$N = 8 \quad \sum_{i=1}^{N} X_{2i} = 16 \quad \sum_{i=1}^{N} X_{3i} = 16 \quad \sum_{i=1}^{N} X_{2i}^2 = 35 \quad \sum_{i=1}^{N} X_{3i}^2 = 36$$

$$\sum_{i=1}^{N} X_{2i}X_{3i} = 34 \quad \sum_{i=1}^{N} Y_i = 16 \quad \sum_{i=1}^{N} X_{2i}Y_i = 32 \quad \sum_{i=1}^{N} X_{3i}Y_i = 32$$

necesitamos calcular:

$$x'x = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^2 & \sum_{i=1}^{N} x_{2i}x_{3i} \\ \sum_{i=1}^{N} x_{2i}x_{3i} & \sum_{i=1}^{N} x_{3i}^2 \end{pmatrix} \quad x'y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}y_i \\ \sum_{i=1}^{N} x_{3i}y_i \end{pmatrix}$$

Por ello, en primer lugar obtenemos los valores medios de las variables explicativas.

$$\bar{X}_2 = \frac{16}{8} = 2 \quad \bar{X}_3 = \frac{16}{8} = 2$$

En segundo lugar, las sumas de datos centrados podemos obtenerlas de las siguientes expresiones:

$$\sum_{i=1}^{N} x_{2i}^2 = \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_2 \right)^2 = \sum_{i=1}^{N} X_{2i}^2 - N\bar{X}_2^2 = 35 - 8 \cdot 2^2 = 3$$

$$\sum_{i=1}^{N} x_{3i}^2 = \sum_{i=1}^{N} \left(X_{3i} - \bar{X}_3 \right)^2 = \sum_{i=1}^{N} X_{3i}^2 - N\bar{X}_3^2 = 36 - 8 \cdot 2^2 = 4$$

$$\sum_{i=1}^{N} x_{2i} x_{3i} = \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_2 \right)\left(X_{3i} - \bar{X}_3 \right) =$$

$$= \sum_{i=1}^{N} X_{2i} X_{3i} - N\bar{X}_3 \frac{\sum_{i=1}^{N} X_{2i}}{N} - N\bar{X}_2 \frac{\sum_{i=1}^{N} X_{3i}}{N} + N\bar{X}_2\bar{X}_3 =$$

$$= \sum_{i=1}^{N} X_{2i} X_{3i} - N\bar{X}_2\bar{X}_3 = 34 - 8 \cdot 2 \cdot 2 = 2$$

Por tanto:

$$x'x = \begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix}$$

Conociendo la media de Y, calculamos los datos correspondientes a la matriz $x'y$ de forma análoga. Como $\bar{Y} = \frac{16}{8} = 2$:

$$\sum_{i=1}^{N} x_{2i} y_i = \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_2 \right)\left(Y_i - \bar{Y} \right) = \sum_{i=1}^{N} X_{2i} Y_i - N\bar{X}_2\bar{Y} = 32 - 8 \cdot 2 \cdot 2 = 0$$

$$\sum_{i=1}^{N} x_{3i} y_i = \sum_{i=1}^{N} \left(X_{3i} - \bar{X}_3 \right)\left(Y_i - \bar{Y} \right) = 0$$

Y, por tanto,

$$x'y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

De esta forma:

$$\hat{\beta} = \left(x'x \right)^{-1} x'y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Para el cálculo de $\hat{\beta}_1$ aplicamos la propiedad de que las coordenadas de los valores medios pertenecen a la recta de regresión del modelo original: $\overline{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3$.

Sustituyendo: $2 = \hat{\beta}_1 + 0\overline{X}_2 + 0\overline{X}_3$, de donde obtenemos que la constante vale 2.

(b) Sustituyendo en la expresión del coeficiente de determinación:

$$R^2 = \frac{\hat{\beta}' x' y}{y' y} = \frac{SCR}{SCT} = \frac{(0 \quad 0)\begin{pmatrix} 0 \\ 0 \end{pmatrix}}{SCT} = 0$$

Como vemos, el modelo no explica nada de la varianza de Y.

(c) El coeficiente de correlación de una regresión lineal cuyos datos son los representados en el gráfico es cero, ya que la relación lineal entre X e Y es completamente inexistente, siendo por tanto la covarianza cero, y consecuentemente, también la correlación:

$$r = \frac{Cov(x, y)}{S_x S_y} = \frac{0}{S_x S_y} = 0$$

EJERCICIO 1.32

Dado el siguiente modelo estandarizado:

$$\hat{y}_i^s = -0.1432 x_{2i}^s + 1.1143 x_{3i}^s$$

(a) ¿Cuál de los dos regresores ejerce mayor influencia sobre la endógena?

(b) Sabiendo que la matriz de datos estandarizados es

$$x^{s\prime} x^s = \begin{pmatrix} 5 & 4.4688 \\ 4.4688 & 5 \end{pmatrix}$$

y que las desviaciones típicas de los datos centrados son: $S_y = 1.4142$, $S_{x_2} = 0.6633$ y $S_{x_3} = 1.7204$, obtenga los coeficientes estimados del modelo centrado.

(c) Dado el coeficiente de determinación corregido del modelo centrado ($\overline{R}^2 = 0.9541$) y sabiendo que $N = 5$, obtenga el coeficiente de determinación.

(d) Con la información disponible, calcule la Suma de Cuadrados de los Errores (SCE) del modelo centrado.

Solución

(a) La variable X_3 tiene mayor peso sobre Y que X_2, ya que su coeficiente estandarizado en términos absolutos es mayor que el de X_2.

(b) Si denominamos los datos estandarizados como x_{2i}^s, x_{3i}^s y los centrados como x_{2i}, x_{3i}, podemos expresar la matriz de datos estandarizados $x^{s\prime}x^s$ en términos de los datos centrados de la siguiente manera:

$$x^{s\prime}x^s = \begin{pmatrix} \sum_{i=1}^N x_{2i}^{s2} & \sum_{i=1}^N x_{2i}^s x_{3i}^s \\ \sum_{i=1}^N x_{2i}^s x_{3i}^s & \sum_{i=1}^N x_{3i}^{s2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N \left(\frac{x_{2i}}{S_{x_2}}\right)^2 & \sum_{i=1}^N \frac{x_{2i}x_{3i}}{S_{x_2}S_{x_3}} \\ \sum_{i=1}^N \frac{x_{2i}x_{3i}}{S_{x_2}S_{x_3}} & \sum_{i=1}^N \left(\frac{x_{3i}}{S_{x_3}}\right)^2 \end{pmatrix} = \begin{pmatrix} 5 & 4.4688 \\ 4.4688 & 5 \end{pmatrix}$$

Por tanto, despejando, obtenemos la matriz de datos centrados en función de los estandarizados:

$$x'x = \begin{pmatrix} \sum_{i=1}^N x_{2i}^2 & \sum_{i=1}^N x_{2i}x_{3i} \\ \sum_{i=1}^N x_{2i}x_{3i} & \sum_{i=1}^N x_{3i}^2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N x_{2i}^{s2}S_{x_2}^2 & \sum_{i=1}^N x_{2i}^s x_{3i}^s S_{x_2}S_{x_3} \\ \sum_{i=1}^N x_{2i}^s x_{3i}^s S_{x_2}S_{x_3} & \sum_{i=1}^N x_{3i}^{s2}S_{x_3}^2 \end{pmatrix} =$$

$$= \begin{pmatrix} 5\cdot 0.6633^2 & 4.4688\cdot 0.6633\cdot 1.7204 \\ 4.4688\cdot 0.6633\cdot 1.7204 & 5\cdot 1.7204^2 \end{pmatrix} = \begin{pmatrix} 2.2 & 5.1 \\ 5.1 & 14.8 \end{pmatrix}$$

Para obtener $x^{s\prime}y^s$, sabemos que:

$$x^{s\prime}y^s = \left(x^{s\prime}x^s\right)\hat{\beta}^s = \begin{pmatrix} 5 & 4.4688 \\ 4.4688 & 5 \end{pmatrix}\begin{pmatrix} -0.1432 \\ 1.1143 \end{pmatrix} = \begin{pmatrix} 4.2635 \\ 4.9315 \end{pmatrix}$$

La matriz $x'y$ de datos centrados la despejamos análogamente a partir de la matriz conocida de datos estandarizados $x^{s\prime}x^s$. Por tanto, a partir de la matriz de datos estandarizados,

$$x^{s\prime}y^{s} = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^{s} y_{3i}^{s} \\ \sum_{i=1}^{N} x_{3i}^{s} y_{i}^{s} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \frac{x_{2i}^{s} y_{i}^{s}}{S_{x_2} S_y} \\ \sum_{i=1}^{N} \frac{x_{3i}^{2} y_{i}^{s}}{S_{x_3} S_y} \end{pmatrix} = \begin{pmatrix} 4.2635 \\ 4.9315 \end{pmatrix}$$

podemos despejar la matriz de datos centrados:

$$x'y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^{s} y_{i}^{s} S_{x_2} S_y \\ \sum_{i=1}^{N} x_{3i}^{s} y_{i}^{s} S_{x_3} S_y \end{pmatrix} = \begin{pmatrix} 4 \\ 12 \end{pmatrix}$$

El vector de coeficientes correspondiente al modelo con datos centrados se obtiene calculando la expresión:

$$\hat{\beta} = (x'x)^{-1} x'y$$

La matriz inversa es:

$$(x'x)^{-1} = \begin{pmatrix} 2.2595 & -0.7786 \\ -0.7786 & 0.3358 \end{pmatrix}$$

De forma que:

$$\hat{\beta} = \begin{pmatrix} -0.3053 \\ 0.9160 \end{pmatrix}$$

Otra forma de resolver el ejercicio mucho más sencilla es teniendo en cuenta la relación existente entre los coeficientes estandarizados y los del modelo centrado: es fácil deducir a partir de las expresiones de ambos modelos que los coeficientes estandarizados se pueden obtener a partir de los del modelo centrado, o viceversa, estando relacionados de la siguiente forma:

$$\hat{\beta}_j = \frac{\hat{\beta}_j^{s} S_y}{S_{x_j}}$$

Por tanto:

$$\hat{\beta}_2 = \frac{\hat{\beta}_2^{s} S_y}{S_{x_2}} = \frac{-0.1432188 \cdot 1.4142}{0.6633} = -0.305$$

$$\hat{\beta}_3 = \frac{\hat{\beta}_3^{s} S_y}{S_{x_3}} = \frac{1.1143 \cdot 1.4142}{1.7204} = 0.9159$$

(c) Conociendo la expresión que relaciona el coeficiente de determinación corregido en función del coeficiente de determinación:

$$\overline{R}^2 = 1 - \left(\frac{N-1}{N-k}\right)\left(1-R^2\right)$$

despejando R^2 obtenemos que:

$$R^2 = \frac{(N-k)\overline{R}^2 - (N-k)}{N-1} + 1 = \frac{(5-3)0.9541 - (5-3)}{4} + 1 = 0.977$$

(d) A partir de la expresión del coeficiente de determinación corregido:

$$\overline{R}^2 = 1 - \frac{SCE/(N-k)}{SCT/(N-1)}$$

despejamos la *SCE*:

$$SCE = \left(\frac{SCT}{N-1} - \frac{SCT}{N-1}\overline{R}^2\right)(N-k)$$

Por otro lado, la *SCT* podemos obtenerla a partir de la varianza de Y, ya que:

$$SCT = S_y^2 N = 1.4142^2 \cdot 5 = 9.999$$

Sustituyendo:

$$SCE = \left(\frac{9.999}{4} - \frac{9.999}{4}0.9541\right)2 = 0.229$$

EJERCICIO 1.33

Dados los siguientes valores para datos centrados:

$$\sum_{i=1}^{N} x_{2i}y_i = 1 \quad \sum_{i=1}^{N} x_{3i}y_i = 52 \quad \sum_{i=1}^{N} x_{2i}^2 = 4 \quad \sum_{i=1}^{N} x_{3i}^2 = 82 \quad \sum_{i=1}^{N} x_{2i}x_{3i} = 3 \quad \sum_{i=1}^{N} y_i^2 = 45.5$$

(a) Obtenga los estimadores MCO del modelo centrado $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 z_i$, donde $z_i = 100x_{3i}$. Utilice todas las matrices necesarias para obtener dicha estimación.

(b) ¿Qué valores tomarán $\hat{\beta}_2$ y $\hat{\beta}_3$ si el modelo estimado por MCO fuera ahora $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$?

(c) Calcule los coeficientes de determinación de ambos modelos y compárelos.

(d) Calcule la varianza estimada de la perturbación aleatoria para el modelo del primer apartado. ¿Cree que coincidirá con la del otro modelo?

Solución

(a) Tendremos que construir la matriz z de esta nueva especificación y obtener los coeficientes estimados del modelo aplicando la expresión: $\hat{\beta} = (z'z)^{-1} z'y$

El nuevo modelo será: $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 100 x_{3i}$, donde

$$z = \begin{pmatrix} x_{21} & 100x_{31} \\ x_{22} & 100x_{32} \\ \dots & \dots \\ x_{2N} & 100x_{3N} \end{pmatrix} \quad z'z = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^2 & \sum_{i=1}^{N} 100x_{2i}x_{3i} \\ \sum_{i=1}^{N} 100x_{2i}x_{3i} & \sum_{i=1}^{N} x_{3i}^2 100^2 \end{pmatrix} \quad z'y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}y_i \\ \sum_{i=1}^{N} 100x_{3i}y_i \end{pmatrix}$$

$$\Rightarrow \hat{\beta} = \begin{pmatrix} 4 & 300 \\ 300 & 820000 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 5200 \end{pmatrix} = \begin{pmatrix} -0.231900 \\ 0.006426 \end{pmatrix}$$

(b) El valor de $\hat{\beta}_2$ coincide en ambos modelos, mientras que $\hat{\beta}_3$ valdrá 0.6426 en el modelo $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$. Como vemos, su valor queda multiplicado por 100 con respecto al modelo del apartado anterior.

(c) El R^2 para el modelo $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 z_i$ es:

$$R^2 = \frac{\hat{\beta}' z'z \hat{\beta}}{y'y} = \frac{(-0.2319 \quad 0.006426)\begin{pmatrix} 4 & 300 \\ 300 & 820000 \end{pmatrix}\begin{pmatrix} -0.231900 \\ 0.006426 \end{pmatrix}}{45.5} = 0.729$$

El R^2 del modelo $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ es el mismo, ya que la única diferencia entre ambos es un cambio de escala en la segunda variable explicativa.

(d) La varianza estimada de las perturbaciones del modelo $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 z_i$ es:

$$e'e = y'y - \hat{\beta}' z'z \hat{\beta} = 45.5 - 33.1849 = 12.31 \Rightarrow \hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{12.31}{6-3} = 4.103$$

La varianza de ambos modelos coincide ya que, como se ha comentado, la transformación que se ha realizado a la variable x_3 es simplemente un cambio de escala que no afecta al ajuste del modelo.

EJERCICIO 1.34

Disponemos del siguiente modelo de regresión simple estimado:

$$\hat{Y}_i = -346.5909 + 1.0113636 X_i \qquad (1.5)$$

donde:

$$N = 6 \quad \sum_{i=1}^{6} Y_i = 5000 \quad \sum_{i=1}^{6} X_i^2 = 8460000 \quad \sum_{i=1}^{6} X_i = 7000 \quad \sum_{i=1}^{6} X_i Y_i = 6130000$$

Se realiza la siguiente transformación a la variable X:

$$Z_i = \frac{50 + X_i}{2}$$

(a) Con la información suministrada, obtenga todas las matrices para la estimación de los coeficientes estimados del siguiente modelo:

$$Y_i = \beta_1 + \beta_2 Z_i + u_i \qquad (1.6)$$

(b) Obtenga la expresión de los coeficientes de determinación de los modelos (1.5) y (1.6) y compare ambos valores razonando los resultados obtenidos.

Solución

(a) La nueva matriz Z del modelo (1.6) será:

$$Z = \begin{pmatrix} 1 & \frac{50 + X_1}{2} \\ 1 & \frac{50 + X_2}{2} \\ \dots & \dots \\ 1 & \dots \\ 1 & \frac{50 + X_6}{2} \end{pmatrix}$$

Los coeficientes estimados se obtendrán calculando: $\hat{\beta} = (Z'Z)^{-1} Z'Y$

Los cálculos necesarios se muestran a continuación:

$$Z'Z = \begin{pmatrix} 1 & 1 & \cdots & \cdots & 1 \\ \frac{50+X_1}{2} & \frac{50+X_2}{2} & \cdots & \cdots & \frac{50+X_6}{2} \end{pmatrix} \begin{pmatrix} 1 & \frac{50+X_1}{2} \\ 1 & \frac{50+X_2}{2} \\ \cdots & \cdots \\ 1 & \cdots \\ 1 & \frac{50+X_6}{2} \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & \sum_{i=1}^{6}\left(\frac{50}{2}+\frac{X_i}{2}\right) \\ \sum_{i=1}^{6}\left(\frac{50}{2}+\frac{X_i}{2}\right) & \sum_{i=1}^{6}\left(\frac{50}{2}+\frac{X_i}{2}\right)^2 \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & \left(\frac{6\cdot 50}{2}+\frac{\sum_{i=1}^{6} X_i}{2}\right) \\ \left(\frac{6\cdot 50}{2}+\frac{\sum_{i=1}^{6} X_i}{2}\right) & \sum_{i=1}^{6}\left(\frac{50^2}{2^2}+\frac{X_i^2}{2^2}+2\frac{50X_i}{2^2}\right) \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & \left(\frac{6\cdot 50}{2}+\frac{\sum_{i=1}^{6} X_i}{2}\right) \\ \left(\frac{6\cdot 50}{2}+\frac{\sum_{i=1}^{6} X_i}{2}\right) & \left(6\frac{50^2}{2^2}+\frac{\sum_{i=1}^{6} X_i^2}{2^2}+2\frac{50\sum_{i=1}^{6} X_i}{2^2}\right) \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & 3650 \\ 3650 & 2293750 \end{pmatrix}$$

$$(Z'Z)^{-1} = \begin{pmatrix} 5.21306800 & -0.00829545 \\ -0.00829545 & 1.3636\cdot 10^{-5} \end{pmatrix}$$

$$Z'Y = \begin{pmatrix} 1 & 1 & \dots & \dots & 1 \\ \frac{50+X_1}{2} & \frac{50+X_2}{2} & \dots & \dots & \frac{50+X_6}{2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ \dots \\ Y_6 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{6} Y_i \\ \sum_{i=1}^{6} \frac{(50+X_i)}{2} Y_i \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{6} Y_i \\ \frac{50}{2}\sum_{i=1}^{6} Y_i + \frac{\sum_{i=1}^{6} X_i Y_i}{2} \end{pmatrix} = \begin{pmatrix} 5000 \\ \frac{50}{2}5000 + \frac{6130000}{2} \end{pmatrix} = \begin{pmatrix} 5000 \\ 3190000 \end{pmatrix}$$

$$\hat{\beta} = (Z'Z)^{-1} Z'Y = \begin{pmatrix} 5.213068 & -0.008295 \\ -0.008295 & 1.36 \cdot 10^{-5} \end{pmatrix} \begin{pmatrix} 5000 \\ 3190000 \end{pmatrix} =$$

$$= \begin{pmatrix} -397.159090 \\ 2.022727 \end{pmatrix}$$

(b) Coeficiente de determinación del modelo (1.5):

$$R_1^2 = \frac{\hat{\beta}' X'Y - N\bar{Y}^2}{Y'Y - N\bar{Y}^2} =$$

$$= \frac{\begin{pmatrix} -346.5909 & 1.0113636 \end{pmatrix} \begin{pmatrix} 5000 \\ 6130000 \end{pmatrix} - N\bar{Y}^2}{Y'Y - N\bar{Y}^2} = \frac{4466704.5 - N\bar{Y}^2}{Y'Y - N\bar{Y}^2}$$

Coeficiente de determinación del modelo (1.6):

$$R_2^2 = \frac{\hat{\beta}' Z'Y - N\bar{Y}^2}{Y'Y - N\bar{Y}^2} =$$

$$= \frac{\begin{pmatrix} -397.15909 & 2.022727 \end{pmatrix} \begin{pmatrix} 5000 \\ 3190000 \end{pmatrix} - N\bar{Y}^2}{Y'Y - N\bar{Y}^2} = \frac{4466704.5 - N\bar{Y}^2}{Y'Y - N\bar{Y}^2}$$

Como vemos, ambos modelos tienen exactamente el mismo ajuste. Se ha realizado un cambio de escala y origen en el modelo inicial, que no afecta a la bondad del ajuste.

2

El modelo básico de regresión lineal múltiple: Contrastación y predicción

EJERCICIO 2.1

La Tabla 2.1 contiene distintos estadísticos de prueba, todos ellos relacionados con el modelo de regresión lineal múltiple (MRLM). Indique cuales son las hipótesis nula y alternativa con las que está asociado cada uno de estos estadísticos.

Tabla 2.1

Estadístico relacionado con el MRLM
$t = \dfrac{(\hat{\beta}_j - \beta_{j0})}{\hat{\sigma}_u \sqrt{a_{jj}}}$
$t = \dfrac{\hat{\beta}_j}{\hat{\sigma}_u \sqrt{a_{jj}}}$
$\dfrac{R^2/(k-1)}{(1-R^2)/(N-k)}$
$\dfrac{\left((R\hat{\beta}-r)'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta}-r)\right)/q}{e'e/(N-k)}$

Solución

Denotando el modelo de regresión lineal múltiple como

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

en la Tabla 2.2 se muestran los contrastes que se asocian con cada uno de los estadísticos.

Tabla 2.2

Estadístico relacionado con el MRLM	Contraste
$t = \dfrac{(\hat{\beta}_j - \beta_{j0})}{\hat{\sigma}_u \sqrt{a_{jj}}}$	$H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j \neq \beta_{j0}$ $H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j > \beta_{j0}$ $H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j < \beta_{j0}$
$t = \dfrac{\hat{\beta}_j}{\hat{\sigma}_u \sqrt{a_{jj}}}$	$H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$ $H_0 : \beta_j = 0$ $H_1 : \beta_j > 0$ $H_0 : \beta_j = 0$ $H_1 : \beta_j < 0$
$\dfrac{R^2/(k-1)}{(1-R^2)/(N-k)}$	$H_0 : \beta_2 = \beta_3 = ... = \beta_i = ... = \beta_k = 0$ H_1 : Al menos un $\beta_j \neq 0$
$\dfrac{\left((R\hat{\beta} - r)'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - r)\right)/q}{e'e/(N-k)}$	$H_0 : R\beta = r$ $H_1 : No\ H_0$

EJERCICIO 2.2

Se ha estimado el siguiente modelo con $N = 100$:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

obteniendo un coeficiente de determinación de 0.95. Este mismo modelo se ha vuelto a estimar eliminando en esta ocasión la variable X_4, obteniéndose un coeficiente de determinación de 0.94.

Con los datos suministrados y sabiendo que la varianza de Y es igual a 50000, ¿es el coeficiente de la variable X_4 significativamente distinto de cero en el primer modelo estimado? Justifique la respuesta mediante la realización de un contraste.

Solución

El ejercicio nos pide un contraste cuya hipótesis nula es

$$H_0 : \beta_4 = 0$$

Con los datos del enunciado es posible calcular el estadístico F siguiente, que, bajo la hipótesis nula como cierta, se distribuye como una F de Fisher-Snedecor con 1 y 96 grados de libertad.

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)}$$

El subíndice R indica que la suma de los cuadrados de los errores se corresponde con la del modelo con restricciones y el NR con la del modelo sin restricciones, siendo m el número de restricciones con el que se ha estimado el modelo restringido con respecto al modelo no restringido.

Para calcular el estadístico de prueba sabemos que se cumple:

$$R^2_{NR} = 1 - \frac{e'e_{NR}}{Y'Y - N\bar{Y}^2} \Rightarrow e'e_{NR} = \left(1 - R^2_{NR}\right)\left(Y'Y - N\bar{Y}^2\right)$$

$$S^2_Y = \frac{Y'Y - N\bar{Y}^2}{N} \Rightarrow Y'Y - N\bar{Y}^2 = S^2_Y \cdot N = 50000 \cdot 100 = 5000000$$

Por tanto,

$$e'e_{NR} = (1 - 0.95)5000000 = 250000$$

De igual forma, partiendo del coeficiente de determinación del modelo con restricciones, tenemos que

$$R^2_R = 1 - \frac{e'e_R}{Y'Y - N\bar{Y}^2} \Rightarrow e'e_R = \left(1 - R^2_R\right)\left(Y'Y - N\bar{Y}^2\right)$$

$$e'e_R = (1 - 0.94) \cdot 5000000 = 300000$$

Partiendo de la información calculada, obtenemos el valor del estadístico de prueba:

$$F = \frac{(300000 - 250000)}{250000/96} = 19.2$$

El valor de la F de Fisher-Snedecor de 1 y 96 grados de libertad que deja a su derecha una probabilidad igual a 0.05, $F^{0.05}_{1,96}$, es 3.94. En consecuencia se rechaza la hipótesis nula, es decir, se rechaza que $\beta_4 = 0$ y que, por tanto, la especificación del modelo sea la restringida.

EJERCICIO 2.3

Con los datos del ejercicio 1.30 y sabiendo que el primer elemento de la diagonal principal de la matriz $(X'X)^{-1}$ toma el valor $a_{11} = 26.87798$, ¿cree que es necesario introducir el término constante en el modelo? Demuéstrelo.

Solución

Para comprobar la necesidad de introducir el término independiente en el modelo, necesitamos realizar un contraste de significación individual en el que contrastemos si el mismo puede tomar valor cero.

El contraste a realizar, por tanto, es el siguiente:

$$\left.\begin{array}{l} H_0:\ \beta_1 = 0 \\ H_1:\ \beta_1 \neq 0 \end{array}\right\}$$

siendo el estadístico de contraste:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S\left(\hat{\beta}_1\right)}$$

En el ejercicio 1.30 obtuvimos el valor de $\hat{\beta}_1$, por lo que sólo tenemos que calcular el valor de $S\left(\hat{\beta}_1\right)$.

Sabemos que

$$S\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}_u^2 \cdot a_{11}} = \sqrt{\frac{e'e}{N-k} \cdot a_{11}} = \sqrt{\frac{y'y - \hat{\beta}'x'y}{N-k} \cdot a_{11}}$$

De esta expresión tan sólo desconocemos el valor de $y'y$, puesto que el resto de valores o los aporta el enunciado o los hemos calculado en el ejercicio 1.30. No obstante, este valor ($y'y$) lo podemos obtener a partir de la varianza de Y, que fue calculada en el ejercicio 1.30.

Trabajando con datos centrados se cumple que

$$S_y^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 = 2.28906891^2 .$$

Por tanto, despejando obtenemos que

$$\sum_{i=1}^{N} y_i^2 = 30 \cdot 2.28906891^2 = 157.19509 = y'y$$

De esta forma, sustituyendo $y'y$ y $S(\hat{\beta}_1)$ en la expresión del estadístico t, obtenemos:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} = \frac{-5.953717 - 0}{\sqrt{\frac{157.19509 - \begin{pmatrix} 0.718155 & 0.218968 \end{pmatrix}\begin{pmatrix} 16.66533 \\ 16.24623 \end{pmatrix}}{30-3}} \cdot 26.97798} =$$

$$= -0.5013415$$

Comparando el valor del estadístico con el valor crítico de una distribución t-Student con 27 grados de libertad ($t_{N-k}^{\alpha/2} = t_{27}^{0.025} = \pm 2.056$), para un nivel de significación del 5% bilateral, no podemos rechazar la hipótesis nula, o lo que es lo mismo, no podemos rechazar que $\beta_1 = 0$, por lo que hablamos de la falta de significatividad estadística de la constante en el modelo.

EJERCICIO 2.4

Se desea estudiar el efecto que, sobre las pernoctaciones hoteleras por habitante en diferentes provincias (P), tiene la renta per capita (R), así como tres variables dicotómicas relativas a si la provincia está ubicada en la costa (ZC), si pertenece a un archipiélago (ZA) o si la provincia pertenece al interior (ZI). Esta última variable queda como modalidad de referencia. Si se obtienen los siguientes resultados de la estimación:

$$\hat{P}_i = -0.24 + 0.24R_i + 0.60ZC_i + 1.6ZA_i$$

y se dispone de la siguiente información: $\hat{\sigma}_u^2 = 1.58$,

$$(XX)^{-1} = \begin{pmatrix} 0.884 & -0.0850 & -0.056 & -0.0010 \\ -0.085 & 0.0090 & 0.002 & 0.0001 \\ -0.056 & 0.0020 & 0.086 & -0.0500 \\ -0.001 & 0.0001 & -0.050 & 0.3830 \end{pmatrix}$$

(a) Indique qué variables son estadísticamente significativas a nivel individual, usando como nivel de significación el 5%. ¿Y si usamos el nivel de significación del 10%?

(b) Construya un intervalo de confianza del 95% para β_3 y otro para β_4, es decir, para el coeficiente de las dicotómicas Zona de Costa y Zona Archipiélago.

Solución

(a) La significación estadística de los coeficientes del modelo se corresponde con los siguientes contrastes:

$$\left.\begin{array}{l} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{array}\right\} \forall j = 2,3,4$$

Los contrastes los realizamos a través de la prueba *t*-Student y para ello calculamos el estadístico de prueba particularizado bajo la hipótesis nula como sigue:

$$\frac{\hat{\beta}_j - \beta_j}{S\left(\hat{\beta}_j\right)} \sim t_{N-k}$$

Previamente tenemos que obtener los elementos de la diagonal principal de la matriz de varianzas y covarianzas de los estimadores mediante la expresión:

$$\hat{V}\left(\hat{\beta}\right) = \hat{\sigma}_u^2 \left(X'X\right)^{-1} = 1.58\left(X'X\right)^{-1}$$

Es decir,

$$S^2\left(\hat{\beta}_2\right) = 0.014 \quad S^2\left(\hat{\beta}_3\right) = 0.136 \quad S^2\left(\hat{\beta}_4\right) = 0.605$$

Con esta información ya es posible calcular los distintos estadísticos de prueba, bajo el cumplimiento de la hipótesis nula.

$$\frac{\hat{\beta}_2 - \beta_2}{S\left(\hat{\beta}_2\right)} \sim t_{N-k} \Rightarrow \frac{0.24 - 0}{\sqrt{0.01}} = 2.013$$

$$\frac{\hat{\beta}_3 - \beta_3}{S\left(\hat{\beta}_3\right)} \sim t_{N-k} \Rightarrow \frac{0.60 - 0}{\sqrt{0.14}} = 1.63$$

$$\frac{\hat{\beta}_4 - \beta_4}{S\left(\hat{\beta}_4\right)} \sim t_{N-k} \Rightarrow \frac{1.60 - 0}{\sqrt{0.60}} = 2.06$$

El valor crítico en las tablas de la *t* de Student es $t_{51-4}^{0.025} = 2.01$ y podemos concluir que las variables Renta (R) y Zona Archipiélago (ZA) son estadísticamente significativas a nivel individual y no lo es la variable Zona Costa (ZC).

Como el valor de $t_{51-4}^{0.05} = 1.68$, la respuesta sería la misma que la aportada si utilizamos como nivel de significación el 10%.

(b) Intervalo de confianza para el coeficiente de la Zona de Costa:

$$P\left(\hat{\beta}_3 - t_{N-k}S\left(\hat{\beta}_3\right) \leq \beta_3 \leq \hat{\beta}_3 + t_{N-k}S\left(\hat{\beta}_3\right)\right) = 0.95$$

$$P\left(0.60 - 2.01 \cdot 0.37 \leq \beta_3 \leq 0.60 + 2.01 \cdot 0.37\right) = 0.95$$

$$P\left(-0.15 \leq \beta_3 \leq 1.35\right) = 0.95$$

El intervalo de confianza del 95% contiene el valor 0, luego la variable no es estadísticamente significativa.

Intervalo de confianza para el coeficiente de la Zona Archipiélago:

$$P\left(\hat{\beta}_4 - t_{N-k}S\left(\hat{\beta}_4\right) \leq \beta_4 \leq \hat{\beta}_4 + t_{N-k}S\left(\hat{\beta}_4\right)\right) = 0.95$$

$$P\left(1.6 - 2.01 \cdot 0.77 \leq \beta_4 \leq 1.6 + 2.01 \cdot 0.77\right) = 0.95$$

$$P\left(0.04 \leq \beta_4 \leq 3.16\right) = 0.95$$

Al no contener el intervalo de confianza del 95% el valor 0 podemos decir que la variable *ZA* es estadísticamente significativa.

EJERCICIO 2.5

Dado el modelo estimado del Cuadro 2.1, donde se explica el comportamiento de las importaciones españolas (*IMPORT*) en función del *PIB*, del consumo (*CONS*) y de la Formación Bruta de Capital (*FBC*), complete los valores desconocidos indicando cómo ha obtenido los mismos.

Cuadro 2.1

Dependent Variable: *IMPORT*
Method: Least Squares
Sample: 1969 1986
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PIB	0.032204	0.186884	A	0.8656
CONS	0.242747	0.285361	0.850667	0.4093
FBC	0.414199	0.322260	1.285296	0.2195
C	−19.72511	4.125253	−4.781551	B
R-squared	0.973043	Sum squared resid		71.390
Adjusted R-squared	C	*F*-statistic		E
S.E. of regression	D	Prob(*F*-statistic)		F

Solución

El estadístico *t*-Student (A) lo calculamos como

$$\frac{\hat{\beta}_j}{S\left(\hat{\beta}_j\right)} = \frac{0.032}{0.186} = 0.172$$

La probabilidad asociada a la significatividad de la constante (B) viene dada por la probabilidad de que una distribución *t*-Student de $18-4=14$ grados de libertad sea inferior a –4.781551. Buscando en las tablas de dicha distribución se observa que la probabilidad es prácticamente igual a cero.

El Coeficiente de determinación corregido (C) se obtiene en función del valor del R^2.

$$\overline{R}^2 = 1-\left(1-R^2\right)\frac{N-1}{N-k} = 1-\left(1-0.973\right)\frac{17}{14} = 0.967$$

La desviación típica estimada de las perturbaciones (D) se obtiene como sigue.

$$\hat{\sigma}_u = \sqrt{\frac{e'e}{N-k}} = \sqrt{\frac{71.39}{18-4}} = 2.258$$

Para el cálculo del estadístico de prueba *F* de significación global (E) se usa aquella expresión en la que es función del Coeficiente de determinación del modelo.

$$\frac{R^2/(k-1)}{\left(1-R^2\right)/(N-k)} = \frac{0.973/(4-1)}{(1-0.973)/(18-4)} = 168.17$$

Por último, la probabilidad asociada al estadístico (*F*) es 0, dado que el valor del estadístico se encuentra en el extremo de la distribución. Recordemos que el punto crítico del contraste de significación global, para un nivel de significación del 5%, viene dado por el valor $F_{3,14}^{0.05} = 3.34$.

EJERCICIO 2.6

Dado el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

sujeto a las siguientes restricciones: $R\beta = r$, escriba el contenido de las matrices *R* y *r* que son necesarias para realizar el siguiente contraste de hipótesis:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_3 \quad \text{y} \quad \beta_4 + \beta_5 = 1 \quad \text{y} \quad 2\beta_2 = \beta_5 \\ H_1: \text{No } H_0 \end{array}\right\}$$

Solución

La matriz de restricciones es la siguiente, colocando cada restricción por filas:

$$R = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 & -1 \end{pmatrix}$$

Y la hipótesis nula se escribiría de la siguiente manera:

$$H_0: R\beta = r \Rightarrow \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

EJERCICIO 2.7

Demuestre que el estadístico de prueba para el siguiente contraste conjunto:

$$\left.\begin{array}{l} H_0: \beta_j = 0 \\ H_1: \text{No } H_0 \end{array}\right\} \forall j = 1,2,...k$$

se puede escribir como

$$F = \frac{\hat{\beta}' X' Y}{k\hat{\sigma}_u^2}.$$

Solución

Partiendo de la expresión

$$\frac{\left(\left(R\hat{\beta} - r\right)' \left(R\left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right)\right) \Big/ q}{\hat{\sigma}_u^2}$$

y teniendo en cuenta que tenemos k restricciones conjuntas de nulidad, que R es una matriz identidad de orden $k \cdot k$ y que r es un vector de ceros, el estadístico de prueba lo podemos escribir como

$$\frac{\left[\hat{\beta}'(X'X)\hat{\beta}\right]}{k\hat{\sigma}_u^2}.$$

Sustituyendo $\hat{\beta}$ por $(X'X)^{-1}X'Y$ demostramos que el estadístico de prueba se puede escribir como

$$F = \frac{\hat{\beta}'X'Y}{k\hat{\sigma}_u^2}.$$

EJERCICIO 2.8

Se ha estimado el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

y se ha obtenido el siguiente resultado:

$$\hat{Y}_i = -0.0636 + 1.44X_{2i} - 0.4898X_{3i}$$

Sabiendo que

$$(X'X)^{-1} = \begin{pmatrix} 0.1011 & -0.0007 & -0.0005 \\ -0.0007 & 0.0231 & -0.0162 \\ -0.0005 & -0.0162 & 0.0122 \end{pmatrix}$$

y que $\hat{\sigma}_u^{\,2} = 3.1799$, contraste la siguiente hipótesis conjunta al 95% de nivel de confianza:

$$\left.\begin{array}{l} H_0: \beta_1 + 3\beta_2 = 2 \quad \text{y} \quad \beta_3 = 1 \\ H_1: \text{No } H_0 \end{array}\right\}$$

Solución

La expresión matricial de la hipótesis nula a contrastar es:

$$H_0: \begin{pmatrix} 1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

El estadístico de contraste es:

$$F = \frac{\left(\left(R\hat{\beta}-r\right)'\left(R\left(X'X\right)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)\right)/q}{e'e/(N-k)}$$

Por tanto,

$$F = \frac{1}{3.1799}\left[\left[\begin{pmatrix}1 & 3 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}-0.0636\\ 1.4400\\ -0.4898\end{pmatrix}-\begin{pmatrix}2\\ 1\end{pmatrix}\right]'\left[\begin{pmatrix}1 & 3 & 0\\ 0 & 0 & 1\end{pmatrix}\left(X'X\right)^{-1}\begin{pmatrix}1 & 0\\ 3 & 0\\ 0 & 1\end{pmatrix}\right]^{-1}\cdot\right.$$

$$\left.\cdot\left[\begin{pmatrix}1 & 3 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}-0.0636\\ 1.4400\\ -0.4898\end{pmatrix}-\begin{pmatrix}2\\ 1\end{pmatrix}\right]\Big/2\right] =$$

$$= \frac{\begin{pmatrix}2.256 & -1.4898\end{pmatrix}\begin{pmatrix}0.3048 & -0.0491\\ -0.0491 & 0.0122\end{pmatrix}^{-1}\begin{pmatrix}2.2564\\ -1.4898\end{pmatrix}\Big/2}{3.1799} =$$

$$= \frac{\begin{pmatrix}2.256 & -1.4898\end{pmatrix}\begin{pmatrix}9.329 & 37.54\\ 37.54 & 233.07\end{pmatrix}\begin{pmatrix}2.2564\\ -1.4898\end{pmatrix}\Big/2}{3.1799} = 49.117$$

Bajo la hipótesis nula como cierta, el estadístico F se distribuye como una F de Fisher-Snedecor de 2 y $N-3$ grados de libertad. Para poder obtener dicho valor crítico necesitamos conocer el valor de N. Para ello sabemos que el elemento (1,1) de la matriz $X'X$ es dicho valor N. Por tanto, como conocemos la matriz $\left(X'X\right)^{-1}$, lo que hacemos es invertirla para calcular $X'X$. El elemento (1,1) nos proporciona el valor de N, que, en este caso, es 10. Por tanto, buscamos en las tablas el valor $F_{2,7}^{0.05}$ y obtenemos que el valor crítico es igual a 4.74. Como dicho valor es más pequeño que el estadístico de prueba, rechazamos la hipótesis nula. Es decir, la evidencia empírica no apoya el cumplimiento conjunto de las restricciones $\beta_1+3, \beta_2=2$ y $\beta_3=1$.

EJERCICIO 2.9

A partir de los datos del ejercicio 1.25 realice los siguientes contrastes de hipótesis:

(a) H_0: $\beta_3=6$

(b) H_0: $\beta_2=\beta_3=0$

Solución

(a) Se trata de un contraste de significación individual, por lo que el estadístico de contraste es

$$t = \frac{\hat{\beta}_3 - \beta_3}{S\left(\hat{\beta}_3\right)}$$

En el ejercicio 1.25 estimamos los parámetros y sus desviaciones típicas, por lo que tenemos todos los valores necesarios para realizar el contraste de hipótesis. En este caso, el estadístico de prueba toma el valor

$$t = \frac{\frac{20}{3} - 6}{\sqrt{\frac{1}{291}}} = 11.372$$

Comparando dicho valor con el valor tabulado para una $t_{100-3}^{0.05/2} = t_{97}^{0.025} \approx 1.96$ comprobamos que el estadístico cae en la región de rechazo, por lo que rechazamos la hipótesis nula de que el parámetro β_3 tome el valor 6.

(b) En este apartado se trata de realizar el contraste de significación global, puesto que la hipótesis nula propone la nulidad de todos los coeficientes, excepto de la constante. El estadístico de contraste en este caso es

$$F = \frac{R^2\left(N-k\right)}{\left(1-R^2\right)\left(k-1\right)}$$

En el apartado (c) del ejercicio 1.25 calculamos el coeficiente de determinación, por lo que tenemos todos los datos necesarios para realizar el contraste de hipótesis. Así, el estadístico de prueba toma el valor

$$F = \frac{0.9939\left(100-3\right)}{\left(1-0.9939\right)\left(3-1\right)} = 7902.32$$

Siendo el valor del estadístico tan alto, al compararlo con el valor tabulado para una $F_{3-1,100-3}^{\alpha} = F_{2,97}^{0.05}$, éste caerá en la región de rechazo, por lo que no podemos aceptar la hipótesis nula que indica que el modelo no es globalmente significativo.

EJERCICIO 2.10

A partir de los siguientes modelos estimados, contraste la hipótesis nula de que la propensión marginal de las ventas respecto del precio de la competencia es igual a 2 mediante la utilización de tres expresiones diferentes.

Cuadro 2.2

Dependent Variable: *VENTAS*
Sample: 1 17
Included observations: 17

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PRECIO	1.879350	0.176388	10.65462	0.0000
PUBLICIDAD	7.517483	2.905409	2.587410	0.0215
C	48436.86	1954.543	24.78167	0.0000
R-squared	0.896844	Mean dependent var		64730.86
Adjusted R-squared	0.882107	S.D. dependent var		6026.041
S.E. of regresión	2069.074	Akaike info criterion		18.26637
Sum squared resid	59934913	Schwarz criterion		18.41341

Cuadro 2.3

Sample: 1 17
Included observations: 17
VENTAS = 2**PRECIO* + *C*(1)* *PUBLICIDAD* + *C*(2)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	7.466297	2.852460	2.617494	0.0194
C(2)	47670.30	1572.672	30.31166	0.0000
R-squared	0.893396	Mean dependent var		64730.86
Adjusted R-squared	0.886289	S.D. dependent var		6026.041
S.E. of regression	2032.041	Akaike info criterion		18.18160
Sum squared resid	61937841	Schwarz criterion		18.27962

Se sabe, además, que la inversa de $x'x$ (con datos centrados) del modelo recogido en el Cuadro 2.2 es:

$$(x'x)^{-1} = \begin{pmatrix} 7.267 \cdot 10^{-9} & -3.0830 \cdot 10^{-9} \\ -3.083 \cdot 10^{-9} & 1.9718 \cdot 10^{-6} \end{pmatrix}$$

Solución

i. La opción más sencilla es realizar el siguiente contraste individual a partir del estadístico de contraste *t*-Student con los datos de la salida del Cuadro 2.2.

$$\left.\begin{array}{l} H_0: \beta_2 = 2 \\ H_1: \beta_2 \neq 2 \end{array}\right\}$$

El estadístico de contraste es:

$$\frac{\hat{\beta}_2 - \beta_2}{S(\hat{\beta}_2)} = \frac{1.8793 - 2}{0.1763} = -0.686$$

El valor crítico t_{17-3} al 95% de nivel de confianza es 2.14. Por tanto, no podemos rechazar H_0.

ii. Otra opción para realizar el contraste es utilizar los modelos restringido y no restringido para obtener las sumas de cuadrados de los errores de ambos modelos.

El estadístico de contraste en este caso será:

$$F = \frac{(SCE_R - SCE_{NR})/m}{SCE_{NR}/(N-k)} = \frac{(61\,937\,841 - 59\,934\,913)/1}{59\,934\,913/(17-3)} = 0.4678$$

El valor crítico de una $F_{1,14}$ al 95% de nivel de confianza es 4.6, llegando, por tanto, a la misma conclusión que en el apartado (i).

Nota: Tener en cuenta que el valor del estadístico F de Fisher-Snedecor coincide con el cuadrado del estadístico t:

$$(t\text{-Student})^2 = F_{m,N-k}, \text{ es decir, } (-0.686)^2 = 0.4678.$$

iii. Por último, siempre podemos utilizar el estadístico general para el contraste de combinaciones lineales de parámetros:

$$\left.\begin{array}{l} H_0: R\beta = r \\ H_1: R\beta \neq r \end{array}\right\} \Rightarrow \left.\begin{array}{l} H_0: \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 2 \\ \\ H_1: \text{No } H_0 \end{array}\right\}$$

El estadístico de contraste es:

$$F = \frac{\left(\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right)\right)/q}{\hat{\sigma}_u^2}$$

Trabajando con datos centrados su expresión quedaría como:

$$F = \frac{\left[\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}1.87935\\7.517483\end{pmatrix}-2\right]'\left[\begin{pmatrix}1 & 0\end{pmatrix}(x'x)^{-1}\begin{pmatrix}1\\0\end{pmatrix}\right]^{-1}\left[\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}1.87935\\7.517483\end{pmatrix}-2\right]}{2\,069.074^2} =$$

$$= 0.4678$$

El valor del estadístico coincide con el del apartado (ii), llegando obviamente a la misma conclusión por cualquiera de las tres vías utilizadas.

EJERCICIO 2.11

Una función de producción se especifica mediante el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

donde Y es el logaritmo de producción, X_2 el logaritmo de trabajo y X_3 es el logaritmo del capital. Los datos siguientes se refieren a una muestra de 23 empresas y las observaciones se miden en forma de desviaciones con respecto a sus medias muestrales.

$$\sum_{i=1}^{N} x_{2i}^2 = 12 \quad \sum_{i=1}^{N} x_{3i}^2 = 12 \quad \sum_{i=1}^{N} x_{2i}x_{3i} = 8$$

$$\sum_{i=1}^{N} y_i^2 = 10 \quad \sum_{i=1}^{N} y_i x_2 = 10 \quad \sum_{i=1}^{N} y_i x_3 = 8$$

Contraste la hipótesis de la existencia de rendimientos constantes a escala.

Solución

La estimación del modelo centrado es la siguiente:

$$x'x = \begin{pmatrix}12 & 8\\8 & 12\end{pmatrix} \quad x'y = \begin{pmatrix}10\\8\end{pmatrix} \quad (x'x)^{-1} = \begin{pmatrix}0.15 & -0.10\\-0.10 & 0.15\end{pmatrix} \Rightarrow$$

$$\Rightarrow \hat{\beta} = (x'x)^{-1}x'y = \begin{pmatrix}0.7\\0.2\end{pmatrix}$$

La estimación de la varianza de las perturbaciones es:

$$e'e = y'y - \hat{\beta}'x'x\hat{\beta} = 10 - 8.6 = 1.4 \quad \Rightarrow \quad \hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{1.4}{23-3} = 0.07$$

Las hipótesis nula y alternativa del contraste son:

$$\left.\begin{aligned} H_0&: \beta_2 + \beta_3 = 1 \\ H_1&: \beta_2 + \beta_3 \neq 1 \end{aligned}\right\}$$

y el estadístico de contraste a utilizar es:

$$\frac{\left(R\hat{\beta} - r\right)' \left(R\hat{\sigma}_u^2 \left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right)}{q}$$

cuyo valor se calcula como:

$$\left(\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 0.7 \\ 0.2 \end{pmatrix} - 1\right)' \left(\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 0.0105 & -0.0070 \\ -0.0070 & 0.0105 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^{-1} \left(\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 0.7 \\ 0.2 \end{pmatrix} - 1\right) = 1.4286$$

El punto crítico al 5% que establece una distribución F de 1 y 20 grados de libertad es 4.35. Por lo tanto, existe evidencia empírica a favor del cumplimiento de la hipótesis nula.

EJERCICIO 2.12

Como resultado de contrastar conjuntamente determinadas restricciones en el siguiente modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad i = 1,...,20 \quad SCE = 5758.092$$

se ha obtenido el siguiente valor para el estadístico $F = 2.395187$.

Conociendo que

$$\left(R\hat{\beta} - r\right)' \left[R\left(X'X\right)^{-1} R'\right]^{-1} \left(R\hat{\beta} - r\right) = 1723.96338$$

¿cuántas restricciones se han contrastado?

Solución

Partiendo de la expresión general del estadístico de prueba

$$F = \frac{\left(\left(R\hat{\beta} - r\right)' \left(R\left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right)\right) / q}{e'e/\left(N - k\right)}$$

lo que nos pide el ejercicio es calcular el valor de q. Es decir,

$$F = \frac{1723.963388}{q \cdot \frac{5758.092}{20-4}} = 2.395187 \Rightarrow q = \frac{1723.963388}{\frac{5758.092}{20-4} \cdot 2.395187} = 2$$

El número de restricciones es igual a 2.

EJERCICIO 2.13

Con información de 70 viviendas a la venta en Las Palmas de Gran Canaria, se estima la influencia en el precio de las mismas de: la superficie (X_2), el número de aseos (X_3) y la ubicación de la vivienda según zonas (Puerto (X_4) y Vegueta (X_5)), tomando como referencia la zona de Ciudad Alta (X_6).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

La estimación del modelo con el programa Eviews proporciona los resultados recogidos en el Cuadro 2.4.

Cuadro 2.4

Dependent Variable: *PRECIOS*
Method: Least Squares
Sample: 1 70
Included observations: 70

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–20382.88			
SUPERFICIE_M2	1108.712			
ASEOS	28152.82			
VEGUETA	43861.50			
PUERTO	33634.22			
R-squared	0.740432	Mean dependent var		136157.20000
Adjusted R-squared	0.724459	S.D. dependent var		51621.71000
S.E. of regression	27097.26	Akaike info criterion		23.32100
Sum squared resid	4.77E+10	Schwarz criterion		23.48161
Log likelihood	–811.2351	*F*-statistic		46.35413
Durbin-Watson stat	2.089120	Prob(*F*-statistic)		0.00000

Si además dispone de la siguiente información:

$$(X'X)^{-1} = \begin{pmatrix} 0.258 & -0.003000 & 0.007 & 0.007000 & 0.074 \\ -0.003 & 0.000055 & -0.002 & -0.000194 & -0.001 \\ 0.007 & -0.002000 & 0.099 & 0.012000 & -0.003 \\ 0.007 & -0.000194 & -0.003 & 0.162000 & 0.025 \\ 0.074 & -0.001000 & 0.012 & 0.025000 & 0.551 \end{pmatrix}$$

analice si las variables son significativas a nivel individual.

Solución

La significación estadística de los coeficientes del modelo la realizamos a través de la prueba *t*-Student y para ello calculamos el estadístico de prueba particularizado bajo la hipótesis nula como sigue:

$$\left.\begin{array}{l} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{array}\right\} \forall j=2,3,4,5$$

El estadístico de prueba para el contraste se calcula a partir de la siguiente expresión:

$$\frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)}$$

Los resultados de la estimación en el Eviews nos proporcionan el valor de $\hat{\sigma}_u = 27097.26$ por lo que $\hat{\sigma}_u^2 = 27097.26^2 = 734261500$.

Para realizar los contrastes de significación individual necesitamos los elementos de la diagonal principal de la matriz de varianzas y covarianzas de los estimadores.

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_u^2 (X'X)^{-1} = 734261500(X'X)^{-1} =$$

$$= \begin{pmatrix} 189439466.900 & -2202784.4990 & 5139830.497 & 5139830.497 & 54335350.9600 \\ -2202784.499 & 40364.8600 & -1468522.999 & -146401.620 & -734261.4995 \\ 5139830.497 & -1468522.9990 & 72691888.450 & -2202784.499 & 8811137.9940 \\ 5139830.497 & -146401.6200 & -2202784.499 & 118950362.900 & 18356537.4900 \\ 54335350.960 & -734261.4995 & 8811137.994 & 18356537.490 & 404578086.2000 \end{pmatrix}$$

Por tanto,

$$S(\hat{\beta}_2) = 13763.7 \quad S(\hat{\beta}_3) = 200.91 \quad S(\hat{\beta}_4) = 8525.95 \quad S(\hat{\beta}_5) = 10906.43$$

Los estadísticos de prueba proporcionan los siguientes resultados para cada uno de los coeficientes:

$$\frac{\hat{\beta}_2 - \beta_2}{S\left(\hat{\beta}_2\right)} = \frac{1108.712 - 0}{13763.7} = 0.08$$

$$\frac{\hat{\beta}_3 - \beta_3}{S\left(\hat{\beta}_3\right)} = \frac{28152.82 - 0}{200.91} = 140.1$$

$$\frac{\hat{\beta}_4 - \beta_4}{S\left(\hat{\beta}_4\right)} = \frac{43861.5 - 0}{8525.95} = 5.14$$

$$\frac{\hat{\beta}_5 - \beta_5}{S\left(\hat{\beta}_5\right)} = \frac{33634.22 - 0}{10906.43} = 3.08$$

Como los tres últimos estadísticos de prueba superan, en valor absoluto, al valor crítico $t_{70-5}^{0.025} \approx 2$, podemos afirmar que cada uno de estos coeficientes es significativamente distinto de cero y, por lo tanto, que las variables relativas al número de aseos y las Zonas de Vegueta y Puerto son estadísticamente significativas. No se puede rechazar, sin embargo, la nulidad del coeficiente correspondiente a la variable Superficie.

EJERCICIO 2.14

Con los datos del ejercicio 2.13 y la información que se le facilita, responda a las siguientes cuestiones:

(a) ¿Cuáles serían las hipótesis nula y alternativa si se quisiera contrastar la existencia de un efecto "zona de localización de la vivienda"?

(b) ¿Cuál sería la expresión analítica del estadístico de prueba que se debe utilizar?

(c) Si además se le facilita la siguiente estimación, ¿puede decir que existe el "efecto zona de localización de la vivienda"?

Cuadro 2.5

Dependent Variable: *PRECIOS*
Method: Least Squares
Sample: 1 70
Included observations: 70

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SUPERFICIE_M2	1217.073	218.5261	5.569464	0.0000
ASEOS	28409.970	9491.8320	2.993096	0.0039
C	–25928.650	15027.1300	–1.725455	0.0891
R-squared	0.668545	Mean dependent var		136157.200000
Adjusted R-squared	0.658651	S.D. dependent var		51621.710000
S.E. of regression	30160.010000	Akaike info criterion		23.508330
Sum squared resid	6.09E+10	Schwarz criterion		23.604700
Log likelihood	–819.791700	*F*-statistic		67.569670
Durbin-Watson stat	1.464275	Prob(*F*-statistic)		0.000000

Solución

(a) Las hipótesis nula y alternativa son:

$$\left.\begin{aligned} H_0&: \beta_{Vegueta} = \beta_{Puerto} = 0 \\ H_1&: \beta_{Vegueta} \neq 0 \ \text{ y/o } \ \beta_{Puerto} \neq 0 \end{aligned}\right\}$$

(b) La expresión analítica del estadístico de contraste en este caso es

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)}$$

(c) Sin más que sustituir los valores de las sumas de cuadrados de los residuos resultantes de las estimaciones de los modelos ampliado y restringido obtendremos

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)} = \frac{\dfrac{6.09 \cdot 10^{10} - 4.77 \cdot 10^{10}}{2}}{\dfrac{4.77 \cdot 10^{10}}{70-5}} = 8.99$$

Ese valor 8.99 tiene que compararse con el valor crítico de las tablas de la F con 2 y 65 grados de libertad, que es 3.14. Por tanto, se rechaza la hipótesis nula, lo que implica que el modelo correcto sería aquél en el que existe un efecto significativo de la zona de localización de la vivienda.

EJERCICIO 2.15

Se supone que la demanda de turismo internacional depende del nivel de renta del turista, del precio del producto turístico internacional y del precio del producto turístico en el país de origen del turista. Con el objeto de estudiar el comportamiento de la demanda de turismo internacional se tomó una muestra de 2000 turistas y se obtuvo, para cada uno de ellos, el volumen de demanda de turismo internacional, medido por el dinero gastado por cada uno de ellos en turismo fuera de su región de residencia. Asimismo, se recopila información del precio del producto turístico internacional y del precio del producto turístico de su región, además del nivel de renta de cada una de las personas seleccionadas. Se especificó un modelo de regresión lineal para el gasto turístico internacional (G) en función del nivel de renta (Y), del precio del turismo internacional (PI) y del precio del turismo regional (PR), con término independiente. Es decir, se especificó el siguiente modelo:

$$G_i = \beta_1 + \beta_2 Y_i + \beta_3 PI_i + \beta_4 PR_i + u_i \quad \forall\, i = \{1,2,...,2000\}$$

La estimación arrojó un coeficiente de determinación igual a 0.7. Estudie la significatividad estadística conjunta de las variables Y, PI y PR.

Solución

Se trata de un análisis de la varianza o, lo que es lo mismo, de un contraste de nulidad conjunta de todos los parámetros correspondientes a las variables explicativas (no incluye la constante).

En este caso, el contraste lo formalizaríamos como:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

que equivaldría a

$$\left.\begin{array}{l} H_0: \text{Las variables } Y, PI \text{ y } PR \text{ conjuntamente no tienen capacidad explicativa para explicar a } G \\ \\ H_1: \text{Las variables } Y, PI \text{ y } PR \text{ conjuntamente tienen capacidad explicativa para explicar a } G \end{array}\right\}$$

El estadístico de prueba es:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)} = \frac{R^2(N-k)}{(1-R^2)(k-1)}$$

Si se cumple la hipótesis nula este estadístico se distribuye como una F de Fisher-Snedecor de $(k-1)$ y $(N-k)$ grados de libertad. Es decir, en nuestro caso, de 3 y 1996 grados de libertad. El punto crítico al 5% de significatividad es 2.6 y el valor del estadístico de prueba es:

$$F = \frac{0.7 \cdot (2000-4)}{(1-0.7)\cdot(4-1)} = \frac{1397.2}{0.9} = 1552.4$$

Por tanto, se rechaza la hipótesis nula. Esto es, las variables Y, PI y PR tienen conjuntamente capacidad de explicar el comportamiento de la variable gasto en turismo internacional.

EJERCICIO 2.16

Dado el modelo

$$G_i = \beta_1 + \beta_2 Y_i + \beta_3 PI_i + \beta_4 PR_i + u_i \quad \forall\, i = \{1,2,...,2000\}$$

correspondiente al ejercicio 2.15, se estima el mismo y se obtiene que la suma de los cuadrados de los errores es igual a 4000, mientras que si se estima el modelo

$$G_i = \beta_1 + \beta_2 Y_i + u_i \quad \forall\, i = \{1,2,...,2000\}$$

se obtiene una suma de los cuadrados de los errores de 5000.

Contraste conjuntamente, para un nivel de significación del 5%, la hipótesis de que las dos variables de precios no tienen capacidad explicativa sobre la variable Gasto en turismo internacional (G). Escriba las matrices R y r de la expresión general del estadístico de dicho contraste.

Solución

El contraste a plantear es:

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

Es por tanto un contraste de significación de un subconjunto de parámetros. El estadístico de prueba lo podemos escribir como

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)}$$

El modelo no restringido es el que tiene todas las variables y el modelo restringido es el que incorpora el conjunto de restricciones que contiene la hipótesis nula del contraste planteado. En nuestro caso hay dos restricciones. En consecuencia, el estadístico de prueba es:

$$F = \frac{(5000 - 4000)/2}{4000/1996} = \frac{1000 \cdot 1996}{4000 \cdot 2} = 249.5$$

Si se cumpliese la hipótesis nula, el estadístico se distribuye como una F de Fisher-Snedecor de 2 y 1996 grados de libertad. El valor crítico tabulado es igual a 3 para un nivel de significación del 5%. Por tanto, se rechaza la hipótesis nula. Es decir, se rechaza que conjuntamente los precios no tengan influencia sobre la demanda turística internacional.

Si se quisiese utilizar la versión general del estadístico de prueba, la matriz R sería igual a

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

y la matriz r sería

$$r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

EJERCICIO 2.17

Dado el modelo de regresión lineal múltiple del ejercicio 2.15

$$G_i = \beta_1 + \beta_2 Y_i + \beta_3 PI_i + \beta_4 PR_i + u_i,$$

¿qué se está contrastando mediante las hipótesis

$$\left.\begin{array}{l} H_0: R\beta = r \\ H_1: \text{No } H_0 \end{array}\right\},$$

siendo R la matriz igual a $(0 \quad 1 \quad 1 \quad -1)$ y $r = (0)$?

Solución

Se está contrastando si la suma de los efectos multiplicadores de la renta y el precio del turismo internacional es igual al efecto multiplicador del precio del turismo regional.

EJERCICIO 2.18

Un grupo empresarial ha estimado un modelo para explicar el número de unidades vendidas en función del número de establecimientos que posee, de los gastos en publicidad (en miles de euros) y del precio unitario de cada unidad vendida (en euros). Se toma una muestra de 34 empresas de dicho grupo y se estima el modelo, obteniendo el siguiente resultado:

$$\hat{\beta} = \begin{pmatrix} 2 \\ 1000 \\ 700 \\ 10 \end{pmatrix} \quad e'e = 100 \quad (X'X)^{-1} = \begin{pmatrix} 0.1 & 0.40 & -0.1 & -0.10 \\ & 0.01 & -0.3 & 0.40 \\ & & 0.1 & 0.20 \\ & & & 0.02 \end{pmatrix}$$

Contraste al nivel de significación del 5% que tiene el mismo efecto sobre el número de unidades vendidas el incrementar 1000 euros en publicidad que el disponer de un establecimiento más.

Solución

El modelo con el que se está trabajando se puede especificar como sigue:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Denotamos por Y a las unidades vendidas, X_2 al número de establecimientos, X_3 a los gastos de publicidad en miles de de euros y X_4 al precio unitario en euros. Según este modelo, lo que se desea contrastar es si existe evidencia empírica para afirmar que $\beta_2 = \beta_3$. O, lo que es lo mismo, $\beta_2 - \beta_3 = 0$. Como se puede ver, éste es un caso de combinación lineal de parámetros. La matriz R es igual a

$$R = \begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}$$

y el contraste se especifica como

$$\left.\begin{array}{l} H_0: R\beta = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

El estadístico de prueba viene dado por

$$F = \frac{\left(\left(R\hat{\beta} - r\right)'\left(R\left(X'X\right)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right)\right)/q}{e'e/(N-k)}$$

Bajo la hipótesis nula como cierta, el estadístico F se distribuye como una F de Fisher-Snedecor de q grados de libertad en el numerador y $(N-k)$ en el denominador. Como sabemos, q es el número de restricciones, que coincide con las columnas de la matriz R, es decir, $q=1$. Además,

$$\left(R\hat{\beta}-r\right)=\left(\begin{pmatrix}0 & 1 & -1 & 0\end{pmatrix}\begin{pmatrix}2\\1000\\700\\10\end{pmatrix}-0\right)=1000-700-0=300$$

y,

$$\left(R\left(X'X\right)^{-1}R'\right)=\left(\begin{pmatrix}0 & 1 & -1 & 0\end{pmatrix}\begin{pmatrix}0.1 & 0.40 & -0.1 & -0.100\\0.4 & 0.01 & -0.3 & 0.400\\-0.1 & -0.30 & 0.1 & 0.200\\-0.1 & 0.40 & 0.2 & 0.002\end{pmatrix}\begin{pmatrix}0\\1\\-1\\0\end{pmatrix}\right)=$$

$$=\left(\begin{pmatrix}0.4+0.1 & 0.01+0.3 & -0.3-0.1 & 0.4-0.2\end{pmatrix}\begin{pmatrix}0\\1\\-1\\0\end{pmatrix}\right)=$$

$$=0.31+0.4=0.71$$

Por tanto, el estadístico de prueba se calcula como

$$F=\frac{\left(\left(R\hat{\beta}-r\right)'\left[R\left(X'X\right)^{-1}R'\right]^{-1}\left(R\hat{\beta}-r\right)\right)/q}{e'e/(N-k)}=$$

$$=\frac{\dfrac{300\cdot(0.71)^{-1}\cdot 300}{1}}{\dfrac{100}{34-4}}=\frac{300\cdot 300\cdot 30}{100\cdot 0.71}=38028.17$$

Si buscamos en las tablas el punto que deja a su derecha una probabilidad de 0.05 en una F de Fisher-Snedecor de 1 grado de libertad en el numerador y 30 en el denominador vemos que es igual 4.17. En consecuencia, dado que la zona de rechazo contiene a todos los valores por encima de 4.17 y el estadístico de prueba es igual a 38028.17 concluimos que se rechaza la hipótesis nula. Es decir, rechazamos que el disponer de un establecimiento más tenga el mismo

efecto sobre el número de unidades vendidas que incrementar en 1000 euros la publicidad.

EJERCICIO 2.19

Se ha obtenido la siguiente estimación con el programa Eviews, donde *SUELDO* es el valor en euros del sueldo mensual, *ANTIG* se corresponde con el número de meses de antigüedad en la empresa e *INTELIG* es la medida del coeficiente de inteligencia medida en una escala de 0 a 1.5. Además, todas las variables se refieren a una muestra de 500 individuos de una misma ciudad.

Cuadro 2.6

Dependent Variable: *SUELDO*
Method: Least Squares
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	448.60990	35.724990		
ANTIG	11.50781	0.734895		
INTELIG	810.79920	21.596750		
R-squared	0.773641	Mean dependent var		1257.64200
Adjusted R-squared	0.772730	S.D. dependent var		482.21530
S.E. of regression	229.885700	Akaike info criterion		13.71902
Sum squared resid	26265182	Schwarz criterion		13.74431
Log likelihood	−3426.756000	*F*-statistic		
Durbin-Watson stat	1.973109	Prob(*F*-statistic)		0.00000

(a) Contraste la significatividad individual de los parámetros (multiplicadores) del modelo. Especifique formalmente cada uno de los contrastes. Indique la forma general del estadístico de prueba y obtenga el valor del estadístico de prueba. Indique su decisión e interprétela.

(b) Plantee el contraste de significación global sin incluir la constante y calcule el estadístico de prueba.

(c) Con la información que se muestra en el Cuadro 2.6, ¿aceptaría o rechazaría la hipótesis nula en un contraste de significación global? Justifique la respuesta e interprete el resultado.

(d) Si un individuo de la misma ciudad de la cual se extrajo la muestra, pero que no pertenece a la misma, tiene una antigüedad en su trabajo de 24 meses y un coeficiente de inteligencia de 0.82, ¿cuál sería su sueldo esperado?

Solución

(a) Denotamos el modelo de la siguiente manera:

$$SUELDO_i = \beta_1 + \beta_2 ANTIG_i + \beta_3 INTELIG_i + u_i$$

Si denotamos por $\hat{\beta}_j$ al estimador mínimo cuadrático ordinario de β_j, con $j = 1,2,3$, y por $S^2(\hat{\beta}_j)$ a su varianza estimada, los contrastes de significatividad individual se pueden escribir de la siguiente forma:

Contraste 1	Contraste 2	Contraste 3
$H_0: \beta_1 = 0$	$H_0: \beta_2 = 0$	$H_0: \beta_3 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_2 \neq 0$	$H_1: \beta_3 \neq 0$

El estadístico de prueba es

$$t_j = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)}$$

Si se cumplen las hipótesis nulas este estadístico se distribuye como una *t* de Student de $(500-3)$ grados de libertad.

Con los datos del Cuadro 2.6 es inmediato calcular los valores de los estadísticos de prueba para los tres contrastes planteados.

Contraste 1	Contraste 2	Contraste 3
$t_1 = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)}$	$t_2 = \frac{\hat{\beta}_2}{S(\hat{\beta}_2)}$	$t_3 = \frac{\hat{\beta}_3}{S(\hat{\beta}_3)}$
$t_1 = \frac{448.61}{35.73} = 12.56$	$t_2 = \frac{11.51}{0.735} = 15.66$	$t_3 = \frac{810.8}{21.6} = 37.54$

Dado que la zona de aceptación se encuentra comprendida entre −1.96 y 1.96, se rechazan todos los contrastes. Esto significa que los tres regresores son individualmente significativos a la hora de explicar la variabilidad de la variable *SUELDO*, o lo que es lo mismo, tienen capacidad explicativa de la misma.

(b) El contraste de significación global se formaliza de la siguiente manera, en función de la notación utilizada en el apartado anterior:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

El estadístico de prueba puede tener varias formas, y una de ellas es la que está en función del coeficiente de determinación y que se corresponde con la expresión siguiente, siendo k el número de regresores del modelo y N el número de observaciones:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)}$$

Utilizando los datos del Cuadro 2.6, es inmediato obtener el valor del estadístico de prueba.

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)} = \frac{\dfrac{0.7736}{3-1}}{\dfrac{1-0.7736}{500-3}} = 849.31$$

(c) En la salida del Cuadro 2.6 no figura el estadístico de contraste, pero sí su probabilidad asociada, que es igual a 0.00. Por tanto, se rechaza la hipótesis nula a cualquier nivel de significación estándar, lo cual implica que conjuntamente las variables *ANTIG* e *INTELIG* tienen capacidad explicativa de la variable *SUELDO*.

(d) En este apartado se pide la predicción del sueldo mensual de una persona que tiene 24 meses de antigüedad en la empresa y un coeficiente de inteligencia de 0.82. Tal y como se muestra a continuación, el sueldo esperado para este individuo es igual a 1389.70 euros mensuales.

$$\widehat{Sueldo}_{501} = 448.6 + 11.51 \cdot 24 + 810.8 \cdot 0.82 = 1389.70 \text{ €}$$

EJERCICIO 2.20

Se ha estimado el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

obteniéndose el siguiente resultado, donde figuran, entre paréntesis, los ratios *t*:

$$\hat{Y}_i = 3.02 + \underset{(12.64)}{0.815} X_{2i} + \underset{(5.12)}{0.33} X_{3i} \qquad \hat{\sigma}_u^2 = 0.83 \qquad N = 20$$

Además, sabemos que:

$$\sum_{i=1}^{20} X_{2i} X_{3i} = 0; \qquad \bar{X}_2 = \bar{X}_3 = 0$$

Si imponemos la restricción $\beta_2 + \beta_3 = 1$, se obtiene que la varianza estimada de la perturbación aleatoria del modelo restringido es $\hat{\sigma}_u^2 = 0.9$.

Contraste la existencia de evidencia empírica a favor de dicha restricción por los dos métodos conocidos.

Solución

Utilizando la suma de cuadrados de los errores de los modelos restringido y no restringido, el estadístico de prueba para el siguiente contraste

$$\left.\begin{aligned} H_0&: \beta_2 + \beta_3 = 1 \\ H_1&: \beta_2 + \beta_3 \neq 1 \end{aligned}\right\}$$

es:

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)} = \frac{(16.2 - 14.11)/1}{14.11/17} = 2.52$$

El valor crítico correspondiente a una $F_{1,17}$ al 95% es 4.45. En consecuencia, no podemos rechazar la hipótesis nula. Por tanto, existe evidencia empírica a favor del cumplimiento de dicha hipótesis.

Otra opción para realizar el mismo contraste consiste en utilizar la expresión general del contraste de combinación lineal de parámetros:

$$F = \frac{\left(\left(R\hat{\beta} - r\right)'\left(R\left(X'X\right)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right)\right)/q}{\hat{\sigma}_u^2}$$

La restricción expresada en términos matriciales es:

$$R\beta = r \quad \Rightarrow \quad H_0 : \beta_2 + \beta_3 = 1$$

con

$$R = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} 3.020 \\ 0.815 \\ 0.330 \end{pmatrix} \quad r = 1$$

Sabemos que:

$$X'X = \begin{pmatrix} 20 & 0 & 0 \\ 0 & \sum_{i=1}^{N} X_{2i}^2 & 0 \\ 0 & 0 & \sum_{i=1}^{N} X_{3i}^2 \end{pmatrix}$$

Los elementos de la diagonal de la matriz $X'X$ los podemos obtener conociendo los ratios t-Student y los coeficientes estimados. De esta forma calculamos las varianzas de los estimadores de los parámetros de posición mediante:

$$t=\frac{\hat{\beta}_j}{S\left(\hat{\beta}_j\right)}\Rightarrow S\left(\hat{\beta}_j\right)=\frac{\hat{\beta}_j}{t}\Rightarrow\begin{cases}S^2\left(\hat{\beta}_2\right)=\left(\dfrac{0.815}{12.64}\right)^2=0.064^2\\ S^2\left(\hat{\beta}_3\right)=\left(\dfrac{0.33}{5.12}\right)^2=0.064^2\end{cases}$$

Por tanto, conociendo la matriz de varianzas y covarianzas de los $\hat{\beta}_j$, y sabiendo que $\hat{\sigma}_u^2=0.83$:

$$\hat{V}\left(\hat{\beta}\right)=\hat{\sigma}_u^2\left(X'X\right)^{-1}=\begin{pmatrix}\dfrac{\hat{\sigma}_u^2}{N} & 0 & 0\\ 0 & \hat{\sigma}_u^2\Big/\sum_{i=1}^{N}X_{2i}^2 & 0\\ 0 & 0 & \hat{\sigma}_u^2\Big/\sum_{i=1}^{N}X_{3i}^2\end{pmatrix}=$$

$$=\begin{pmatrix}\dfrac{0.83}{20} & 0 & 0\\ 0 & 0.064^2 & 0\\ 0 & 0 & 0.064^2\end{pmatrix}$$

Igualando términos y despejando obtenemos:

$$\frac{\hat{\sigma}_u^2}{\sum_{i=1}^{N}X_{2i}^2}=0.064^2\Rightarrow\sum_{i=1}^{N}X_{2i}^2=\frac{\hat{\sigma}_u^2}{0.064^2}\Rightarrow\sum_{i=1}^{N}X_{2i}^2=\frac{0.83}{0.064^2}=202.637$$

$$\frac{\hat{\sigma}_u^2}{\sum_{i=1}^{N}X_{3i}^2}=0.064^2\Rightarrow\sum_{i=1}^{N}X_{3i}^2=\frac{\hat{\sigma}_u^2}{0.064^2}\Rightarrow\sum_{i=1}^{N}X_{3i}^2=\frac{0.83}{0.064^2}=202.637$$

Luego la matriz $\left(X'X\right)^{-1}$ será:

$$(X'X)^{-1} = \begin{pmatrix} 1/20 & 0 & 0 \\ 0 & 1/\sum_{i=1}^{N} X_{2i}^2 & 0 \\ 0 & 0 & 1/\sum_{i=1}^{N} X_{3i}^2 \end{pmatrix} = \begin{pmatrix} 0.05 & 0 & 0 \\ 0 & 4.93\cdot 10^{-3} & 0 \\ 0 & 0 & 4.93\cdot 10^{-3} \end{pmatrix}$$

Ahora ya disponemos de todos los datos para calcular el estadístico de prueba.

$$F = \frac{\left(\left(R\hat{\beta}-r\right)'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)\right)/q}{\hat{\sigma}_u^2} =$$

$$= \frac{1}{0.83}\left[\left(\left((0 \quad 1 \quad 1)\begin{pmatrix} 3.020 \\ 0.815 \\ 0.330 \end{pmatrix} - 1\right)'\left[(0 \quad 1 \quad 1)\begin{pmatrix} 0.05 & 0 & 0 \\ 0 & 4.93\cdot 10^{-3} & 0 \\ 0 & 0 & 4.93\cdot 10^{-3} \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right]^{-1}\right.$$

$$\left.\left((0 \quad 1 \quad 1)\begin{pmatrix} 3.020 \\ 0.815 \\ 0.330 \end{pmatrix} - 1\right)\right)/1\Bigg] = 2.52$$

Como vemos, el resultado coincide con la primera solución.

EJERCICIO 2.21

La estimación de un modelo con datos centrados correspondiente a 100 observaciones proporciona los siguientes resultados:

$$y_i = x_{2i} + 6.6x_{3i}$$

obtenidos a partir de la matriz

$$(x'x)^{-1} = \begin{pmatrix} 0.033 & 0 \\ 0 & 0.333 \end{pmatrix}.$$

Se sabe además que $\hat{\sigma}_u^2 = 0.0103$.

Obtenga un intervalo de confianza del 95% para la predicción que ofrece el modelo del valor individual de la endógena correspondiente a los valores de las variables explicativas $x_2 = 0.13$ y $x_3 = 0.25$.

Solución

El valor predicho se obtiene a partir de

$$\hat{y}_{101} = x'_{101}\hat{\beta} = \begin{pmatrix} 0.13 & 0.025 \end{pmatrix}\begin{pmatrix} 1 \\ 6.6 \end{pmatrix} = 0.295$$

y el intervalo de confianza se obtiene sin más que sustituir en

$$\hat{y}_{101} \pm t^{\alpha/2}_{N-k}\hat{\sigma}_{e_{101}}$$

donde $\hat{\sigma}_{e_{101}} = \hat{\sigma}_u\sqrt{x'_{101}\left(x'x\right)^{-1}x_{101}+1}$, valor que calculamos a continuación:

$$x'_{101}\left(x'x\right)^{-1}x_{101} = \begin{pmatrix} 0.13 & 0.025 \end{pmatrix}\begin{pmatrix} 0.033 & 0 \\ 0 & 0.333 \end{pmatrix}\begin{pmatrix} 0.130 \\ 0.025 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.00429 & 0.00832 \end{pmatrix}\begin{pmatrix} 0.130 \\ 0.025 \end{pmatrix} = 0.0007658$$

$$\Rightarrow \hat{\sigma}_{e_{101}} = \sqrt{0.0103}\cdot\sqrt{0.0007658+1} = 0.1015$$

El intervalo pedido se corresponde con:

$$IC = \hat{y}_{101} \pm t^{\alpha/2}_{N-k}\hat{\sigma}_{e_{101}} = 0.295 \pm 2\cdot 0.1015 \Rightarrow \begin{pmatrix} 0.092 & 0.498 \end{pmatrix}$$

EJERCICIO 2.22

Sean dos individuos A y B que se han quedado fuera de la muestra utilizada para estimar el siguiente modelo:

$$\widehat{LSALARIO}_i = 8.128619 + 0.023524\cdot EDAD_i -$$
$$- 0.0171777\cdot EXPER_i + 0.081381\cdot EDUCA_i$$

El individuo A tiene 30 años de edad, 10 años de experiencia en la empresa y tiene 12 años de escolarización completados; mientras que el individuo B tiene 59 años, lleva 23 años en la empresa y tiene 16 años de escolarización completados. ¿Qué salario predice el modelo estimado para estos dos trabajadores?

Solución

El salario predicho para el trabajador A es:

$$lsalario_A = 8.128619 + 0.023524\cdot 30 - 0.0171777\cdot 10 + 0.081381\cdot 12 = 9.639141$$

$$\Rightarrow Salario_A = e^{9.639141} = 15354.15\ u.m.$$

Por su parte, el salario predicho para el trabajador B es:

$$lsalario_A = 8.128619 + 0.023524 \cdot 59 - 0.0171777 \cdot 23 + 0.081381 \cdot 16 = 10.42356$$

$$\Rightarrow Salario_B = e^{10.42356} = 33642.45 \ u.m.$$

EJERCICIO 2.23

Estudie la capacidad predictiva del modelo del Ejercicio 2.22 a través del cálculo del error cuadrático medio, sabiendo que el individuo A obtiene un salario real de 15000 u.m. y el B de 33000 u.m.

Solución

Llamamos error cuadrático medio a la media de los errores al cuadrado. Por tanto, en este caso tomará el siguiente valor:

$$ECM = \frac{(15000 - 15354.04)^2 + (33000 - 33642.44)^2}{2} = 269042.753$$

EJERCICIO 2.24

Obtenga la predicción puntual de la variable X para cada uno de los individuos (i, j, k, l, m) a partir de la siguiente predicción por intervalos:

$$P(-228.33 \le X_i \le -111.77) = 0.95$$

$$P(-63.92 \le X_j \le 54.69) = 0.95$$

$$P(112.68 \le X_k \le 234.22) = 0.90$$

$$P(56.22 \le X_l \le 174.49) = 0.99$$

$$P(30.52 \le X_m \le 149.68) = 0.99$$

Solución

La predicción por intervalos se calcula a partir de la siguiente expresión:

$$IC = \left(\hat{Y}_j \pm t^{\alpha/2}_{N-k}\hat{\sigma}_e\right) \quad j = \text{individuo extramuestral}$$

por lo que el predictor puntual coincidirá siempre con el punto medio del intervalo, sea cual sea el nivel de significación.

Así pues, los valores predichos para estos intervalos son:

Tabla 2.3

Extremo inferior	Extremo superior	Predicción puntual
−228.33	−111.77	−170.05
−63.92	54.69	−4.62
112.68	234.22	173.45
56.22	174.49	115.36
30.52	149.68	90.10

EJERCICIO 2.25

Se ha estimado el siguiente modelo con datos centrados:

$$y_i = \underset{(6.794116)}{0.635794}\, x_{2i} + \underset{(7.881834)}{0.713270}\, x_{3i} \qquad \overline{Y} = 11.94 \quad N = 20$$

Entre paréntesis se presenta el estadístico t para el contraste de significación individual. Obtenga la predicción por intervalo del valor medio de Y_{N+1} en el caso de que $x_{2,N+1} = 2$ y $x_{3,N+1} = 2$, sabiendo que $Cov\left(\hat{\beta}_2, \hat{\beta}_3\right) = 0.000706$.

Solución

El valor pedido se obtiene como:

$$\hat{y}_{N+1} \pm t_{17}^{\alpha/2} \hat{\sigma}_{e_{N+1}}^{(m)}$$

La predicción puntual centrada se calcula como

$$\hat{y}_{N+1} = 2 \cdot (0.635794 + 0.713270) = 2.698128$$

Por otra parte,

$$\hat{\sigma}_{e_{N+1}}^{(m)} = x'_{N+1} \hat{\sigma}_u^2 \left(x'x\right)^{-1} x_{N+1} \text{ donde } x'_{N+1} = \begin{pmatrix} 2 & 2 \end{pmatrix}$$

Para su cálculo necesitamos,

$$\hat{\sigma}_u^2 \left(x'x\right)^{-1} = \begin{pmatrix} \left(\dfrac{0.635794}{6.794116}\right)^2 & 0.000706 \\ & \left(\dfrac{0.713270}{7.881834}\right)^2 \end{pmatrix} = \begin{pmatrix} 0.0087572 & 0.0007060 \\ & 0.0081894 \end{pmatrix}$$

y

$$x'_{N+1} \hat{\sigma}_u^2 \left(x'x\right)^{-1} = \begin{pmatrix} 0.0189265 & 0.0177908 \end{pmatrix}$$

$$\Rightarrow x'_{N+1}\hat{\sigma}_u^2\left(x'x\right)^{-1}x_{N+1} = 0.0734346$$

Es decir,

$$\hat{\sigma}_{e_{N+1}}^{(m)} = \sqrt{0.0734346} = 0.2709882$$

Finalmente, para un nivel de significación del 5%:

$$\hat{y}_{N+1} \pm t_{17}^{5\%}\hat{\sigma}_{e_{N+1}}^{(m)} = 2.698128 \pm 2.11 \cdot 0.2709882 =$$

$$= \begin{cases} (2.1263428 \quad 3.2699132) & \text{para } y_i \\ (14.0663428 \quad 15.2099132) & \text{para } Y_i \end{cases}$$

EJERCICIO 2.26

Se ha recogido información del gasto de personal (Y), del número de personas que componen la plantilla (X_1) y del número medio de horas extras que se realizan a la semana (X_2) y se ha estimado el gasto en función de estos factores, resultando:

$$\hat{Y}_i = 303.5 + 2.3X_{1i} - 25.07X_{2i} + 0.2X_{1i}X_{2i}$$

Se sabe además que

$$\hat{V}(\hat{\beta}) = \begin{pmatrix} 1022.379 & -0.622 & -14.458 & 0.0900 \\ -0.622 & 0.005 & 0.092 & -0.0010 \\ -14.458 & 0.092 & 2.638 & -0.0160 \\ 0.090 & -0.001 & -0.016 & 0.0001 \end{pmatrix}$$

(a) Con la información que se facilita en la Tabla 2.4, y sabiendo que el coeficiente de determinación es 0.963, realice un contraste de significación conjunta de los coeficientes

Tabla 2.4

Fuente de la variación explicada por:	Grados de libertad	Sumas de cuadrados
El modelo $\hat{Y}$	3	¿?
El error residual e	20-4	¿?
Total	20-1	2254327

(b) Obtenga una predicción puntual y por intervalos del gasto total para el caso en el que el tamaño de la plantilla sea de 250 y el número medio de horas extras semanales sea de 4.

Solución

(a) Como el modelo tiene constante,

$$R^2 = 0.963 = \frac{SCR}{SCT} = \frac{SCR}{2254327} \Rightarrow SCR = 0.963 \cdot 2254327 = 2170916.9$$

Luego, los resultados que faltan en la tabla serán $SCR = 2170916.9$ y SCE que se obtendrá de la siguiente diferencia

$$SCE = SCT - SCR = 2254327 - 2170916.9 = 83410.1$$

Para obtener el valor del estadístico de contraste F necesitamos las medias de las Sumas de Cuadrados, que se obtienen dividiendo SCE y SCR entre sus correspondientes grados de libertad.

Tabla 2.5

Fuente de la variación explicada por:	Grados de libertad	Sumas de cuadrados	Media de los cuadrados
El modelo $\hat{Y}$	3	2170916.9	2170916.9 / 3 = 723638.9
El error residual e	20-4	83410.1	83410.1 / 16 = 5213.1
Total	20-1	2254327	

El valor del estadístico de contraste y las hipótesis nula y alternativa vienen dados por

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: H_0 \text{ no se cumple} \end{array}\right\} \Rightarrow F = \frac{723638.9}{5213.1} = 138.81$$

Si lo comparamos con el valor en las tablas de $F_{3,16}^{0.05} = 3.24$, nos lleva a rechazar la hipótesis nula, lo cual significa que alguna/s o todas las variables explicativas contribuyen a explicar la variabilidad de la endógena.

(b) Para obtener la predicción puntual calculamos

$$\hat{Y}_{N+1} = X'_{N+1}\hat{\beta} = \begin{pmatrix} 1 & 250 & 4 & 1000 \end{pmatrix} \begin{pmatrix} 303.50 \\ 2.30 \\ -25.07 \\ 0.20 \end{pmatrix} = 978.22$$

Y para obtener la predicción por intervalos, como $e'e = 83410.1$, podemos calcular $\hat{\sigma}_u^2$.

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{83410.1}{20-4} = 5213.1$$

Éste es el valor que encontramos en la tabla anterior como media de cuadrados de los residuos.

El valor del estimador de la varianza de las perturbaciones lo necesitamos para, a partir de la matriz de varianzas y covarianzas de los estimadores de los parámetros, obtener la matriz $X'X$ sin más que dividir todos sus elementos entre el valor de $\hat{\sigma}_u^2$. De ahí que dicha matriz venga dada por

$$(X'X)^{-1} = \begin{pmatrix} 0.196116 & -0.0001190 & -0.002773 & 0.00001700 \\ -0.000119 & 0.0000010 & 0.000018 & -0.00000010 \\ -0.002773 & 0.0000180 & 0.000506 & -0.00000300 \\ 0.000017 & -0.0000001 & -0.000003 & 0.00000002 \end{pmatrix}$$

El intervalo de confianza se obtiene sustituyendo en $IC = \hat{Y}_{N+1} \pm t_{17}^{\alpha/2}\hat{\sigma}_e$ y para ello hemos de calcular

$$\hat{\sigma}_{e_{N+1}} = \hat{\sigma}_u \sqrt{X'_{N+1}(X'X)^{-1}X_{N+1} + 1} = 77.4567892\,.$$

Sustituyendo, tendremos el intervalo de confianza requerido:

$$IC = \hat{Y}_{N+1} \pm t_{17}^{\alpha/2}\hat{\sigma}_{e_{N+1}} = 978.22 \pm 2 \cdot 77.4567892 \Rightarrow (823.31 \quad 1133.13)$$

EJERCICIO 2.27

Con una muestra aleatoria simple se obtienen las matrices que se muestran a continuación:

$$(X'X)^{-1} = \begin{pmatrix} 2 & 4 & 5 \\ a & 5 & 7 \\ 5 & b & 1 \end{pmatrix} \quad X'Y = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

Sabiendo, además, que la primera columna de la matriz X es una columna de unos, si un individuo toma valor cero para todas las variables exógenas del modelo de regresión lineal múltiple, ¿qué valor predeciría para dicho individuo? Justifique su respuesta.

Solución

El valor predicho para el individuo i se obtiene como $X_i'\hat{\beta}$, siendo X_i el vector de valores de las exógenas del individuo i. Dadas las matrices que nos da el enunciado, el modelo lo podemos escribir como,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Por tanto, el valor predicho para el individuo que tiene como matriz $X_i' = (1 \quad 0 \quad 0)$ vendrá dado por el estimador de β_1, tal y como se demuestra a continuación:

$$X_i'\hat{\beta} = (1 \quad 0 \quad 0)\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \hat{\beta}_1$$

A partir de las matrices que nos da el enunciado es inmediato obtener el valor de la estimación del parámetro β_1:

$$\hat{\beta}_1 = 2 \cdot 1 + 4 \cdot 3 + 5 \cdot 5 = 39$$

EJERCICIO 2.28

Se dispone de dos modelos, uno para la variable renta y otro para la variable número de coches matriculados. Se ha estudiado la capacidad predictiva para ambos modelos y se ha obtenido un Error Absoluto Medio (EAM) de 23 para el modelo de la variable renta y de 12 para el modelo correspondiente al número de coches matriculados.

(a) ¿Cuál de los dos modelos presenta una mejor capacidad predictiva? Justifique su respuesta.

(b) Si los datos del enunciado se correspondiesen con la medida Error Absoluto Medio en Porcentaje de Media (*PMEA*), ¿modificaría su respuesta? ¿Por qué?

Solución

(a) El estadístico que se nos proporciona para analizar la capacidad predictiva no permite la comparación entre modelos que tengan distinta variable endógena. Por tanto, simplemente con ellos no podemos contestar a la pregunta.

(b) En este caso sí se modificaría la respuesta. La razón es que el error absoluto medio en porcentaje de media es una medida relativa de la capacidad predictiva. No tiene unidades y lo que mide es el error medio que se comete en la predicción por cada 100 unidades de la variable endógena. Como se puede deducir es adimensional, tal y como se observa en la siguiente expresión:

$$PMEA = \frac{\sum_{j=1}^{h} \left| e_j / y_j \right|}{h} \cdot 100$$

por lo que el hecho de que las variables no se midan en las mismas unidades no afecta a la medida. En este caso, el modelo correspondiente al número de coches presenta una mejor capacidad predictiva entre ambos modelos, aunque tampoco se puede calificar de buena.

EJERCICIO 2.29

En la Tabla 2.6 se muestran los valores de la renta para cinco individuos pertenecientes a una población para la cual se ha estimado un modelo para su renta con el siguiente resultado:

$$\widehat{RENTA}_i = 3 + 2 \cdot SEXO_i + 10 \cdot ANTIGÜEDAD_i$$

siendo *SEXO* una variable que toma el valor 1 si el individuo es varón y 0 en caso contrario y *ANTIGÜEDAD* una variable que mide los meses de antigüedad en el trabajo.

Además, la Tabla 2.6 recoge el valor de ambas variables explicativas para cada uno de los cinco individuos.

Tabla 2.6

Renta	Sexo	Antigüedad
200	Varón	15
220	Varón	20
190	Hembra	14
150	Hembra	12
210	Hembra	19

Analice la capacidad predictiva de modelo mediante una medida relativa. Interprete los resultados.

Solución

La medida relativa que tenemos que calcular es el Error Absoluto Medio en Porcentaje de Media. Su fórmula es la que se muestra a continuación:

$$PMEA = \frac{\sum_{j=1}^{h} \left| e_j / y_j \right|}{h} \cdot 100$$

El primer paso consiste en predecir la renta de cada uno de estos cinco individuos; en segundo lugar, obtener su error, para, en tercer lugar, dividir cada uno de estos errores por el valor real de la renta y el valor absoluto de este cociente multiplicarlo por 100. El valor del *PMEA* se obtiene como media de estos

últimos valores. En la Tabla 2.7 se muestran todos los cálculos necesarios para obtener esta medida de la capacidad predictiva del modelo.

Tabla 2.7

Renta	Sexo	Antigüedad	Renta Predicha	Error	|Error/renta|*100
200	1	15	155	45	22.50
220	1	20	205	15	6.82
190	0	14	143	47	24.74
150	0	12	123	27	18.00
210	0	19	193	17	8.10
Error Absoluto Medio en Porcentaje de Media: 16.03					

Los resultados indican que el modelo predice con un error medio de 16 unidades por cada 100 unidades de media. En consecuencia podemos decir que el modelo tiene una mala capacidad predictiva.

EJERCICIO 2.30

Un individuo desea invertir 18.000 euros en la bolsa. Sin embargo, tiene dudas sobre si hacerlo en acciones de la empresa A o en acciones de la empresa B. En principio, preferirá aquella empresa en la que espera obtener un rendimiento por euro invertido más alto y con una mayor seguridad (la seguridad de la inversión se mide en términos de la varianza del error de predicción).

El individuo cree que la rentabilidad por euro de las acciones de cada empresa en un momento dado dependerá de dos variables: el volumen de beneficios reales obtenidos por la misma durante ese período y el volumen de activos medios mantenido en ese mismo período. Por ello, se han estimado los siguientes modelos:

$$\hat{Y}_t^{(A)} = 1.808845 + 1.937910 X_{2t}^{(A)} + 1.024075 X_{3t}^{(A)} \quad t = 1,...,T \qquad (2.1)$$

$$\hat{Y}_t^{(B)} = 2.00178 + 1.54558 X_{2t}^{(B)} + 1.09319 X_{3t}^{(B)} \quad t = 1,...,T \qquad (2.2)$$

donde:

$Y_t^{(i)}$ son los rendimientos por cada 100 euros invertidos en acciones de la empresa i ($i = A, B$), en el período t.

$X_{2t}^{(i)}$ son los beneficios reales de la empresa i ($i = A, B$), en el período t, en millones de euros.

$X_{3t}^{(i)}$ es el volumen de activos de la empresa i ($i = A, B$), en el período t, en millones de euros.

$u_t^{(i)}$ es la perturbación aleatoria del modelo correspondiente, ($i = A, B$).

En la estimación se han usado datos mensuales, generados durante los últimos 20 años, que cumplen:

$$X'Y^{(A)} = \begin{pmatrix} 270 \\ 295 \\ 1300 \end{pmatrix} \qquad X'Y^{(B)} = \begin{pmatrix} 250 \\ 280 \\ 1200 \end{pmatrix}$$

(a) Estime la varianza de las perturbaciones de los modelos (2.1) y (2.2), sabiendo que: $\sum Y_t^{2(A)} = 3000$ y que $\sum Y_t^{2(B)} = 2300$.

(b) Suponiendo que el individuo conoce los siguientes datos del período $T+1$:

$$X_{2,T+1}^{(A)} = 2 \quad X_{3,T+1}^{(A)} = 5 \quad X_{2,T+1}^{(B)} = 2.5 \quad X_{3,T+1}^{(B)} = 4.5$$

y que dispone de la siguiente información, ¿en cuál de las dos empresas decidirá invertir? (En el resultado de las predicciones de los rendimientos despreciar la parte decimal, sólo considerar la parte entera).

$$X_{T+1}^{\prime(A)}(X'X)^{-1} \quad X_{T+1}^{(A)} = 0.107079 \qquad X_{T+1}^{\prime(B)}(X'X)^{-1} \quad X_{T+1}^{(B)} = 0.140112$$

(c) Obtenga los intervalos de predicción para cada una de las empresas.

Solución

(a) Sabiendo que

$$\hat{\sigma}_u^2 = \frac{Y'Y - \hat{\beta}'X'Y}{N-k}$$

es inmediato obtener la estimación de la varianza de la perturbación aleatoria para los modelos de ambas empresas.

Empresa A

$$\hat{\sigma}_u^2 = \frac{3000 - \begin{pmatrix} 1.808845 & 1.937910 & 1.024075 \end{pmatrix} \begin{pmatrix} 270 \\ 295 \\ 1300 \end{pmatrix}}{20-3} = 35.80182$$

Empresa B

$$\hat{\sigma}_u^2 = \frac{2300 - \begin{pmatrix} 2.00178 & 1.54558 & 1.09319 \end{pmatrix} \begin{pmatrix} 250 \\ 280 \\ 1200 \end{pmatrix}}{20-3} = 3.23321$$

(b) Cálculos previos:

$$\hat{Y}_t^{(A)} = 1.808845 + 1.937910 \cdot 2 + 1.024075 \cdot 5 = 10.81$$

$$\hat{Y}_t^{(B)} = 2.00178 + 1.54558 \cdot 2.5 + 1.09319 \cdot 4.5 = 10.79$$

Si se desprecia la parte decimal, y sólo se considera la parte entera, ambas empresas predicen una misma rentabilidad (10%). Por tanto, sólo queda como criterio para tomar una decisión el riesgo de invertir en cada una de las empresas.

El riesgo se valora a partir de la varianza del error de predicción. Éste se calcula como:

$$\widehat{Var}(e_{T+1}) = \hat{\sigma}_u^2 \left[1 + X'_{T+1} (X'X)^{-1} X_{T+1} \right]$$

Para la empresa A

$$\widehat{Var}\left(e_{T+1}^{(A)}\right) = 35.80182(1 + 0.107079) = 39.63544$$

Para la empresa B

$$\widehat{Var}\left(e_{T+1}^{(B)}\right) = 3.23324(1 + 0.140112) = 3.686222$$

La varianza del error de predicción de la empresa A es mayor que la correspondiente a la de la empresa B, por tanto, el individuo invertirá en la empresa B al soportar un menor riesgo (menor varianza).

(c) La expresión para la predicción por intervalos es:

$$\hat{Y}_{T+1} \pm t_{N-k}^{\alpha/2} \hat{\sigma}_{e_{T+1}}$$

Por tanto, los intervalos de predicción son:

Para la empresa A: $IC = 10.81 \pm 2.1098 \cdot 6.2957 \Rightarrow (-2.48 \quad 24.09)$

Para la empresa B: $IC = 10.79 \pm 2.1098 \cdot 1.92 \Rightarrow (6.73 \quad 14.84)$

EJERCICIO 2.31

Una firma tiene un total de 20 industrias establecidas en un mercado único. Los datos de beneficios, costes de la materia prima y número de trabajadores se muestran en la Tabla 2.8. La multinacional está pensando en montar una industria nueva en otra ciudad y, dada la población y la disponibilidad de las materias primas, considera que la nueva industria debe tener 3 trabajadores con un coste de la materia prima de 2500 euros. Dados los datos del resto de industrias, ¿cuáles son los beneficios que esperaría se produjesen con la nueva industria?

Tabla 2.8

Beneficios (€)	C. Mat. Prima (€)	Nº trabajadores
2233	7256	1
2698	8822	6
887	2759	5
1530	4947	8
2365	7726	2
2423	7945	5
563	1465	8
1349	4222	6
1262	4025	1
274	643	4
218	358	1
2519	8082	7
1564	5065	7
1953	6215	5
2006	6265	7
1870	5986	4
1194	3603	5
2861	9102	8
1123	3379	1
1514	4871	7

Solución

Si denotamos por B_i a los beneficios de la empresa i, por C al coste de la materia prima, por T al número de trabajadores y suponiendo que se cumplen todas las hipótesis básicas del modelo de regresión lineal múltiple, el modelo econométrico que representa al beneficio se puede escribir como:

$$B_i = \beta_0 + \beta_1 C_i + \beta_2 T_i + u_i$$

El ejercicio nos pide el valor predicho para el beneficio de una empresa ajena a la muestra que tuviera unos costes de 2500 euros con 3 trabajadores. Es decir, conocemos el vector de valores de las explicativas, que es igual a

$$X_{N+1} = \begin{pmatrix} 1 \\ 2500 \\ 3 \end{pmatrix}.$$

Si denotamos por $\hat{B}_{N+1}$ al beneficio individual predicho, sabemos por teoría que coincide con el beneficio predicho medio y que es igual a $X'_{N+1}\hat{\beta}$, siendo $\hat{\beta}$ el vector de parámetros estimados del modelo anterior. En consecuencia, el primer paso es estimar dicho modelo.

Las distintas matrices se muestran a continuación, junto con el modelo estimado.

$$X'X = \begin{pmatrix} 20 & 102736 & 98 \\ & 657743848 & 531953 \\ & & 600 \end{pmatrix} \Rightarrow$$

$$\Rightarrow \quad (X'X)^{-1} = \begin{pmatrix} 0.37828767 & -3.22156 \cdot 10^{-5} & -0.03322499 \\ & 8.1164 \cdot 10^{-9} & -1.934 \cdot 10^{-6} \\ & & 0.00880809 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 32406 \\ 205088959 \\ 167562 \end{pmatrix} \quad \Rightarrow \quad \hat{\beta} = \begin{pmatrix} 84.4755300 \\ 0.2965360 \\ 2.5667685 \end{pmatrix}$$

$$\Rightarrow \quad \hat{B}_i = 84.48 + 0.297C_i + 2.567T_i$$

Teniendo en cuenta el valor de X_{N+1} y los parámetros estimados, el predictor individual y medio es igual a

$$\hat{B}_{N+1} = X'_{N+1}\hat{\beta} = \begin{pmatrix} 1 & 2500 & 3 \end{pmatrix} \begin{pmatrix} 84.480 \\ 0.297 \\ 2.567 \end{pmatrix} = 833.516 \text{ euros}$$

EJERCICIO 2.32

Con los datos del ejercicio 2.31 y para un nivel de confianza del 95%, prediga qué intervalo contiene al verdadero valor, tanto individual como medio.

Solución

El intervalo pedido para el caso individual es

$$IC = \left(\hat{B}_j \pm t_{17}^{0.025} \hat{\sigma}_u \sqrt{X_j' \left(X'X \right)^{-1} X_j + 1} \right)$$

y para el caso medio

$$IC = \left(\hat{B}_j \pm t_{17}^{0.025} \hat{\sigma}_u \sqrt{X_j' \left(X'X \right)^{-1} X_j} \right),$$

siendo $t_{17}^{0.025} = 2.11$ y

$$\hat{\sigma}_u = \sqrt{\frac{Y'Y - \hat{\beta}'X'Y}{N-k}}$$

Con los datos del enunciado es inmediato calcular $Y'Y$, cuyo valor es 64000214, y $\hat{\beta}'X'Y = 63983913.53$, con lo cual

$$\hat{\sigma}_u = \sqrt{\frac{64000214 - 6398913.53}{20-3}} = 30.96$$

Teniendo en cuenta que

$$X_j = \begin{pmatrix} 1 \\ 2500 \\ 3 \end{pmatrix}$$

y conociendo la matriz $\left(X'X \right)^{-1}$ obtenida en el ejercicio 2.31, es inmediato obtener los siguientes intervalos:

Para la predicción individual:

$$IC = \left(833.519 \pm 2.11 \cdot 30.96 \sqrt{1.11884966} \right) \Rightarrow \left(764.4 \quad 902.6 \right)$$

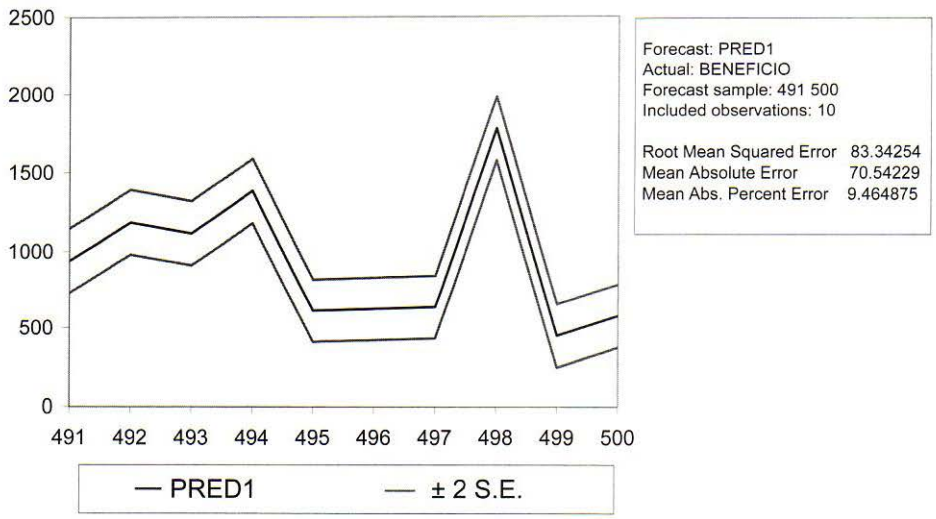
Para la predicción en media:

$$IC = \left(833.519 \pm 2.11 \cdot 30.96 \sqrt{0.11884966} \right) \Rightarrow \left(811 \quad 856 \right)$$

EJERCICIO 2.33

El Gráfico 2.1 se corresponde con una salida del programa Eviews.

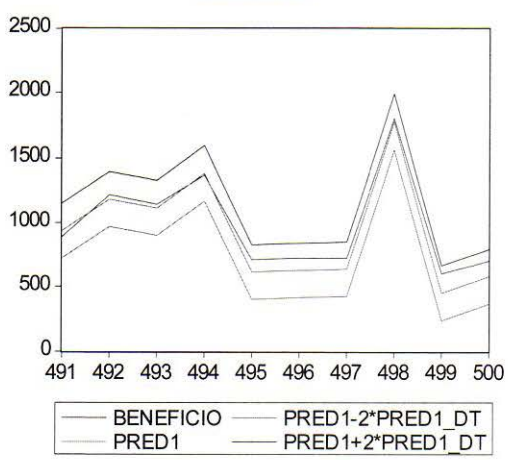
Gráfico 2.1



(a) Interprete toda la información que se encuentra en dicha salida.

(b) ¿Se corresponde el Gráfico 2.2 con la salida anterior? Interprete el gráfico.

Gráfico 2.2



(c) Los datos de *BENEFICIO* y *PRED*1 (con dos decimales) se muestran en la Tabla 2.9. Calcule el Error Absoluto Medio en Porcentaje de Media.

Tabla 2.9

Individuo	BENEFICIO	PRED1
491	893.02	939.84
492	1219.97	1184.87
493	1144.31	1114.86
494	1369.98	1383.24
495	711.49	614.87
496	725.70	626.54
497	723.26	638.21
498	1805.18	1778.87
499	603.36	451.51
500	701.69	579.87

Solución

(a) El Gráfico 2.1 se corresponde con una salida de predicción del programa Eviews. En concreto se predicen los datos correspondientes a la variable *BENEFICIO* para los individuos que se encuentran en la base de datos en las posiciones 491 a la 500. Estas predicciones se almacenan en la variable *PRED*1. En el gráfico se muestran los valores de las predicciones individuales tanto de forma puntual como para un nivel de confianza del 95%. Las dos primeras medidas que figuran en el Gráfico 2.1 son la Raíz del Error Cuadrático Medio y el Error Absoluto Medio. Ambas son medidas de la capacidad predictiva, pero, dado su carácter absoluto, sólo son válidas para comparar entre modelos alternativos para una misma variable endógena. La tercera medida es el Error Absoluto Medio en Porcentaje de Media. Éste nos dice que por cada 100 unidades de beneficio el modelo comete un error de predicción de casi 9.5 unidades. En consecuencia, el modelo presenta una capacidad predictiva normal tendiendo a mala.

(b) Numéricamente el Gráfico 2.2 es congruente con el Gráfico 2.1 del apartado anterior y se deduce que la variable *PRED*1_*DT* contiene la información correspondiente a la desviación típica de la predicción. Como se puede observar, la capacidad predictiva no es muy mala puesto que al menos la predicción por intervalos contiene a los valores reales de los beneficios de las 10 empresas para las cuales se realizó la predicción.

(c) La fórmula del estadístico pedido es:

$$PMEA = \frac{\sum_{j=1}^{h} \left| e_j / y_j \right|}{h} \cdot 100$$

La Tabla 2.10 contiene todos los datos necesarios para obtener el valor de *PMEA*.

Tabla 2.10

Individuo	BENEFICIO	PRED1	error	|Error/BENEFICIO|*100
491	893.02	939.84	−46.82	5.24288370
492	1219.97	1184.87	35.10	2.87711993
493	1144.31	1114.86	29.45	2.57360331
494	1369.98	1383.24	−13.26	0.96789734
495	711.49	614.87	96.62	13.57995190
496	725.70	626.54	99.16	13.66404850
497	723.26	638.21	85.05	11.75925670
498	1805.18	1778.87	26.31	1.45747239
499	603.36	451.51	151.85	25.16739590
500	701.69	579.87	121.82	17.36094290
Error Absoluto Medio en Porcentaje de Media=				9.46505726

Como se puede observar, el valor es muy similar al de la salida del Eviews. La diferencia se debe al número de decimales que usa el programa Eviews, frente a los dos decimales que se han utilizado para el cálculo de los errores.

EJERCICIO 2.34

Para un modelo con 5 parámetros, y disponiendo de una muestra de 20 datos, se ha realizado la predicción por intervalos individual y media para el individuo i de una población para un nivel de confianza del 95%. Sus resultados fueron $(2 \quad 6)$ y $(3 \quad 5)$, respectivamente.

(a) Determine cuál es la predicción individual y cuál en media.

(b) Obtenga el valor de la expresión $X_i'(X'X)^{-1}X_i$.

(c) Si se dispone de las tablas de la distribución t de Student, ¿cuál es la estimación de la varianza de la perturbación aleatoria?

Solución

(a) El error de predicción individual tiene una varianza mayor que el de la predicción en media. Recordemos que para el error individual la varianza es igual a $Var(e_i) = \sigma_u^2\left(X_i'(X'X)^{-1}X_i + 1\right)$ y para la predicción media

$Var\left(e_i^{(m)}\right) = \sigma_u^2\left(X_i'(X'X)^{-1}X_i\right)$. Dado que la amplitud del intervalo de predicción depende de estos valores, es evidente que el que tenga un intervalo más pequeño se corresponde con la predicción en media y el más amplio con la predicción individual.

(b) Si tenemos en cuenta que el predictor individual puntual coincide con el predictor en media puntual y que ambos coinciden con el punto central de la predicción por intervalos, es inmediato deducir que el predictor puntual es igual a 4. Además, teniendo en cuenta que el predictor individual por intervalos se obtiene como

$$IC = \left(\hat{Y}_i \pm t_{N-k}^{\alpha/2}\hat{\sigma}_u\sqrt{X_i'(X'X)^{-1}X_i + 1}\right)$$

y que el predictor en media puntual, como

$$IC = \left(\hat{Y}_i \pm t_{N-k}^{\alpha/2}\hat{\sigma}_u\sqrt{X_i'(X'X)^{-1}X_i}\right)$$

sabemos que se tiene que cumplir:

$$\frac{t_{N-k}^{\alpha/2}\hat{\sigma}_u\sqrt{X_i'(X'X)^{-1}X_i}}{t_{N-k}^{\alpha/2}\hat{\sigma}_u\sqrt{X_i'(X'X)^{-1}X_i + 1}} = \frac{1}{2}$$

Operando con esta última expresión se concluye que $X_i'(X'X)^{-1}X_i = \frac{1}{3}$.

(c) Sabiendo que $t_{N-k}^{\alpha/2}\hat{\sigma}_u\sqrt{X_i'(X'X)^{-1}X_i} = 1$ y teniendo en cuenta que $t_{15}^{0.025} = 2.131$, valor que se obtiene de la tabla de la t de Student de 15 grados de libertad que deja a su derecha una probabilidad de 0.025, despejando $\hat{\sigma}_u$, se obtiene:

$$\hat{\sigma}_u = \frac{1}{t_{15}^{0.025}\sqrt{X_i'(X'X)^{-1}X_i}} = \frac{1}{2.131\cdot\sqrt{\frac{1}{3}}} = 0.81$$

Por tanto, el estimador de la varianza de la perturbación aleatoria es igual a 0.81^2.

EJERCICIO 2.35

Se ha estimado el siguiente modelo por mínimos cuadrados restringidos, aplicando directamente la fórmula de los estimadores Mínimos Cuadrados Restringidos (MCR):

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \text{sujeto a} \quad \beta_2 + \beta_3 = 2$$

El vector de estimadores es

$$\hat{\beta}_{MCR} = \begin{pmatrix} 6 \\ 2 \\ 0 \end{pmatrix}$$

Plantee el modelo transformado que emplearía si quisiera obtener los estimadores del modelo restringido imponiendo directamente la restricción y estimando por MCO.

Solución

En el modelo transformado sustituimos $\beta_3 = 2 - \beta_2$, obteniendo así:

$$Y_i = \beta_1 + \beta_2 X_{2i} + (2 - \beta_2) X_{3i} + u_i$$

$$Y_i - 2X_3 = \beta_1 + \beta_2 (X_{2i} - X_{3i}) + u_i$$

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i$$

Estimando por MCO el modelo anterior:

$$\hat{\beta}_{MCR} = (X'^* X^*)^{-1} X'^* Y^* = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$$

EJERCICIO 2.36

¿El estimador mínimo cuadrático restringido es sesgado o insesgado? Demuéstrelo.

Solución

El estimador mínimo cuadrático restringido es insesgado si las restricciones son ciertas. En caso contrario, es sesgado. Para demostrarlo tenemos en cuenta el siguiente desarrollo:

$$\hat{\beta}_{MCR} = \hat{\beta}_{MCO} + (X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} \left(r - R\hat{\beta}_{MCO} \right)$$

$$E\left(\hat{\beta}_{MCR}\right) = \beta + (X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} \left(r - R\hat{\beta}_{MCO} \right)$$

Cuando la restricción es cierta ocurre que $(r - R\beta) = 0$, ya que en ese caso se cumple que $R\beta = r$. Por tanto, bajo la restricción, se obtiene:

$$E\left(\hat{\beta}_{MCR}\right) = \beta$$

siendo el estimador MCR insesgado. En caso contrario, el estimador MCR sería sesgado.

EJERCICIO 2.37

El modelo $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ fue estimado por mínimos cuadrados, utilizándose 23 observaciones y obteniéndose la ecuación:

$$\hat{Y}_i = 1.5 + 3.5X_{2i} - 0.7X_{3i} \qquad R^2 = 0.982$$

El mismo modelo fue estimado bajo la restricción $\beta_3 = 0$, con el siguiente resultado:

$$\hat{Y}_i = 1.2 + 3.8X_{2i} \qquad R^2 = 0.876$$

¿Podemos decir que la restricción fue impuesta correctamente? Justifique la respuesta mediante la realización de un contraste.

Solución

Si la restricción impuesta fuese correcta, sería preferible la estimación del modelo restringido a la del modelo ampliado, pues sólo ésta permite obtener estimadores lineales, insesgados y óptimos (ELIO).

Para comprobar cuál de estas dos estimaciones es preferida, basta con hacer un contraste de hipótesis de subconjunto de parámetros, con el fin de comprobar si el regresor X_3 es o no significativo.

La hipótesis nula del contraste es H_0: $\beta_3 = 0$, y el estadístico de contraste es el siguiente:

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)}$$

Dado que el modelo presenta constante, se puede obtener el valor de la *SCE* a partir del valor del R^2 de la siguiente forma:

$$R^2 = 1 - \frac{SCE}{SCT} \Rightarrow SCE = \left(1 - R^2\right) \cdot SCT$$

Por lo que la *SCE* del modelo ampliado valdrá

$$SCE = (1 - 0.982) \cdot SCT = 0.018 \cdot SCT$$

y la del modelo restringido valdrá

$$SCE = (1 - 0.876) \cdot SCT = 0.124 \cdot SCT .$$

Sustituyendo estos valores en la expresión del estadístico F, se obtiene el valor del estadístico de contraste.

$$F = \frac{(0.124 \cdot SCT - 0.018 \cdot SCT)/1}{0.018 \cdot SCT/(23-3)} = \frac{0.106 \cdot SCT}{0.018 \cdot SCT/20} = 117.77$$

Dado que el valor del estadístico es superior al punto crítico $\left(F^{\alpha}_{m,N-k} = F^{0.05}_{1,20} = 4.35\right)$, rechazamos la hipótesis nula y, por tanto, no podemos decir que la restricción impuesta para estimar el segundo modelo haya sido correcta.

3

Problemas provocados por los datos económicos. Multicolinealidad, *outliers* y no normalidad

EJERCICIO 3.1

¿Por qué la presencia de multicolinealidad aproximada conlleva que se tienda a no rechazar la hipótesis de no significatividad de las variables que la provocan?

Solución

Esta falta de rechazo se debe a que la multicolinealidad aproximada incrementa la desviación típica de las estimaciones de los coeficientes. De esta manera, el ratio t de significación individual tiende a disminuir, toda vez que el mismo se obtiene como:

$$t = \frac{\hat{\beta}_j}{S\left(\hat{\beta}_j\right)}$$

Recordemos que la disminución de este ratio hace que se incremente la posibilidad de no rechazar la hipótesis nula de significación individual.

Si hay problemas de multicolinealidad aproximada, es porque existe una cuasi combinación lineal entre todas o algunas de las variables explicativas del modelo,

entre las que puede estar la constante. Ello implica que la matriz $X'X$ presente un determinante muy próximo a cero, lo que, a su vez, supone que los elementos de la matriz $(X'X)^{-1}$ tiendan a ser muy grandes. No olvidemos que en todos los elementos de dicha matriz inversa interviene, multiplicando, la inversa de este determinante. De esta manera, y dado que la matriz $(X'X)^{-1}$ se usa para calcular, igualmente, en la matriz de varianzas y covarianzas de las estimaciones de los coeficientes del modelo, mediante la expresión que ya conocemos $V(\hat{\beta}_j) = \hat{\sigma}_u^2 (X'X)^{-1}$, las varianzas de estas estimaciones se verán incrementadas como consecuencia de la presencia de multicolinealidad aproximada.

EJERCICIO 3.2

¿Cómo afecta a las propiedades de los estimadores obtenidos por MCO el hecho de que el modelo adolezca de multicolinealidad aproximada? ¿Y si se tratase de multicolinealidad exacta?

Solución

Cuando el modelo presenta multicolinealidad aproximada los estimadores MCO siguen siendo lineales, insesgados y óptimos (ELIO). Sin embargo, sus varianzas son más altas de lo que serían sin este problema, por lo que resultan ser más imprecisos y las estimaciones estarán limitadas para llevar a cabo análisis estructural, aunque sí serán válidos para realizar predicciones.

En cambio, si en lugar de multicolinealidad aproximada estamos hablando de un problema de multicolinealidad exacta, nos encontramos ante la presencia de una combinación lineal exacta entre todas o algunas de las variables explicativas del modelo. Esto conlleva que el determinante de la matriz $X'X$ valga cero, por lo que la matriz $X'X$ deja de ser invertible y no podemos obtener los estimadores MCO[1]. Si no podemos obtener dichos estimadores, mucho menos podremos hablar de sus propiedades.

[1] Recordemos que los estimadores MCO se obtienen mediante la expresión $\hat{\beta} = (X'X)^{-1} X'Y$.

EJERCICIO 3.3

Dada la siguiente información.

$$X'X = \begin{pmatrix} 35.000000 & 193.057941 & 772.251242 \\ 193.057941 & 1067.570520 & 4270.38961 \\ 772.251242 & 4270.389610 & 17081.9886 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 2437.90284 \\ 13469.82610 \\ 53880.66190 \end{pmatrix} \qquad Y'Y = (170002.814)$$

¿Podríamos afirmar que el modelo presenta multicolinealidad aproximada?

Solución

Las consecuencias de la multicolinealidad aproximada son:

- Imprecisión de sus estimadores, debido a la inflación de su varianza
- Falta de significación individual de los estimadores, pero modelo globalmente significativo.

Imprecisión de los estimadores

La precisión de un estimador es la mitad de la amplitud de su intervalo de confianza. Dado que el intervalo de confianza para $\hat{\beta}_j$ es $\hat{\beta}_j \pm t_{N-k}^{\alpha/2} S\left(\hat{\beta}_j\right)$, la precisión del intervalo es $t_{N-k}^{\alpha/2} S\left(\hat{\beta}_j\right)$. En nuestro caso, el modelo tiene tres parámetros de posición $(k=3)$ y N es igual a 35.

Dado que $\hat{V}\left(\hat{\beta}_j\right) = \hat{\sigma}_u^2 \left(X'X\right)^{-1}$, necesitamos calcular el estimador de la varianza de la perturbación aleatoria, $\hat{\sigma}_u^2$, y la matriz $\left(X'X\right)^{-1}$.

La inversa de $X'X$ es

$$\left(X'X\right)^{-1} = \begin{pmatrix} 11.4104 & 48.6772 & -12.6848 \\ 48.6772 & 548760.1560 & -137188.7850 \\ -12.6848 & -137188.7850 & 34296.9061 \end{pmatrix}$$

En cuanto a $\hat{\sigma}_u^2$, dado que es igual a $e'e/(N-k)$, será necesario calcular previamente el valor de $e'e$. Éste se puede obtener de la expresión $Y'Y - \hat{\beta}'X'Y$,

donde $Y'Y$ nos lo da el enunciado y $\hat{\beta}$ lo podemos calcular como $(X'X)^{-1}X'Y$.

$$\hat{\beta} = \begin{pmatrix} 11.4104 & 48.6772 & -12.6848 \\ 48.6772 & 548760.1560 & -137188.7850 \\ -12.6848 & -137188.7850 & 34296.9061 \end{pmatrix} \begin{pmatrix} 2437.90284 \\ 13469.82610 \\ 53880.66190 \end{pmatrix} = \begin{pmatrix} 23.250 \\ 6.910 \\ 0.375 \end{pmatrix}$$

De esta forma obtenemos que

$$e'e = Y'Y - \hat{\beta}'X'Y =$$

$$= 170002.814 - \begin{pmatrix} 23.25 & -6.91 & 0.375 \end{pmatrix} \begin{pmatrix} 2437.90284 \\ 13469.82610 \\ 53880.66190 \end{pmatrix} = 2.9747$$

por lo que $\hat{\sigma}_u^2 = \dfrac{2.9747}{35-3} = 0.0929$.

Buscando en las tablas de la *t*-Student, se obtiene un valor igual a 2.03 para $t_{32}^{0.025}$. En consecuencia, la precisión para un nivel de significación del 5% se muestra en la Tabla 3.1 para cada uno de los parámetros estimados. En la misma también se incluye la estimación por intervalos.

Tabla 3.1

Parámetro	Precisión	Intervalo
β_1	2.09	(21.16 25.34)
β_2	458.34	(−451.43 465.26)
β_3	114.62	(−114.25 115)

Como se puede observar, la precisión de los estimadores es muy mala, puesto que, a excepción de la constante, toma valores muy altos con respecto a los estimadores.

Significación individual

En cuanto a la significación de los parámetros la podemos realizar bien con el contraste de significación individual o bien comprobando si el cero está incluido en la estimación por intervalos. Si esto es así, ello implica que no se rechaza la hipótesis de que el parámetro en cuestión toma el valor cero y que,

por tanto, la variable a la que acompaña en el modelo no tiene capacidad explicativa sobre la variable endógena.

En nuestro caso, claramente los parámetros β_2 y β_3 son no significativos al 5% de nivel de significación. Recordemos que, si se quisiera realizar el contraste de significación individual, el estadístico de prueba no es más que el cociente entre el estimador y su desviación típica estimada. En la Tabla 3.2 se muestra el valor de los parámetros estimados, su desviación típica estimada, el estadístico de contraste y el valor crítico de la *t*-Student de 32 grados de libertad para un nivel de significación bilateral del 5%.

Tabla 3.2

Parámetro	Estimador	Desviación típica estimada	Estadístico de prueba	Punto crítico
β_1	23.2511221	1.02991506	22.57576670	2.03
β_2	6.9137392	225.86146600	0.03061053	2.03
β_3	0.3746984	56.46486960	0.00663596	2.03

La conclusión con los datos de la Tabla 3.2 es la misma que la alcanzada con los intervalos de confianza.

Significatividad conjunta

La significatividad conjunta la podemos estudiar con la expresión de la F de Fisher-Snedecor en términos del coeficiente de determinación. El estadístico de prueba es

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)}$$

Para calcular el coeficiente de determinación, dado que ya se ha calculado $e'e$, únicamente resta por calcular la *SCT*. Para ello tenemos en cuenta el siguiente desarrollo:

$$SCT = Y'Y - N\left(\frac{\sum_{i=1}^{N} y_i}{N}\right)^2 = 170002.814 - 35\cdot\left(\frac{2437.90284}{35}\right)^2 = 192.235$$

Por tanto, $R^2 = 0.9845$ y el estadístico de prueba se calcula como:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)} = \frac{0.9845\cdot 32}{2\cdot(1-0.9845)} = 1017.95$$

El punto crítico para una F de Fisher-Snedecor de 2 y 32 grados de libertad y una significación del 5% es igual a 3.30. En consecuencia, rechazamos la hipótesis nula de que las dos variables conjuntamente no tengan capacidad explicativa de la endógena.

En conclusión, como se puede observar de los resultados anteriores, parece que existe una contradicción en el sentido de que individualmente ninguna de las dos variables parece aportar nada significativo en la explicación de la endógena, pero conjuntamente son muy significativas. De hecho, conjuntamente llegan a explicar más del 98% de las variaciones de la endógena. Claramente nos encontramos con un problema de multicolinealidad aproximada.

EJERCICIO 3.4

Para una muestra de empresas de un determinado sector económico se ha ajustado la siguiente función de producción Cobb-Douglas:

$$Q_i = \beta_1 L_i^{\beta_2} K_i^{\beta_3} e^{u_i} \quad N = 23$$

Para que el modelo sea lineal en los parámetros tomamos logaritmos, obteniendo así la siguiente especificación:

$$\ln Q_i = \ln \beta_1 + \beta_2 \ln L_i + \beta_3 \ln K_i + u_i$$

El resultado de la estimación de este último modelo ha sido:

$$\ln \hat{Q}_i = 0.5 + \underset{(0.71)}{0.76} \ln L_i + \underset{(0.14)}{0.19} \ln K_i \quad R^2 = 0.969 \tag{3.1}$$

en donde los valores entre paréntesis recogen las desviaciones típicas estimadas de los coeficientes estimados.

(a) Realice los contrastes individuales de significación para β_2 y β_3 y el contraste de significación conjunto de ambos coeficientes.

(b) La estimación de los modelos de regresión simple de $\ln Q$ en función de $\ln L$ y de $\ln Q$ en función de $\ln K$ ha producido los siguientes resultados:

$$\ln \hat{Q}_t = -5.5 + \underset{(0.09)}{1.71} \ln L_t \quad R^2 = 0.964 \tag{3.2}$$

$$\ln \hat{Q}_t = 5.3 + \underset{(0.02)}{0.34} \ln K_t \quad R^2 = 0.966 \tag{3.3}$$

Realice el contraste de significación sobre cada uno de los coeficientes de los modelos. Razone la aparente contradicción entre los resultados de los contrastes obtenidos en el apartado (a) con los obtenidos en este apartado (b).

Solución

(a) Contrastes de significación individual de los coeficientes:

$$\left.\begin{array}{l} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{array}\right\} \rightarrow \text{Estadístico de contraste: } t = \frac{\hat{\beta}_2}{\hat{S}(\hat{\beta}_2)} = \frac{0.76}{0.71} = 0.93$$

$$\left.\begin{array}{l} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{array}\right\} \rightarrow \text{Estadístico de contraste: } t = \frac{\hat{\beta}_3}{\hat{S}(\hat{\beta}_3)} = \frac{0.19}{0.14} = 1.35$$

Como el valor crítico de una distribución *t*-Student con 20 grados de libertad al 95% de nivel de confianza es 2.09, no podemos rechazar la hipótesis nula. Por tanto, según este resultado, ni el trabajo ni el capital parecen aportar información significativa a la explicación de la variable endógena.

Contraste de significación global del modelo:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \exists\, \beta_j \neq 0 \end{array}\right\}$$

$$\rightarrow \text{Estadístico de contraste: } F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)} = \frac{0.969/2}{(1-0.969)/(23-3)} = 312.5$$

Como el valor crítico de la distribución $F_{2,20}$ al 95% es 3.49, rechazamos claramente la hipótesis nula, por lo que el modelo es globalmente significativo.

(b) A la vista de las estimaciones (3.2) y (3.3), se puede concluir que los factores capital y trabajo, considerados aisladamente para explicar la producción resultan estadísticamente significativos. Ello se deduce de que la ratio entre el valor del coeficiente estimado y el error estándar supera, en ambos casos, los valores críticos tabulados.

El hecho de que los coeficientes del modelo (3.1) no sean individualmente significativos, pero sí lo sean globalmente, es uno de los síntomas de la presencia de multicolinealidad aproximada en el modelo estimado. Esta incongruencia observada entre los contrastes de significación individual y el contraste de significación global en el modelo (3.1) vuelve a observarse al estimar con regresiones simples la producción (ver (3.2) y (3.3)). Todo indica claramente la existencia de una fuerte multicolinealidad entre el factor capital y trabajo.

EJERCICIO 3.5

Dada únicamente la matriz $X'X$ siguiente:

$$X'X = \begin{pmatrix} 20 & 9.1275 & 9.1374 \\ & 5.8820 & 5.8860 \\ & & 5.8900 \end{pmatrix}$$

¿existen indicios de presencia de multicolinealidad en el modelo del que procede?

Solución

Como se puede observar las columnas segunda y tercera son muy parecidas, lo que implica una relación lineal aproximada entre ellas. Ello nos muestra que las dos variables explicativas del modelo en estudio presentan una correlación importante, lo cual nos estaría dando indicios de presencia de multicolinealidad aproximada en dicho modelo.

Si además calculamos el determinante de dicha matriz, obtenemos un valor bastante pequeño, concretamente de –0.0001736, lo cual nos vuelve a indicar que el modelo en estudio presenta problemas de multicolinealidad aproximada.

EJERCICIO 3.6

Dado el modelo (3.4), donde explicamos la productividad de las empresas ($PROD$) en función del número de trabajadores ($TRAB$), de los años medios de formación de la plantilla ($ESTUD$), de una variable cualitativa que indica la nacionalidad de la empresa ($DNAC$), tomando valor 1 cuando la empresa es nacional y 0 cuando es extranjera, y de otra variable cualitativa ($DEXT$), que toma valor 1 cuando la empresa es extranjera y 0 cuando es nacional,

$$PROD_i = \beta_1 + \beta_2 TRAB_i + \beta_3 ESTUD_i + \beta_4 DNAC_i + \beta_5 DEXT_i + u_i \quad (3.4)$$

¿encuentra algún problema en la especificación del modelo? En ese caso, ¿cuáles serían las consecuencias de dicho problema?

Solución

Si se especifica el modelo (3.4), se estaría incurriendo en un problema de multicolinealidad exacta. De hecho, al analizar la matriz X, ésta podría tomar la siguiente forma:

$$X = \begin{pmatrix} 1 & Trab_1 & Est_1 & 1 & 0 \\ 1 & Trab_2 & Est_2 & 0 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & 0 & 1 \\ 1 & Trab_N & Est_N & 1 & 0 \end{pmatrix}$$

En ella vemos que, por la especificación del modelo, la primera columna coincide con la suma de las columnas cuarta y quinta. Por tanto, no podríamos calcular los coeficientes estimados, puesto que estaríamos ante un caso de combinación lineal exacta entre variables explicativas y, por tanto, el determinante de la matriz $X'X$ valdrá cero, con lo que resultaría imposible calcular $(X'X)^{-1}$ (necesario para estimar los valores de β por MCO).

EJERCICIO 3.7

Sea el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 V_i + \beta_3 W_i + u_i \tag{3.5}$$

en donde

$$W_i = \frac{50 + V_i}{2}$$

¿Cree que la estimación de este modelo presentaría algún problema?

Solución

La estimación del modelo (3.5) presenta todos los problemas derivados de la combinación lineal exacta existente entre V_i y W_i. De hecho, nos encontramos ante un problema de multicolinealidad exacta y, por tanto, no podremos estimar el modelo por MCO, ya que no se pueda calcular la inversa de $X'X$, debido a que el determinante de $X'X$ vale cero.

EJERCICIO 3.8

¿Existiría algún problema si se quisiera estimar el siguiente modelo por MCO?

$$Y_i = \beta_2 \ln X_{2i} + \beta_3 \ln X_{2i}^2 + u_i$$

Solución

El modelo propuesto no podría estimarse, porque existiría multicolinealidad exacta entre los dos regresores. De hecho $\ln X_{2i}^2 = 2\ln X_{2i}$, por tanto, el segundo regresor es el doble del primero $(\ln X_{2i})$.

EJERCICIO 3.9

Dado el siguiente modelo:

$$Y_i = \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{2i}^2 + u_i \tag{3.6}$$

en donde

$$X_{4i} = X_{2i} - 2X_{3i} \tag{3.7}$$

(a) Indique si es posible estimar de forma única algún parámetro del modelo original.

(b) ¿Qué problema presentaría la estimación por Mínimos Cuadrados Ordinarios (MCO) del modelo (3.6)? ¿Qué consecuencias tiene este problema?

(c) ¿Qué pasaría y cuál es su propuesta de actuación en caso de darse la relación $\beta_4 = \beta_2 - 2\beta_3$ en vez de $X_{4i} = X_{2i} - 2X_{3i}$?

Solución

(a) El enunciado de este ejercicio nos plantea un modelo con un problema de multicolinealidad exacta, puesto que existe una combinación lineal entre algunas de sus variables explicativas ($X_{4i} = X_{2i} - 2X_{3i}$). En estos casos, para poder estimar el modelo, habrá que transformar previamente el mismo introduciendo la relación conocida entre las variables explicativas.

Introduciendo (3.7) en (3.6) obtenemos:

$$\begin{aligned} Y_i &= \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{2i} - 2X_{3i}) + \beta_5 X_{2i}^2 + u_i = \\ &= \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i} - 2\beta_4 X_{3i} + \beta_5 X_{2i}^2 + u_i = \\ &= (\beta_2 + \beta_4) X_{2i} + (\beta_3 - 2\beta_4) X_{3i} + \beta_5 X_{2i}^2 + u_i \end{aligned}$$

La estimación de dicho modelo será:

$$\hat{Y}_i = \hat{\alpha}_1 X_{2i} + \hat{\alpha}_2 X_{3i} + \hat{\alpha}_3 X_{2i}^2$$

Por tanto, no es posible estimar directamente los parámetros β_2, β_3 y β_4 del modelo (3.6), mientras el parámetro β_5 quedará estimado directamente a través de $\hat{\alpha}_3$.

(b) Al tratarse de un problema de multicolinealidad exacta se estaría rompiendo una de las hipótesis básicas de los modelos de regresión lineal múltiple, la hipótesis de rango pleno. En este caso el rango de la matriz X sería menor al número de variables explicativas, puesto que una de ellas (X_4) se obtiene como combinación lineal de otras dos. Esto conlleva que el determinante de la matriz $X'X$ valga cero y que, por tanto, no se pueda calcular $(X'X)^{-1}$, de forma que no es posible obtener los $\hat{\beta}_{MCO}$.

(c) En caso de darse la relación $\beta_4 = \beta_2 - 2\beta_3$ estaríamos ante un caso de combinación lineal de parámetros, pero no de combinación lineal de variables. Por tanto, no nos enfrentaríamos a un problema de multicolinealidad exacta y sí sería posible estimar el modelo. No obstante, si la relación planteada entre los parámetros fuera cierta, lo correcto sería estimar el modelo por Mínimos Cuadrados Restringidos (MCR) en lugar de por MCO, de forma que pudiéramos obtener estimadores eficientes.

EJERCICIO 3.10

En un modelo de regresión $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, se conoce que $X_{3i} = a + bX_{2i}$.

(a) ¿Qué problema presenta la estimación de este modelo?

(b) Bajo el supuesto de que se conocen los valores a y b, especifique el modelo susceptible de ser estimado, si se tiene en cuenta la relación existente entre X_2 y X_3.

(c) Demuestre, utilizando las expresiones con datos centrados, que el valor del determinante de la matriz $X'X$ es cero.

Solución

(a) Los datos presentan un problema de multicolinealidad exacta.

(b) Si tenemos en cuenta la relación entre ambas variables, el modelo susceptible de ser estimado sería el siguiente:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (a + bX_{2i}) + u_i = (\beta_1 + \beta_3 a) + (\beta_2 + \beta_3 b) X_{2i} + u_i$$

lo que supone estimar el modelo de regresión lineal simple siguiente:

$$Y_i = \alpha + \beta X_{2i} + u_i \quad \text{donde} \quad \alpha = \beta_1 + \beta_3 a \quad \text{y} \quad \beta = \beta_2 + \beta_3 b$$

Ahora bien, el modelo presenta un problema de identificación de los parámetros, puesto que no podremos obtener los estimadores de los parámetros iniciales β_1, β_2 y β_3, sino tan sólo una combinación de ellos, α y β.

(c) El estimador obtenido por mínimos cuadrados ordinarios, expresado en términos de datos centrados, es igual a

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^2 & \sum_{i=1}^{N} x_{2i}x_{3i} \\ \sum_{i=1}^{N} x_{2i}x_{3i} & \sum_{i=1}^{N} x_{3i}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} x_{2i}y_i \\ \sum_{i=1}^{N} x_{3i}y_i \end{pmatrix}$$

Si $X_{3i} = a + bX_{2i}$, entonces

$$(X'X)^{-1} = \begin{pmatrix} \sum_{i=1}^{N} x_{2i}^2 & b\sum_{i=1}^{N} x_{2i}^2 \\ b\sum_{i=1}^{N} x_{2i}^2 & b^2\sum_{i=1}^{N} x_{2i}^2 \end{pmatrix}^{-1} = \frac{1}{\sum_{i=1}^{N} x_{2i}^2} \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix}^{-1},$$

puesto que $x_{3i} = X_{3i} - \overline{X}_3 = (a + bX_{2i}) - (a + b\overline{X}_2) = b(X_{2i} - \overline{X}_2) = bx_{2i}$.

A partir de esta expresión, resulta inmediato comprobar que el determinante de esta matriz vale cero, puesto que

$$\begin{vmatrix} 1 & b \\ b & b^2 \end{vmatrix} = b^2 - b^2 = 0.$$

EJERCICIO 3.11

Calcule el factor de inflación de la varianza (*FIV*) del modelo (3.8) e interprete el resultado:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{3.8}$$

siendo Y el consumo de leche, X_2 la renta disponible y X_3 el precio del litro de leche, y sabiendo además que:

$$S_Y^2 = 17.2 \quad S_{X_2}^2 = 12.9 \quad S_{X_3}^2 = 10.25 \quad S_{X_2,X_3} = 5.65$$

Solución

El *FIV* correspondiente a una variable X_j se calcula a partir del coeficiente de determinación de la regresión auxiliar que estima cada variable X_j en función del resto de variables explicativas. En este caso existen tan sólo dos variables explicativas, por lo que la regresión auxiliar correspondiente es:

$$X_{2i} = \gamma_1 + \gamma_2 X_{3i} + v_i$$

Con los datos del enunciado podemos calcular la correlación entre X_2 y X_3 y, a partir de la correlación, podemos obtener el coeficiente de determinación de la regresión auxiliar R_j^2, ya que en regresión simple se verifica que:

$$r_{X_2,X_3}^2 = R_j^2$$

Sabiendo que la correlación lineal simple se obtiene como el cociente entre la covarianza entre las dos variables y el producto de sus desviaciones típicas, obtenemos:

$$r_{X_2,X_3} = \frac{S_{X_2,X_3}}{S_{X_2} S_{X_3}} = \frac{5.65}{\sqrt{12.9} \cdot \sqrt{10.25}} = 0.49135$$

Por tanto, el coeficiente de determinación de la regresión auxiliar será:

$$R_j^2 = r_{X_2,X_3}^2 = 0.49135^2 = 0.2414$$

Finalmente, el *FIV* se calcula a partir de la expresión:

$$FIV\left(X_j\right) = \frac{1}{1 - R_j^2} = 1.318$$

Con lo que las varianzas de las estimaciones MCO de los coeficientes del modelo (3.8) se incrementan en un 31.8% respecto a las que se obtendrían en un modelo con total ausencia de correlación entre las variables explicativas, o sea en ausencia de multicolinealidad.

EJERCICIO 3.12

Con la información de la Tabla 3.3, analice el problema de multicolinealidad en un modelo en el que la variable Y viene explicada por las variables X_2 y X_3 y una constante[2].

[2] Se recomienda hacer este ejercicio mediante el uso de una hoja de cálculo o mediante un programa de tratamiento estadístico. La solución que presentamos está realizada mediante el programa EViews.

Tabla 3.3

Y	X_2	X_3
66.5562523	5.18306490	20.7326689
66.1628862	5.12208277	20.4892643
66.5469353	5.11795051	20.4725990
70.2076572	5.52731675	22.1102175
68.7278073	5.38991927	21.5602575
72.1692334	5.83009034	23.3212261
71.1843840	5.71482074	22.8601716
65.2835742	5.01038480	20.0424441
67.7667431	5.28004430	21.1204676
72.9124242	5.87482970	23.4999873
69.7311680	5.51007425	22.0403871
69.4277154	5.49844505	21.9940137
73.7211779	5.97359954	23.8952564
66.0939196	5.13143581	20.5266037
73.4523458	5.94661115	23.7870439
68.2152258	5.35405473	21.4163777
70.1947336	5.57248788	22.2906936
69.6662676	5.46245345	21.8506186
69.7273432	5.58251257	22.3303430
70.3378558	5.60180349	22.4073330
72.0782670	5.79681218	23.1874253
72.5115581	5.93247535	23.7301700
69.1935676	5.50045602	22.0027497
69.2570862	5.40971063	21.6392672
73.0187767	5.91034769	23.6416919
68.9096920	5.41984926	21.6803091
70.6179649	5.57985620	22.3199248
66.0271491	5.10222971	20.4091493
72.3925089	5.85890540	23.4364941
70.4508941	5.56203475	22.2486348
69.2840803	5.49609042	21.9850485
68.6248823	5.33201888	21.3284603
66.0522571	5.12105412	20.4844635
68.9258414	5.44874499	21.7952034
72.4726603	5.90337383	23.6142752

Solución

Con estos datos el primer cálculo consistiría en obtener la matriz de correlación lineal entre las tres variables (la endógena y las dos exógenas). Dicha matriz se muestra en el Cuadro 3.1.

Cuadro 3.1

Matriz Correlaciones

	Y	X_2	X_3
Y	1	0.99223021	0.99223239
X_2	0.99223021	1	0.99999997
X_3	0.99223239	0.99999997	1

Como se puede observar ambas variables explicativas están muy correlacionadas con la endógena, pero también lo están entre ellas, con una correlación que es casi igual a 1.

Aparte de la correlación, podemos calcular el factor de inflación de la varianza. Para ello realizamos una regresión de la variable X_2 en función de X_3. Los resultados de la estimación realizada con el EViews se muestran en el Cuadro 3.2.

Cuadro 3.2

Dependent Variable: X2
Method: Least Squares
Sample: 1 35
Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–9.01E–05	0.000250	–0.360171	0.7210
X3	0.249998	1.13E–05	222085.49	0.0000
R-squared	1.000000	Mean dependent var		5.515941
Adjusted R-squared	1.000000	S.D. dependent var		0.280455
S.E. of regression	7.40E–05	Akaike info criterion		–16.128350
Sum squared resid	1.81E–07	Schwarz criterion		–16.039480
Log likelihood	284.246200	F-statistic		4.88E+08
Durbin-Watson stat	1.711263	Prob(F-statistic)		0.000000

A partir de estos resultados podemos ver que la correlación entre ambas variables es igual a 1. Por tanto, y dado que el factor de inflación de la varianza se calcula como

$$FIV\left(X_j\right)=\frac{1}{1-R_j^2},$$

su valor tendería a infinito.

De esta forma hemos comprobado por dos métodos diferentes que la estimación del modelo propuesto está sujeta a la presencia de multicolinealidad aproximada muy alta, casi exacta.

EJERCICIO 3.13

Dado el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

se ha procedido a estimar con 100 datos el siguiente modelo auxiliar para la variable X_2:

$$X_{2i} = \lambda_1 + \lambda_3 X_{3i} + \lambda_4 X_{4i} + \lambda_5 X_{5i} + v_i \tag{3.9}$$

La estimación de este último modelo presenta un estadístico F de Fisher-Snedecor igual a 1 para el contraste

$$\left.\begin{array}{l} H_0: \lambda_3 = \lambda_4 = \lambda_5 = 0 \\ H_1: H_0 \text{ no se cumple} \end{array}\right\} \tag{3.10}$$

Calcule el factor de inflación de la varianza de la variable X_2 para el modelo de Y.

Solución

El factor de inflación de la varianza de la variable X_2 se calcula como

$$FIV(X_2) = \frac{1}{(1 - R_2^2)},$$

siendo R_2^2 el coeficiente de determinación de la regresión de la variable X_2 en función del resto de variables explicativas. Es decir, el coeficiente de determinación de la regresión (3.9).

Para calcular este valor únicamente debemos tener en cuenta que el estadístico de prueba del contraste (3.10) se puede escribir como

$$F = \frac{R^2}{1 - R^2} \cdot \frac{N - k}{k - 1}$$

Despejando R^2 de esta última expresión se obtiene

$$R^2 = \frac{F \cdot (k-1)}{F \cdot (k-1) + (N-k)}$$

Teniendo en cuenta que estamos hablando de datos de la ecuación (3.9), tendremos que $F=1$, $k=4$ y $N=100$. Sustituyendo estos valores en la expresión del coeficiente de determinación se obtiene un valor igual a $3/99=0.03$.

Por último, sustituyendo dicho valor en la expresión que nos permite calcular el *FIV* se obtiene un valor igual a 1.031. Su reducido valor nos indica que la variable X_2 no presenta problemas de multicolinealidad en la ecuación que permite estimar el modelo para la variable Y.

EJERCICIO 3.14

Dado el siguiente modelo:

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+u_i$$

obtenga el factor de inflación de la varianza para la variable X_2 a partir de la siguiente estimación, donde $N=10$:

$$\hat{X}_{2i}=13.69+0.1364X_{3i} \tag{3.11}$$

sabiendo que el estadístico F correspondiente al contraste global del modelo (3.11) vale 18.26.

Solución

El factor de inflación de la varianza se calcula como

$$FIV\left(X_2\right)=\frac{1}{1-R_2^2},$$

donde R_2^2 es el coeficiente de determinación de la regresión auxiliar (3.11).

A partir del estadístico del contraste global de (3.11), podemos obtener R_2^2, puesto que la expresión del mismo es:

$$F=\frac{R_2{}^2/(k-1)}{\left(1-R_2{}^2\right)/(N-k)}$$

Despejando de la misma el valor de R_2^2, obtenemos:

$$R_2^2=\frac{(k-1)F}{F(k-1)+(N-k)}=\frac{(2-1)\cdot 18.26}{18.26\cdot(2-1)+(10-2)}=0.695$$

Por tanto, el factor de inflación de la varianza vale

$$FIV(X_2) = \frac{1}{1-R_2^2} = \frac{1}{1-0.695} = 3.27$$

A la vista del resultado, podríamos calificar de moderada la presencia de multicolinealidad.

EJERCICIO 3.15

Con relación al modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

obtenga el factor de inflación de la varianza para la variable X_2 a partir de la estimación del Cuadro 3.3.

Cuadro 3.3

Dependent Variable: X_2
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X_3	0.136486	0.031937	4.273640	0.0005
C	13.692760	8.683002	1.576962	0.1322

Solución

Para obtener el *FIV* se utiliza la expresión:

$$FIV(X_2) = \frac{1}{1-R_2^2}$$

con lo que será necesario calcular el valor de R_2^2.

Éste se puede obtener a partir de la expresión del contraste de significación global:

$$F = \frac{R_2^2(N-k)}{(1-R_2^2)(k-1)} \tag{3.12}$$

En los modelos de regresión lineal simple, el valor del estadístico de significación global *F*, coincide con el cuadrado del estadístico de significación individual *t*. Dado que la regresión auxiliar del Cuadro 3.3 es un modelo con una única variable explicativa, podemos aplicar esta relación y así, sustituyendo los valores conocidos en la expresión (3.12), obtenemos el valor de R_2^2:

$$(4.273640)^2 = \frac{R_2^2(20-2)}{(1-R_2^2)(2-1)} = \frac{18R_2^2}{(1-R_2^2)} \Rightarrow R_2^2 = 0.50369957$$

Por tanto, el factor de inflación de la varianza para la variable X_2 valdrá:

$$FIV(X_2) = \frac{1}{1-0.50369957} = 2.01467$$

EJERCICIO 3.16

Determine los valores de los factores de inflación de la varianza, si en todas las regresiones auxiliares de las exógenas, la probabilidad asociada al estadístico de prueba del contraste del análisis de la varianza es igual a 1.

Solución

Dados los posibles valores que puede tomar una distribución *F* de Fisher-Snedecor, el que la probabilidad asociada a un estadístico determinado sea la unidad, implica que dicho estadístico toma el valor cero.

Teniendo en cuenta este resultado, y sabiendo que el estadístico del análisis de la varianza se calcula como

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)}$$

es inmediato demostrar que el coeficiente de determinación se calcula como

$$R^2 = \frac{F\cdot(k-1)}{F\cdot(k-1)+(N-k)}$$

Dado que $F = 0$, el coeficiente de determinación será igual a

$$R^2 = \frac{F\cdot(k-1)}{F\cdot(k-1)+(N-k)} = \frac{0}{N-k} = 0$$

Por último, teniendo en cuenta que

$$FIV(X_j) = \frac{1}{(1-R_j^2)},$$

es inmediato concluir que el valor del *FIV* será siempre igual a 1.

EJERCICIO 3.17

Se ha estimado el modelo

$$\hat{Y}_i = \underset{(2.24)}{1.72} + \underset{(5.30)}{3.38} X_{2i}$$

siendo el valor que figura entre paréntesis, el ratio t.

Sin embargo, se está estudiando la posibilidad de incluir como variable explicativa X_3, con lo que el modelo estimado quedaría tal que:

$$\hat{Y}_i = \underset{(3.41)}{1.80} + \underset{(1.02)}{0.74} X_{2i} + \underset{(4.57)}{0.51} X_{3i} \tag{3.13}$$

La elevada variación que presenta el coeficiente de X_2 cuando se comparan ambas estimaciones hace sospechar la existencia de multicolinealidad. Para cuantificar esta posibilidad, calcule el factor de inflación de la varianza para el modelo (3.13) y comente el resultado.

Solución

Sabemos que el Factor de Inflación de la Varianza de la estimación de un parámetro β_j es el número de veces en que se multiplica su varianza en el modelo de regresión lineal múltiple respecto al modelo de regresión lineal simple. De esta manera

$$Var\left(\hat{\beta}_j\right) = \underbrace{\frac{\sigma_u^2}{\sum_{i=1}^{N} x_{ji}^2}}_{Var\left(\beta_j'\right)} \cdot \underbrace{\frac{1}{1-R_j^2}}_{FIV} \quad \Rightarrow \quad Var\left(\hat{\beta}_j\right) = Var\left(\hat{\beta}_j'\right) FIV\left(X_j\right)$$

siendo $Var\left(\hat{\beta}_j\right)$ la varianza de la estimación de la regresión lineal múltiple y $Var\left(\hat{\beta}_j'\right)$ la varianza de la regresión lineal simple.

De la expresión anterior podemos deducir que:

$$FIV\left(X_j\right) = \frac{Var\left(\hat{\beta}_j\right)}{Var\left(\hat{\beta}_j'\right)} \tag{3.14}$$

Además, sabemos que el estadístico de significación individual se obtiene a través de la expresión:

$$t = \frac{\hat{\beta}_j}{S\left(\hat{\beta}_j\right)} \tag{3.15}$$

Por tanto, despejando (3.15) podemos obtener el valor de la varianza de los estimadores y sustituyendo estos valores en (3.14) obtenemos que:

$$FIV\left(X_j\right)=\frac{Var\left(\hat{\beta}_j\right)}{Var\left(\hat{\beta}'_j\right)}=\frac{\left(\frac{\beta_j}{t}\right)^2}{\left(\frac{\beta'_j}{t}\right)^2}=\frac{\left(\frac{0.74}{1.02}\right)^2}{\left(\frac{3.38}{5.30}\right)^2}=1.2941$$

Dado el reducido valor obtenido, no podemos hablar de la presencia de multicolinealidad preocupante.

EJERCICIO 3.18

Los resultados de la estimación de un modelo en el que se explica el logaritmo de los kilos de pescado capturado por 30 embarcaciones en la cofradía de pescadores de Arguineguín, $LOG(CAPTURAS)$, en función de la capacidad en bodega ($BODEGA$), el número de días de autonomía del barco ($AUTONOMIA$), la potencia del motor ($POTENCIA$), dos medidas de longitud como son la manga ($MANGA$) y la eslora ($ESLORA$), así como los factores de inversión en capital y trabajo, medidos a través de la inversión en artes de pesca, $LOG(INV\_ARTES\_PESCA)$ y el número de marineros, $LOG(N\_MARINEROS)$, vienen dados en el Cuadro 3.4.

Cuadro 3.4

Dependent Variable: *LOG(CAPTURAS)*
Method: Least Squares
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BODEGA	0.000426	0.000221	1.928848	0.0668
AUTONOMIA	–0.030341	0.112215	–0.270382	0.7894
POTENCIA	–0.000343	0.005036	–0.068067	0.9463
MANGA	–0.131514	0.630370	–0.208630	0.8367
ESLORA	–0.292778	0.223548	–1.309691	0.2038
LOG(INV_ARTES_PESCA)	0.219872	0.263304	0.835048	0.4127
LOG(N_MARINEROS)	1.381093	0.754851	1.829623	0.0809
C	5.944171	2.360748	2.517918	0.0196

(*continúa en la página siguiente*)

Cuadro 3.4 (continuación)

R-squared	0.715895	Mean dependent var	7.343551
Adjusted R-squared	0.625497	S.D. dependent var	1.421831
S.E. of regression	0.870112	Akaike info criterion	2.782789
Sum squared resid	16.65609	Schwarz criterion	3.156442
Log likelihood	–33.741840	F-statistic	7.919436
Durbin-Watson stat	2.529244	Prob(F-statistic)	0.000080

(a) Comente, utilizando toda la información posible del Cuadro 3.4, qué problema sospecha que presentan los datos.

(b) Obtenga los intervalos de confianza de la estimación de los parámetros, para un nivel de confianza del 90%, para las variables *AUTONOMIA*, *POTENCIA*, *MANGA* y *ESLORA* y comente qué implicaciones tienen los resultados obtenidos.

(c) De las regresiones auxiliares se obtienen los coeficientes de determinación de la Tabla 3.4:

Tabla 3.4

Variable dependiente de la regresión	Coeficiente de determinación con el resto de exógenas como explicativas
BODEGA	0.960
AUTONOMIA	0.906
POTENCIA	0.943
MANGA	0.951
ESLORA	0.958
LOG(INV_ARTES_PESCA)	0.715
LOG(N_MARINEROS)	0.826

Calcule el Factor de Inflación de la Varianza e indique qué variable/s causa/n el problema mencionado en el apartado (a).

Solución

(a) Es lógico sospechar que los datos presentan un problema de multicolinealidad aproximada entre las variables exógenas del modelo, ya que las medidas de bondad global del ajuste (R^2, R^2 corregido en grados de libertad) son elevadas y el estadístico F de Fisher-Snedecor de significación global de los coeficientes indica que el modelo es globalmente significativo,

mientras que los contrastes de significación individual nos indican que hay muchas variables no significativas a niveles estándar. Además, los signos que presentan algunos de los coeficientes son contrarios a lo que nos sugiere la teoría económica en cualquier función de producción al uso.

Hay que tener en cuenta que las variables *BODEGA*, *POTENCIA*, *AUTONOMIA*, *ESLORA* y *MANGA* miden todas ellas prácticamente lo mismo, el "tamaño" del barco, y es seguro que estarán altamente correlacionadas.

(b) A continuación calculamos los intervalos de confianza solicitados:

AUTONOMIA

$$P\left(\hat{\beta}_2 - t_{N-k}\hat{S}\left(\hat{\beta}_2\right) \le \beta_2 \le \hat{\beta}_2 + t_{N-k}\hat{S}\left(\hat{\beta}_2\right)\right) = 0.90$$

$$P\left(-0.030341 - 1.7171 \cdot 0.112215 \le \beta_2 \le -0.030341 + 1.7171 \cdot 0.112215\right) = 0.90$$

$$P\left(-0.223 \le \beta_2 \le 0.162\right) = 0.90$$

POTENCIA

$$P\left(\hat{\beta}_3 - t_{N-k}\hat{S}\left(\hat{\beta}_3\right) \le \beta_3 \le \hat{\beta}_3 + t_{N-k}\hat{S}\left(\hat{\beta}_3\right)\right) = 0.90$$

$$P\left(-0.000343 - 1.7171 \cdot 0.005036 \le \beta_3 \le -0.000343 + 1.7171 \cdot 0.005036\right) = 0.90$$

$$P\left(-0.009 \le \beta_3 \le 0.008\right) = 0.90$$

MANGA

$$P\left(\hat{\beta}_4 - t_{N-k}\hat{S}\left(\hat{\beta}_4\right) \le \beta_4 \le \hat{\beta}_4 + t_{N-k}\hat{S}\left(\hat{\beta}_4\right)\right) = 0.90$$

$$P\left(-0.131514 - 1.7171 \cdot 0.630370 \le \beta_4 \le -0.131514 - 1.7171 \cdot 0.630370\right) = 0.90$$

$$P\left(-1.214 \le \beta_4 \le 0.951\right) = 0.90$$

ESLORA

$$P\left(\hat{\beta}_5 - t_{N-k}\hat{S}\left(\hat{\beta}_5\right) \le \beta_5 \le \hat{\beta}_5 + t_{N-k}\hat{S}\left(\hat{\beta}_4\right)\right) = 0.90$$

$$P\left(-0.292778 - 1.7171 \cdot 0.223548 \le \beta_5 \le -0.292778 + 1.7171 \cdot 0.223548\right) = 0.90$$

$$P\left(-0.677 \le \beta_5 \le 0.091\right) = 0.90$$

Los intervalos de confianza son muy amplios con respecto a los valores de los parámetros, lo que conlleva una gran incertidumbre acerca de sus valores y, además, todos ellos contienen al cero, por lo que ninguna de estas variables es significativa.

(c) Calculamos los Factores de Inflación de la Varianza a partir de la expresión

$$FIV(X_j) = \frac{1}{(1-R_j^2)}$$

y obtenemos los resultados de la Tabla 3.5:

Tabla 3.5

Variable dependiente de la regresión	R^2	*FIV*
BODEGA	0.960	25.00
AUTONOMIA	0.906	10.64
POTENCIA	0.943	17.54
MANGA	0.951	20.41
ESLORA	0.958	23.81
LOG(INV_ARTES_PESCA)	0.715	3.51
LOG(N_MARINEROS)	0.826	5.75

Las variables que causan mayores problemas de multicolinelidad son las que tienen un mayor *FIV*. Estas variables, concretamente, son las que están relacionadas con el "tamaño" del barco, lo que indica que estamos midiendo con demasiadas variables una misma dimensión. En la Tabla 3.5 se observa que para el caso de los coeficientes estimados de estas variables, relativas al tamaño, sus correspondientes varianzas de estimación son entre 10.64 y 25 veces más grandes de las que se obtendrían ante la ausencia de multicolinealidad, en un modelo con correlaciones iguales a cero entre todas variables explicativas, marcando estos límites las variables *AUTONOMIA* y *BODEGA*, respectivamente. En el resto de las variables explicativas, $LOG(INV\_ARTES\_PESCA)$ y $LOG(N\_MARINEROS)$, el efecto de la multicolinealidad es menor.

EJERCICIO 3.19

La estimación del siguiente modelo de regresión:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (3.16)$$

en el que Y representa el importe total (en millones de euros) de las hipotecas en el año 2002 de cada una de las 50 provincias españolas, X_2 representa el ahorro familiar bruto (en millones de euros) y X_3 el precio del metro cuadrado de las viviendas, proporciona los siguientes resultados:

$$\hat{Y}_i = 0.17 + 1.34X_{2i} - 0.33X_{3i}$$

No se dispone de regresiones auxiliares, pero se conoce la siguiente información para datos centrados y no centrados:

Datos centrados	Datos no centrados
$\sum_{i=1}^{N} x_2^2 = 53 \quad \sum_{i=1}^{N} x_3^2 = 5 \quad \sum_{i=1}^{N} x_2 x_3 = 8$	$\sum_{i=1}^{N} X_2 = 36.38456 \quad \sum_{i=1}^{N} X_3 = 47.21491$ $\sum_{i=1}^{N} X_2^2 = 78.31992 \quad \sum_{i=1}^{N} X_3^2 = 49.6359$

(a) Calcule el Factor de Inflación de la Varianza y comente la posible presencia de multicolinealidad

(b) Señale cuáles son las consecuencias de la presencia de multicolinealidad en un modelo de regresión lineal múltiple.

Solución

(a) Como el modelo (3.16) tiene tan sólo dos variables explicativas, las regresiones auxiliares del mismo tendrán tan sólo una variable explicativa, por lo que estas regresiones auxiliares serán modelos de regresión lineal simple que tomarán la siguiente especificación:

$$X_{2i} = \gamma_1 + \gamma_2 X_{3i} + \nu_i$$
$$X_{3i} = \mu_1 + \mu_2 X_{2i} + \upsilon_i$$

Además, en los modelos de regresión lineal simple se cumple que $R_j^2 = r_{X_2,X_3}^2$.

Por tanto, comenzamos tratando de obtener el valor de dicho coeficiente de correlación lineal simple:

$$r_{X_2,X_3} = \frac{S_{X_2,X_3}}{S_{X_2} S_{X_3}} = \frac{\sum_{i=1}^{N} (X_2 - \bar{X}_2)(X_3 - \bar{X}_3)}{\sqrt{\sum_{i=1}^{N} (X_2 - \bar{X}_2)^2} \sqrt{\sum_{i=1}^{N} (X_3 - \bar{X}_3)^2}} =$$

$$= \frac{\sum_{i=1}^{N} (x_2 x_3)}{\sqrt{\sum_{i=1}^{N} x_2^2} \sqrt{\sum_{i=1}^{N} x_3^2}} = \frac{8}{\sqrt{53} \cdot \sqrt{5}} = 0.4914$$

A partir de este resultado podemos obtener el coeficiente de determinación como

$$R^2 = r^2_{X_2,X_3} = 0.4914^2 = 0.2415$$

Y, finalmente, a partir del coeficiente de determinación podemos obtener el *FIV*:

$$FIV\left(X_j\right) = \frac{1}{1-R_j^2} = 1.3184$$

concluyendo que no hay indicios de presencia de multicolinealidad preocupante en el modelo especificado.

(b) Consecuencias de la multicolinealidad:

- La multicolinealidad genera varianzas de los estimadores que aumentan cuando se incrementa el grado de la multicolinealidad. Ello genera falta de precisión de los estimadores.
- Las altas varianzas de los estimadores generan gran incertidumbre sobre el valor del parámetro.
- Muchos contrastes de significación individual indican "erróneamente" coeficientes no significativos.
- Los modelos son válidos para la predicción, pero presentan problemas a la hora de realizar análisis estructural, debido a que no se pueden aislar los efectos de cada variable individualmente.

EJERCICIO 3.20

Se estima el modelo (3.17) que explica la variabilidad en el número de pernoctaciones en 51 provincias españolas en función del número total de visitantes, así como del número de plazas en hoteles y hostales, además de una dicotómica que recoge el efecto "zona costera" (tomando el valor 1 cuando se trata de una provincia costera y cero cuando se trata de una provincia del interior).

$$PERNOCTA_i = \beta_1 + \beta_2 VISIT_i + \beta_3 HOTEL_i + \beta_4 HOSTAL_i + \beta_5 COSTA_i + u_i \qquad (3.17)$$

Los resultados de la estimación están recogidos en el Cuadro 3.5.

Cuadro 3.5

Dependent Variable: *PERNOCTA*
Method: Least Squares
Sample: 1 51
Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
VISIT	1.142788	0.146158	7.818828	0.0000
HOTEL	–5.800268	4.289428	–1.352224	0.1829
HOSTAL	–27.985910	36.460440	–0.767569	0.4467
COSTA	755571.2	215109.7	3.512493	0.0010
C	314034.0	153822.0	2.041541	0.0470
R-squared	0.8291420	Mean dependent var		1700361
Adjusted R-squared	0.8142840	S.D. dependent var		1585166
S.E. of regression	683123.4	Akaike info criterion		29.799630
Sum squared resid	2.15E+13	Schwarz criterion		29.989030
Log likelihood	–754.8906000	F-statistic		55.807160
Durbin-Watson stat	1.4382610	Prob(F-statistic)		0.000000

(a) ¿Qué problema cree que presentan los datos? Justifique su respuesta.

(b) Si dispone de la información sobre regresiones auxiliares de la Tabla 3.6, ¿cuál/es es/son la/s variable/s que está/n provocando el problema?

Tabla 3.6

Endógena	Explicativas	Estadístico $F_{3,47}^{0.05} = 2.8$
HOTEL	*HOSTAL, VISIT, COSTA, C*	65.2
HOSTAL	*HOTEL, VISIT, COSTA, C*	63.5
VISITANTES	*HOTEL, HOSTAL, COSTA, C*	32.5

(c) ¿Qué modelo alternativo sugiere que debe especificarse?

Solución

(a) Los datos presentan problemas de multicolinealidad, ya que el modelo estimado presenta un R^2 alto, que explica el 82.9% de la variabilidad de la endógena, mientras que algunos de los contrastes de significación individual revelan variables no significativas en el análisis cuando la teoría indica que son variables importantes. Además, los signos de los coeficientes relativos a las variables *HOSTAL* y *HOTEL* no son correctos.

(b) A partir del valor del estadístico F de Fisher-Snedecor podemos obtener el valor del coeficiente de determinación de cada una de las regresiones auxiliares

$$F = \frac{R_j^2/(k-1)}{(1-R_j^2)/(N-k)} \tag{3.18}$$

Operando en (3.18) obtenemos

$$F\frac{(1-R_j^2)}{(N-k)} = \frac{R_j^2}{(k-1)}$$

$$\frac{(F-F\cdot R_j^2)}{(51-4)} = \frac{R_j^2}{(4-1)}$$

$$(F-F\cdot R_j^2)(4-1) = R_j^2(51-4)$$

$$3F-3F\cdot R_j^2 = 47R_j^2$$

$$3F = 47R_j^2 + 3F\cdot R_j^2$$

$$R_j^2 = \frac{3F}{47+3F}$$

Mediante esta expresión podemos obtener los coeficientes de determinación necesarios para calcular el Factor de Inflación de la Varianza para cada variable, tal y como se muestra en la Tabla 3.7.

Tabla 3.7

Endógena	Estadístico *F*	R_j^2	*FIV*
HOTEL	65.2	0.80626546	5.16170213
HOSTAL	63.5	0.80210526	5.05319149
VISITANTES	32.5	0.67474048	3.07446809

Las variables que causan una mayor multicolinealidad son *HOTEL*, *HOSTAL* y *VISIT*, ya que son las que presenta un mayor *FIV*.

(c) Una posible solución a este problema de multicolinealidad puede pasar por crear una nueva variable que aglutine el número de plazas totales ofertadas $TOTAL\_PLAZAS = HOTEL + HOSTAL$, lo que sugiere una especificación del modelo como la que sigue:

$$PERNOCTA_i = \beta_1 + \beta_2 VISIT_i + \beta_3 TOTAL\_PLAZAS_i + \beta_5 COSTA_i + u_i$$

EJERCICIO 3.21

Si un modelo de regresión lineal múltiple presenta la matriz de datos de la Tabla 3.8, estudie la presencia de influencia potencial.

Tabla 3.8

Regresor Ficticio	X_2	X_3
1	1	3
1	3	2
1	4	3
1	3	4
1	2	3
1	3	4
1	4	5
1	5	6

Solución

Las dos medidas que se han estudiado para el análisis de la influencia potencial son el apalancamiento o *leverage* y la distancia de Cook. Para calcular esta última se necesitaría estimar el modelo y, dado que no disponemos de los datos de la variable endógena, no es posible obtener dicha medida. En consecuencia, la medida que tenemos que calcular es el apalancamiento, teniendo en cuenta que ello únicamente nos permitirá analizar la presencia de observaciones con influencia potencial.

La medida de apalancamiento se obtiene a partir de la matriz $H = X\left(X'X\right)^{-1}X'$, denominada *Hat Matrix*. Concretamente, en su diagonal principal se encuentra el *leverage* de cada observación.

Con los datos del enunciado es inmediato obtener dicha matriz. El único problema que puede plantearse proviene del tamaño de las matrices con las que se está trabajando. Esto obliga a utilizar un programa informático que nos facilite su cálculo. Uno de los programas más accesibles es el Excel, pero para poder utilizarlo es necesario saber cuál es la dimensión de cada una de las matrices con las que estamos trabajando. Recordemos que la matriz $X'X$ y su inversa es de orden $k \cdot k$, mientras que X es de orden $N \cdot k$ y, por tanto, X' es de orden $k \cdot N$. Esto significa que la matriz $X\left(X'X\right)^{-1}$ es el producto de dos matrices, la primera es de orden $N \cdot k$ y la segunda de orden $k \cdot k$. Por tanto, su resultado es de orden $N \cdot k$. Al multiplicar $X\left(X'X\right)^{-1}$ por X', resulta

que la matriz H es de orden $N \cdot N$. En este ejercicio $N = 8$, con lo cual la matriz H es de orden $8 \cdot 8$.

A continuación se incluyen todas las matrices necesarias para obtener la matriz H, calculadas con la hoja electrónica Excel haciendo uso de las funciones MMULT y MINVERSA de dicho programa.

$$X'X = \begin{pmatrix} 8 & 25 & 30 \\ 25 & 89 & 101 \\ 30 & 101 & 124 \end{pmatrix} \quad (X'X)^{-1} = \begin{pmatrix} 1.43965517 & -0.12068966 & -0.25 \\ -0.12068966 & 0.15862069 & -0.10 \\ -0.25000000 & -0.10000000 & 0.15 \end{pmatrix}$$

$$X(X'X)^{-1} = \begin{pmatrix} 0.56896552 & -0.26206897 & 0.10 \\ 0.57758621 & 0.15517241 & -0.25 \\ 0.20689655 & 0.21379310 & -0.20 \\ 0.07758621 & -0.04482759 & 0.05 \\ 0.44827586 & -0.10344828 & -3.3307 \cdot 10^{-16} \\ 0.07758621 & -0.04482759 & 0.05 \\ -0.29310345 & 0.01379310 & 0.10 \end{pmatrix}$$

$$H = X(X'X)^{-1}X' =$$

$$= \begin{pmatrix} 0.607 & -0.017 & -0.179 & 0.183 & 0.345 & 0.183 & 0.021 & -0.141 \\ -0.017 & 0.543 & 0.448 & 0.043 & 0.138 & 0.043 & -0.052 & -0.147 \\ -0.179 & 0.448 & 0.462 & 0.048 & 0.034 & 0.048 & 0.062 & 0.076 \\ 0.183 & 0.043 & 0.048 & 0.143 & 0.138 & 0.143 & 0.148 & 0.153 \\ 0.345 & 0.138 & 0.034 & 0.138 & 0.241 & 0.138 & 0.034 & -0.069 \\ 0.183 & 0.043 & 0.048 & 0.143 & 0.138 & 0.143 & 0.148 & 0.153 \\ 0.021 & -0.052 & 0.062 & 0.148 & 0.034 & 0.148 & 0.262 & 0.376 \\ -0.141 & -0.147 & 0.076 & 0.153 & -0.069 & 0.153 & 0.376 & 0.598 \end{pmatrix}$$

Para estudiar el apalancamiento se analizan los valores de la diagonal principal de H. Existen dos criterios. El primero, el de Hoaglin y Welsch (1978), los compara con el resultado que se obtiene con la operación $2 \cdot k/N$. En nuestro caso $2 \cdot 3/8 = 0.75$. Según este criterio ninguno de los individuos de la muestra presenta síntomas de ser una observación potencialmente influyente. El criterio alternativo es el de Huber (1981), que viene dado por la siguiente regla:

$$\begin{cases} \text{Si } \max(h_{ii}) \leq 0.20 & \Rightarrow & \text{No hay influencia potencial} \\ \text{Si } 0.2<\max(h_{ii})<0.5 & \Rightarrow & \text{Existe riesgo de influencia potencial} \\ \text{Si } \max(h_{ii}) \geq 0.5 & \Rightarrow & \text{Existe influencia potencial} \end{cases}$$

Según la matriz H, el primer individuo tiene un valor del apalancamiento igual a 0.607, siendo el valor máximo de toda la diagonal principal y ello implicaría que existen problemas de influencia potencial en el conjunto de datos.

EJERCICIO 3.22

Utilizando los datos del Ejercicio 3.21, y para una variable endógena determinada, se ha estimado el modelo con los 8 individuos y, posteriormente, se ha vuelto a estimar seleccionando únicamente a los siete últimos. Los parámetros estimados en ambos casos han sido los siguientes:

Con todos los datos: $\hat{\beta} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$ Eliminando el primer dato: $\hat{\beta}^{(1)} = \begin{pmatrix} 1.9 \\ 1.1 \\ 2.9 \end{pmatrix}$.

Además, se sabe que la suma de los cuadrados de los errores del modelo con todos los datos es igual a 10.

Estudie la posibilidad de que el individuo 1 de la muestra sea un individuo con influencia real.

Solución

Teniendo en cuenta la información del Ejercicio 3.21 junto con la que nos da el enunciado de este problema es posible calcular la distancia de Cook para el individuo primero de la muestra.

El valor de la distancia de Cook para el individuo 1 se calcula como

$$D_1 = \frac{\left(\hat{\beta} - \hat{\beta}^{(1)}\right)' (X'X) \left(\hat{\beta} - \hat{\beta}^{(1)}\right) / k}{\hat{\sigma}_u^2} \qquad (3.19)$$

Como todos los valores son conocidos, sustituyéndolos en (3.19) se obtiene que la distancia de Cook para el primer individuo es igual a:

$$D_1 = \frac{\begin{pmatrix}0.1 & -0.1 & 0.1\end{pmatrix}' \begin{pmatrix}8 & 25 & 30\\ 25 & 89 & 101\\ 30 & 101 & 124\end{pmatrix} \begin{pmatrix}0.1\\ -0.1\\ 0.1\end{pmatrix} \Big/ 3}{\frac{10}{8-3}} = 0.048$$

Para decidir si el valor de D_1 es estadísticamente significativo, a la hora de indicarnos si el individuo 1 ejerce una influencia real sobre la regresión del modelo tenemos que comparar dicho valor con una distribución F de Fisher-Snedecor de 3 y 5 grados de libertad. El valor que deja a su derecha una probabilidad del 5% en esta distribución es 5.41. Por tanto concluimos que no existe evidencia de que el individuo 1 sea un individuo con influencia real.

EJERCICIO 3.23

Se ha estimado un modelo con un tamaño muestral de $N = 10$, pero se ha considerado que la observación 6 puede estar produciendo problemas de influencia real. Bajo esta consideración, se han obtenido los siguientes valores predichos de la variable endógena:

Tabla 3.9

$\hat{Y}$	$\hat{Y}^{(6)}$
−15.4677	−5.4089
3.6896	2.7627
6.8911	4.1283
2.7559	2.3644
10.7094	5.7569
26.9589	12.6882
21.9506	10.5519
7.8280	4.5279
14.0940	7.2007
6.3876	3.9135

donde $\hat{Y}^{(6)}$ representa la predicción de la variable endógena, eliminando previamente, para la estimación del modelo, la observación 6.

Con la información suministrada contraste, mediante el contraste de Cook, si la observación número 6 tiene influencia real, sabiendo que el modelo sólo tiene una variable explicativa y que el sumatorio del cuadrado de sus errores es igual a 897.5609.

Solución

Para calcular la distancia de Cook hay que realizar previamente los cálculos que figuran en la Tabla 3.10.

Tabla 3.10

$\hat{Y}$	$\hat{Y}^{(6)}$	$\left(\hat{Y}-\hat{Y}^{(6)}\right)$	$\left(\hat{Y}-\hat{Y}^{(6)}\right)^2$
−15.4677	−5.4089	−10.0588	101.1795
3.6896	2.7627	0.9269	0.8591
6.8911	4.1283	2.7628	7.6331
2.7559	2.3644	0.3915	0.1533
10.7094	5.7569	4.9524	24.5263
26.9589	12.6882	14.2707	203.6529
21.9506	10.5519	11.3987	129.9304
7.8280	4.5279	3.3001	10.8907
14.0940	7.2007	6.8934	47.5190
6.3876	3.9135	2.4741	6.1212
		Suma =	532.465237

El valor de la distancia de Cook para el individuo 6 se puede calcular como

$$D_6 = \frac{\left(\hat{Y}-\hat{Y}^{(6)}\right)'\left(\hat{Y}-\hat{Y}^{(6)}\right)}{k\hat{\sigma}_u^2} \tag{3.20}$$

Sustituyendo los valores conocidos, se obtiene una distancia de Cook de

$$D_6 = \frac{\left(\hat{Y}-\hat{Y}^{(6)}\right)'\left(\hat{Y}-\hat{Y}^{(6)}\right)}{k\hat{\sigma}_u^2} = \frac{\left(\hat{Y}-\hat{Y}^{(6)}\right)'\left(\hat{Y}-\hat{Y}^{(6)}\right)}{k\cdot\dfrac{e'e}{N-k}} = \frac{532.465237}{2\cdot\dfrac{897.5609}{10-2}} = 2.3729$$

Como el valor tabulado para una F de Fisher-Snedecor con 2 y 8 grados de libertad, para un nivel de significación del 5%, es 4.459, el estadístico de Cook cae en la región de aceptación y no tenemos evidencia que nos haga sospechar que la observación sexta ejerza influencia real.

EJERCICIO 3.24

Con los datos del Ejercicio 3.12, estudie la posible presencia de *outliers* en el modelo en donde la variable Y viene explicada por la variable X_2.

Solución

La estimación del modelo propuesto, a partir de los datos del Ejercicio 3.12, es la siguiente:

$$\hat{Y}_i = 23.25131 + 8.412536 X_{2i}$$

A partir de dicho resultado podemos obtener todos los valores estimados para la variable endógena y, dado que conocemos sus valores reales, también podemos calcular el vector de errores —como diferencia entre el valor real y el valor estimado—. Los datos de los errores se recogen en la Tabla 3.11.

Tabla 3.11

Errores									
−0.298	−0.178	0.241	0.458	0.134	−0.128	−0.143	−0.118	0.097	0.239
0.126	−0.079	0.217	−0.326	0.175	−0.077	0.065	0.462	−0.487	−0.039
0.061	−0.647	−0.331	0.496	0.046	0.064	0.426	−0.147	−0.147	0.409
−0.203	0.518	−0.280	−0.163	−0.441					

A partir de esta serie es inmediato calcular su media —que toma el valor cero— y la *SCE* —que es igual a 2.9757—. Por tanto, el estimador de la varianza de la perturbación aleatoria vale

$$S_e^2 = \frac{e'e}{N} = \frac{2.9757}{35} = 0.085 \Rightarrow S_e = 0.29158$$

Para detectar la presencia de *outliers* necesitamos estandarizar los errores, dividiendo cada uno de ellos por la estimación de la desviación típica de la perturbación aleatoria.

Los datos de los errores estandarizados se recogen en la Tabla 3.12.

Tabla 3.12

Errores estandarizados									
−1.021	−0.611	0.825	1.569	0.458	−0.439	−0.491	−0.404	0.332	0.819
0.433	−0.273	0.743	−1.117	0.600	−0.265	0.222	1.584	−1.670	−0.133
0.209	−2.219	−1.134	1.702	0.159	0.218	1.461	−0.504	−0.504	1.402
−0.697	1.776	−0.961	−0.560	−1.512					

Para que exista un *outlier*, utilizando un nivel de significación del 1%, su error estandarizado debe ser superior a 2.5 en términos absolutos. Como se puede concluir de la Tabla 3.12, no parece que estemos en presencia de ninguna observación que pueda ser considerada *outlier*.

EJERCICIO 3.25

Al calcular los errores estandarizados del Ejercicio 3.24 y fijar el punto de corte en 2.5, se concluye que no había evidencia empírica a favor de la existencia de *outliers* entre los datos. Sin embargo, la Tabla 3.12 muestra un error estandarizado que supera en valor absoluto el 1.96 (valor crítico, para un nivel de significación del 5%, para un contraste de dos colas de una distribución normal estándar). Suponiendo que este valor pueda estar indicando que este individuo tiene un comportamiento diferencial con respecto al resto, ¿cómo habría que modificar el modelo para incluir la presencia del *outlier*?

Solución

En caso de considerar que una de las observaciones es un *outlier*, habría que crear una variable dicotómica que recogiera su carácter diferenciado. Esta variable tomaría siempre el valor cero, excepto para el individuo que presenta el error estandarizado alto (en este ejercicio es de -2.219), en cuyo caso la dicotómica tomaría el valor 1.

A continuación habría que estimar el modelo en el que Y estuviera en función de X_2 y de la variable dicotómica creada. Como la observación que presentaba un error estandarizado anormalmente alto era la observación 22, podemos llamar a la dicotómica $D22$. De esta forma, la nueva matriz de datos para estimar el modelo sería la que se muestra en la Tabla 3.13.

Tabla 3.13

Y	Regresor ficticio	X_2	$D22$
66.5562523	1	5.18306490	0
66.1628862	1	5.12208277	0
66.5469353	1	5.11795051	0
70.2076572	1	5.52731675	0
68.7278073	1	5.38991927	0
72.1692334	1	5.83009034	0
71.1843840	1	5.71482074	0
65.2835742	1	5.01038480	0
67.7667431	1	5.28004430	0
72.9124242	1	5.87482970	0
69.7311680	1	5.51007425	0
69.4277154	1	5.49844505	0
73.7211779	1	5.97359954	0
66.0939196	1	5.13143581	0
73.4523458	1	5.94661115	0

(*continúa en la página siguiente*)

Tabla 3.13 (continuación)

Y	Regresor ficticio	X_2	D22
68.2152258	1	5.35405473	0
70.1947336	1	5.57248788	0
69.6662676	1	5.46245345	0
69.7273432	1	5.58251257	0
70.3378558	1	5.60180349	0
72.0782670	1	5.79681218	0
72.5115581	1	5.93247535	1
69.1935676	1	5.50045602	0
69.2570862	1	5.40971063	0
73.0187767	1	5.91034769	0
68.9096920	1	5.41984926	0
70.6179649	1	5.57985620	0
66.0271491	1	5.10222971	0
72.3925089	1	5.85890540	0
70.4508941	1	5.56203475	0
69.2840803	1	5.49609042	0
68.6248823	1	5.33201888	0
66.0522571	1	5.12105412	0
68.9258414	1	5.44874499	0
72.4726603	1	5.90337383	0

La estimación del nuevo modelo ofrece el resultado del Cuadro 3.6.

Cuadro 3.6

Dependent Variable: Y
Method: Least Squares
Sample: 1 35
Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.65862	0.977677	23.175970	0.0000
X2	8.523684	0.177425	48.040930	0.0000
D22	−0.713603	0.294383	−2.424060	0.0212
R-squared	0.986922	Mean dependent var		69.654370
Adjusted R-squared	0.986105	S.D. dependent var		2.377813
S.E. of regression	0.280291	Akaike info criterion		0.375840
Sum squared resid	2.514020	Schwarz criterion		0.509156
Log likelihood	−3.577208	F-statistic		1207.449000
Durbin-Watson stat	2.282709	Prob(F-statistic)		0.000000

Como se puede observar, la variable dicotómica es significativa al 5%, pero no al 1%. En consecuencia, si trabajamos con un nivel de significación del 5%

el individuo 22 es considerado un *outlier*, que se caracteriza por reducir el valor medio de Y en 0.7136 unidades.

EJERCICIO 3.26

El Cuadro 3.7 contiene los valores observados y estimados para cuatro observaciones. Se sospecha que las mismas puedan ser *outliers*. Contraste la veracidad de dicha sospecha, a partir de los datos contenidos en el Cuadro 3.7 y sabiendo que la desviación típica de los errores es de 170594.

Cuadro 3.7

	Y	$\hat{Y}$
AGOSTO 1995	2382227	2285541
OCTUBRE 1995	1676466	1670428
JULIO 1998	3196895	2866337
NOVIEMBRE 1999	2626740	2182833

Solución

Para saber si se trata de *outliers* habrá que realizar un contraste de hipótesis con cada observación.

El contraste se plantea en los siguientes términos:

$$\left.\begin{array}{l} H_0\text{: La observación no es } outlier \\ H_1\text{: La observación sí es } outlier \end{array}\right\}$$

y se resuelve a través de la estandarización de los errores y su comparación con el valor crítico de una distribución Normal Estándar.

Los valores estandarizados de los errores están recogidos en el Cuadro 3.8.

Cuadro 3.8

	Error	Error estandarizado
AGOSTO 1995	96686	0.5668
OCTUBRE 1995	6038	0.0354
JULIO 1998	330558	1.9377
NOVIEMBRE 1999	443907	2.6021

Si comparamos los errores estandarizados con el valor crítico para un contraste de dos colas de la distribución Normal Estándar al 5% de nivel de significación (1.96), podemos considerar que la observación de noviembre de 1999 es *outlier*, mientras que las de agosto y octubre de 1995 y julio de 1998 no lo

son, es decir, sólo la observación correspondiente a noviembre de 1999 tiene un residuo MCO anormalmente alto.

EJERCICIO 3.27

Tras la estimación de un modelo por MCO en el programa Eviews, se obtienen los siguientes residuos mínimo cuadrático ordinarios. El Gráfico 3.1 recoge sus principales estadísticos descriptivos. Detecte la posible existencia de *outliers* a partir de estos datos y de los contenidos en la Tabla 3.14.

Gráfico 3.1

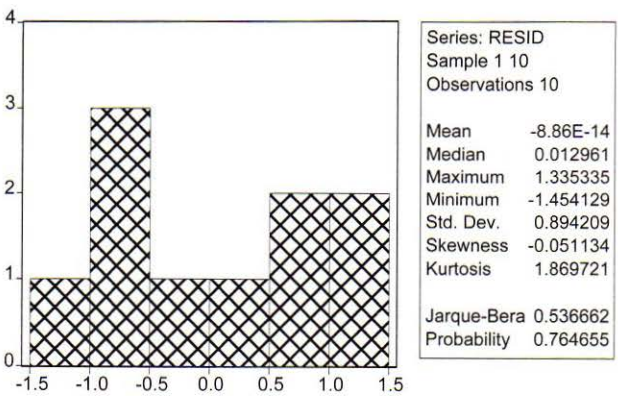


Tabla 3.14

Observación	e
1	–0.77
2	1.34
3	–0.68
4	–0.65
5	0.58
6	–0.18
7	0.21
8	–1.45
9	1.00
10	0.63

Solución

En primer lugar calculamos los residuos estandarizados, para lo que utilizamos la información relativa a la cuasi-desviación típica de los residuos proporcionada por el Gráfico 3.1, que nos indica que la misma vale

0.894209[3]. Ello implica que la desviación típica de los errores es igual a 0.8483 (obtenido como: $S_e = \sqrt{\tilde{S}_e^2 (N-1)/N}$). Estos residuos estandarizados están recogidos en la tercera columna de la Tabla 3.15.

Tabla 3.15

Observación	e	e/S_e
1	−0.77	−0.91
2	1.34	1.58
3	−0.68	−0.80
4	−0.65	−0.77
5	0.58	0.68
6	−0.18	−0.21
7	0.21	0.25
8	−1.45	−1.71
9	1.00	1.18
10	0.63	0.74

Las hipótesis nula y alternativa del contraste vienen dadas por:

$$\left.\begin{array}{l} H_0\text{: La observación no es } \textit{outlier} \\ H_1\text{: La observación sí es } \textit{outlier} \end{array}\right\}$$

y el estadístico de prueba bajo la hipótesis nula, $(e_i - \bar{e})/S_e$, se distribuye como una Normal Estándar.

Dado que no hay ninguna observación con un residuo MCO estandarizado con valor superior a 1.96 en valor absoluto (valor de la Normal Estándar al 5%), para dicho nivel de significación, no tenemos evidencia que nos haga sospechar de la presencia de ninguna observación atípica o *outlier*.

EJERCICIO 3.28

Dados los datos de la Tabla 3.16, calcule la serie de errores estandarizados y estudentizados del modelo que explica a la variable "gasto", como una función lineal sin restricciones de la variable "renta". A la luz de los resultados, comente si alguna observación puede ser considerada como *outlier*.

[3] Nótese que el programa Eviews, al proporcionar las estadísticas descriptivas de cualquier variable, en lugar de la desviación típica, proporciona la cuasi-desviación típica (Std. Dev).

Tabla 3.16

Gasto	Renta
8	10
25	30
14	20
35	40

Solución

El modelo para el cual se quieren obtener los residuos es el siguiente:

$$GASTO_i = \alpha + \beta \cdot RENTA_i + u_i$$

Como se puede observar, se trata de un modelo de regresión lineal simple, cuyo vector de parámetros es $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ y su estimación es $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1} X'Y$, siendo las matrices:

$$X'X = \begin{pmatrix} 4 & 100 \\ 100 & 3000 \end{pmatrix} \quad (X'X)^{-1} = \begin{pmatrix} 1.5 & -0.050 \\ & 0.002 \end{pmatrix} \quad X'Y = \begin{pmatrix} 82 \\ 2510 \end{pmatrix}$$

Con estas matrices se obtiene un vector de parámetros estimados igual a $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} -2.50 \\ 0.92 \end{pmatrix}$. Con estos valores de los parámetros estimados, y sabiendo que $\hat{Y} = X \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$ y que $e = Y - \hat{Y}$, es inmediato obtener el siguiente vector de errores:

$$e = \begin{pmatrix} 1.3 \\ -0.1 \\ -1.9 \\ 0.7 \end{pmatrix}$$

Los errores estandarizados se obtienen como e_i / S_e y los estudentizados como $e_i / \hat{\sigma}_u \sqrt{1 - h_{ii}}$, siendo h_{ii} el valor del apalancamiento o *leverage* de cada individuo, valores que se corresponden con la diagonal principal de la matriz $H = X(X'X)^{-1} X'$.

Los errores ya se han calculado, por lo que sólo faltaría su desviación típica y el estimador de la desviación típica de la perturbación aleatoria. Estos valores se calculan como

$$S_e = \sqrt{\frac{e'e}{N}} = \sqrt{\frac{1.3^2 + (-0.1)^2 + (-1.9)^2 + 0.7^2}{4}} = \sqrt{\frac{5.8}{4}} = 1.2$$

$$\hat{\sigma}_u = \sqrt{\frac{e'e}{N-k}} = \sqrt{\frac{1.3^2 + (-0.1)^2 + (-1.9)^2 + 0.7^2}{4-2}} = \sqrt{\frac{5.8}{2}} = 1.7$$

y la matriz H viene dada por

$$H = \begin{pmatrix} 0.7 & 0.1 & 0.4 & -0.2 \\ & 0.3 & 0.2 & 0.4 \\ & & 0.3 & 0.1 \\ & & & 0.7 \end{pmatrix}$$

Con esta información y las expresiones anteriores se obtienen los errores estandarizados y estudentizados que se muestran en la Tabla 3.17.

Tabla 3.17

e	e^2	e estandarizados	h_{ii}	e estudentizados
1.3	1.69	1.07959	0.7	1.39370
–0.1	0.01	–0.08305	0.3	–0.07020
–1.9	3.61	–1.57786	0.3	–1.33535
0.7	0.49	0.58132	0.7	0.75050

Para que estos errores sean significativamente distintos de cero para un nivel de significación del 5% deben ser, en el caso de los estandarizados, superiores en valor absoluto a 1.96, y en el caso de los estudentizados nos debemos fijar en el valor de la t-Student de $(N-k) = 4-2 = 2$ grados de libertad, que deja a su derecha una probabilidad del 2.5%. Ese valor es 4.303.

Como se puede observar, todos los residuos —tanto estandarizados como estudentizados— presentan valores inferiores en valor absoluto a los valores críticos, por lo que parece que ninguna de las observaciones puede ser considerada como *outlier* bajo ninguno de los dos criterios.

EJERCICIO 3.29

A partir del siguiente modelo de regresión

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, 2, ..., 63$$

y para llevar a cabo el análisis de *outliers*, influencia potencial y real, se ha obtenido la información de la Tabla 3.18, que reproduce la de los seis primeros individuos de la muestra.

Tabla 3.18

Observación	Distancia Cook	Distancia de Mahalanobis
1	1.47833	0.02843
2	2.26599	16.13000
3	1.08684	1.81120
4	2.55320	2.00000
5	4.64077	1.29540
6	2.78390	0.82320

(a) Determine qué observaciones ejercen influencia real.

(b) Determine cuáles ejercen influencia potencial.

Solución

(a) El estadístico de prueba para detectar qué observaciones ejercen influencia real es la distancia de Cook, que bajo la hipótesis nula se distribuye como una F de Fisher-Snedecor con k y $N-k$ grados de libertad. La hipótesis que se contrasta en este caso es

$$\left.\begin{array}{l} H_0: \beta = \beta^{(i)} \quad (\text{La observación excluida no es influyente}) \\ H_1: \beta \neq \beta^{(i)} \quad (\text{La observación excluida sí es influyente}) \end{array}\right\}$$

En este caso, al ser el valor crítico $F_{3,63-3}^{0.05} = 2.76$, sólo el estadístico de prueba correspondiente a la observación número 5 verifica que $F = 4.64077 > F_{3,63-3}^{0.05} = 2.76$ por lo que será ésa la única observación que ejerza influencia real. Esta influencia puede venir determinada no sólo por los *leverage* o "pesos" de las variables explicativas sino también por los valores de la endógena y tienen como consecuencia que estas observaciones tienen la capacidad de alterar sustancialmente las estimaciones o, en otras palabras, tienen capacidad para "tirar" hacia ellas o "escorar" la recta de regresión, llegando a tener, en ocasiones, residuos MCO incluso menores de los que les corresponderían por su situación en la nube de puntos.

(b) Para identificar las observaciones que ejercen influencia potencial hemos de calcular la medida de apalancamiento o *leverage* para cada una. Esta medida se calcula utilizando la siguiente expresión:

$$h_{ii} = \frac{1}{N}(1 + d_i)$$

siendo d_i la distancia de Mahalanobis.

El valor resultante se compara, siguiendo a Hoaglin y Welsch (1978), con el resultado que se obtiene de la operación $2 \cdot k/N$. En caso de obtener algún valor superior a éste, tendríamos que sospechar la presencia de alguna observación potencialmente influyente. En nuestro caso, el valor de comparación es $2 \cdot k/N = 2 \cdot 3/63 = 0.1$.

Los cálculos realizados están recogidos en la Tabla 3.19.

Tabla 3.19

Observación	Distancia Cook	Distancia Mahalanobis	*Leverage* h_{ii}
1	1.47833	0.02843	0.02
2	2.26599	16.13000	0.27
3	1.08684	1.81120	0.04
4	2.55320	2.00000	0.05
5	4.64077	1.29540	0.04
6	2.78390	0.82320	0.03

A la vista de estos resultados, la única observación que parece ejercer influencia potencial es la número 2, puesto que presenta una medida de apalancamiento, o *leverage*, superior al límite establecido de 0.1. Las razones de la existencia de esta observación con influencia potencial es que los valores que toman las variables explicativas para esta observación difieren sustancialmente de los valores medios. No obstante, estas observaciones por sí solas no tienen capacidad para alterar significativamente los resultados de la regresión.

EJERCICIO 3.30

A partir de los datos recogidos en la Tabla 3.20 se ha obtenido el siguiente modelo estimado:

$$\hat{Y}_i = 4.97 - 0.126X_i \qquad (3.21)$$

Tabla 3.20

Y	X	Distancia de Mahalanobis
1	10	2.30
2	1	1.05
4	2	0.50
5	5	0.01
10	5	0.01

(a) Determine si alguna observación tiene influencia potencial.

(b) Si al eliminar la observación 1 se obtiene la siguiente estimación:

$$\hat{Y}^{(1)}{}_i = 0.98 + 1.313X_i \tag{3.22}$$

determine si esta observación tiene influencia real.

Solución

(a) Para determinar si alguna observación tiene influencia potencial tenemos que comprobar si su medida de apalancamiento o *leverage* es superior a $2 \cdot k/N = 2 \cdot 2/5 = 0.8$. En ese caso, la observación tendría influencia potencial, es decir, los valores de sus variables explicativas distarían de los valores medios de las mismas, sin que ello suponga una capacidad de la observación de alterar los resultados de la estimación.

Obteniendo el *leverage* a través de la expresión

$$h_{ii} = \frac{1}{N}(1 + d_i)$$

comprobamos que el valor para cada observación es $h_{11} = 0.66$; $h_{22} = 0.41$; $h_{33} = 0.30$; $h_{44} = 0.20$ y $h_{55} = 0.20$, lo que indica que ninguna de ellas ejerce influencia potencial.

(b) El estadístico de la distancia de Cook nos permite determinar si una observación presenta influencia real. Éste se obtiene a través de la expresión

$$D_j = \frac{\sum_{i=1}^{N}\left(\hat{Y}_i - \hat{Y}_i^{(j)}\right)^2}{k\hat{\sigma}_u^2}$$

Los valores estimados de $\hat{Y}_i$ y de $\hat{Y}_i^{(j)}$ los obtenemos sustituyendo en los modelos (3.21) y (3.22) respectivamente. De esta forma obtenemos los datos recogidos en la Tabla 3.21.

Tabla 3.21

$\hat{Y}$	$\hat{Y}^{(1)}$	$\left(\hat{Y}-\hat{Y}^{(1)}\right)^2$
3.710	14.110	108.1600
4.844	2.293	6.5070
4.718	3.606	1.2365
4.340	7.545	10.2720
4.340	7.545	10.2720
	$\sum_{i=1}^{5}\left(\hat{Y}_i-\hat{Y}_i^{(1)}\right)^2=$	136.448

También necesitamos calcular la estimación de la varianza de la perturbación aleatoria. Ésta se obtendrá de la expresión siguiente:

$$\hat{\sigma}_u^2=\frac{Y'Y-\hat{\beta}'X'Y}{N-k}=\frac{146-\begin{pmatrix}4.97 & -0.126\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{N}Y_i\\ \sum_{i=1}^{N}X_iY_i\end{pmatrix}}{5-2}=$$

$$=\frac{146-\begin{pmatrix}4.97 & -0.126\end{pmatrix}\begin{pmatrix}22\\ 95\end{pmatrix}}{5-2}=\frac{146-97.37}{5-2}=16.21$$

Por tanto, la distancia de Cook para la observación primera valdrá

$$D_1=\frac{\sum_{i=1}^{N}\left(\hat{Y}_i-\hat{Y}_i^{(1)}\right)^2}{k\hat{\sigma}_u^2}=\frac{136.448}{2\cdot 16.21}=4.2$$

El valor crítico tabulado para una distribución F de Fisher-Snedecor con 2 y 3 grados de libertad al 95% de nivel de confianza es 9.55. Por tanto, el estadístico cae en la región de aceptación y no podemos rechazar la hipótesis nula. De esta manera, no tenemos evidencia que nos permita afirmar que la observación primera ejerce influencia real. No obstante, y si miramos los resultados de las estimaciones que incluyen y excluyen la observación 1, podríamos haber sospechado que dicha observación tenía influencia real debido a que los coeficientes estimados son realmente diferentes en signo y valor. El resultado inesperado del contraste se debe al escaso tamaño muestral y se encuentra afectado por el hecho de que para las estimaciones únicamente contamos con 5 y 4 observaciones respectivamente.

EJERCICIO 3.31

Dado el modelo de regresión lineal múltiple $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ para el que se dispone de la información muestral de la Tabla 3.22, donde d_i es la distancia de Mahalanobis:

Tabla 3.22

Y	X_2	X_3	d_i
11.191	3.1420	2.113	0.43880
15.382	3.3450	1.980	0.20390
0.391	0.0710	0.053	0.43629
19.725	5.0910	2.785	0.07418
7.150	15.5930	5.476	7.87184
0.114	0.0310	0.021	0.44189
1.425	0.2630	0.100	0.41629
6.724	1.4610	0.866	0.28899
29.690	6.7560	3.960	0.33745
189.669	33.0430	17.622	7.49037

$$X'X = \begin{pmatrix} 10.00 & 68.80 & 34.98 \\ 68.80 & 1429.81 & 723.16 \\ 34.98 & 723.16 & 373.11 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 0.14958 & -0.00534 & -0.00368 \\ -0.00534 & 0.03568 & -0.06866 \\ -0.00368 & -0.06866 & 0.13610 \end{pmatrix} \qquad e'e = 485.41$$

(a) Determine qué observaciones tienen influencia potencial.

(b) Si una vez eliminada de la muestra la observación con mayor influencia potencial la estimación del vector de los coeficientes es la siguiente:

$$\hat{\beta}' = \begin{pmatrix} -1.395 & 14.744 & -16.867 \end{pmatrix}$$

indique si esta observación presenta influencia real, utilizando para ello la distancia de Cook.

(c) ¿Qué consecuencias tiene la existencia de observaciones con influencia real en la estimación por MCO? ¿Qué habría que hacer para solucionar la posible influencia real?

Solución

(a) La medida descriptiva que detecta la influencia potencial es el *leverage* (h_{ii}). Ésta se obtiene a través de la siguiente expresión:

$$h_{ii} = \frac{1}{N}(1 + d_i)$$

donde d_i representa la distancia de Mahalanobis para el individuo *i*.

El valor que se obtenga para cada individuo, que siempre estará acotado entre 0 y 1, nos indicará si esa observación está ejerciendo influencia potencial o no.

Según Hoaglin y Welsch (1978), cuando ese valor sea mayor que $2 \cdot k/N$ estamos ante una observación potencialmente influyente. En cambio Huber (1981) observa únicamente el mayor de todos los *leverage* y, en función de su resultado, decide si la muestra tiene alguna observación potencialmente influyente o no, estableciendo el límite que asegura la presencia de este tipo de observaciones en el 0.5.

En este caso, los diferentes valores de *leverage* son los recogidos en la Tabla 3.23:

Tabla 3.23

d_i	h_{ii}
0.43880	0.143880
0.20390	0.120390
0.43629	0.143629
0.07418	0.107418
7.87184	0.887184
0.44189	0.144189
0.41629	0.141629
0.28899	0.128899
0.33745	0.133745
7.49037	0.849037

Según el criterio de Hoaglin y Welsch, que establecen el límite en $2k/N = 2 \cdot 3/10 = 0.6$, las observaciones quinta y décima son potencialmente influyentes.

De la misma forma, siguiendo el criterio de Huber, dado que en la muestra existen observaciones con *leverage* superiores a 0.5, podemos afirmar que

existe influencia potencial para las mismas observaciones de la muestra detectadas con el criterio de Hoaglin y Welsch.

(b) Para detectar la existencia de influencia real habría que contrastar la siguiente hipótesis nula:

$$H_0: \beta = \beta^{(i)}$$

Siendo $\beta^{(i)}$ el vector de estimadores del modelo una vez eliminada la observación *i*-ésima.

Para resolver dicho contraste utilizaremos el estadístico de la distancia de Cook:

$$D_i = \frac{\left(\hat{Y} - \hat{Y}^{(i)}\right)'\left(\hat{Y} - \hat{Y}^{(i)}\right)}{k\hat{\sigma}_u^2} = \frac{\left(\hat{\beta} - \hat{\beta}^{(i)}\right)'(X'X)\left(\hat{\beta} - \hat{\beta}^{(i)}\right)}{k\hat{\sigma}_u^2} \qquad (3.23)$$

Los únicos valores que desconocemos de dicha expresión son $\hat{\beta}$ y $\hat{\sigma}_u^2$, por lo que procedemos a calcularlos.

En primer lugar necesitamos calcular $X'Y$, para poder obtener el valor de $\hat{\beta}$

$$X'Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 3.142 & 3.345 & \cdots & 33.043 \\ 2.113 & 1.980 & \cdots & 17.622 \end{pmatrix} \begin{pmatrix} 11.191 \\ 15.382 \\ \vdots \\ 189.669 \end{pmatrix} = \begin{pmatrix} 281.461 \\ 6776.573 \\ 3614.099 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y =$$

$$= \begin{pmatrix} 0.14958 & -0.00534 & -0.00368 \\ -0.00534 & 0.03568 & -0.06866 \\ -0.00368 & -0.06866 & 0.13610 \end{pmatrix} \begin{pmatrix} 281.461 \\ 6776.573 \\ 3614.099 \end{pmatrix} = \begin{pmatrix} -7.358 \\ -7.841 \\ 25.573 \end{pmatrix} \qquad (3.24)$$

Por su parte $\hat{\sigma}_u^2$ valdrá:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{485.41}{10-3} = 69.344 \qquad (3.25)$$

Sustituyendo (3.24) y (3.25) en (3.23) obtenemos el valor del estadístico:

$$D_i = \frac{1}{3 \cdot 69.344}\left[\left[\begin{pmatrix} -7.358 \\ -7.841 \\ 25.573 \end{pmatrix} - \begin{pmatrix} -1.395 \\ 14.744 \\ -16.867 \end{pmatrix}\right]' \cdot \right.$$

$$\left. \cdot \begin{pmatrix} 0.14958 & -0.00534 & -0.00368 \\ -0.00534 & 0.03568 & -0.06866 \\ -0.00368 & -0.06866 & 0.13610 \end{pmatrix}\left[\begin{pmatrix} -7.358 \\ -7.841 \\ 25.573 \end{pmatrix} - \begin{pmatrix} -1.395 \\ 14.744 \\ -16.867 \end{pmatrix}\right]\right] =$$

$$= \frac{16219.04}{208.03} = 77.96$$

Comparando el valor del estadístico con el valor tabulado para la distribución $F_{k,N-k}$ ($F_{3,7}^{0.05} = 4.3468$), observamos que el valor del estadístico cae en la región de rechazo, por lo que no podemos aceptar que los estimadores $\hat{\beta}$ sean iguales con o sin esa observación, lo que implica que la quinta observación parece ejercer una influencia real sobre la estimación del modelo.

(c) La existencia de una observación con influencia real en un modelo implica que dicha observación puede llegar a cambiar los resultados de la estimación. De hecho, una única observación con influencia real podría llegar a ser responsable de que determinadas variables explicativas resulten no significativas, del cambio de signo en los estimadores, de la falta de linealidad en el ajuste, de la no normalidad de los residuos y, en definitiva, del deterioro de la capacidad predictiva de un modelo.

En caso de detectar la existencia de una observación con influencia real habría que revisar en primer lugar los datos de origen, no sea que se trate sólo de una errata a la hora de introducir los mismos. Si se comprueba que el dato es correcto, cabe la posibilidad de probar la estimación del modelo con formas funcionales alternativas o de introducir una variable ficticia que recoja el efecto de la observación "problemática". Por último, si esto no soluciona el problema, se puede optar, siempre que sea posible, por eliminar dicha observación del modelo.

EJERCICIO 3.32

Dado el modelo de regresión lineal múltiple

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

se dispone de la información muestral de la Tabla 3.24

Tabla 3.24

Y	X_2	X_3	d_i
96	56	1	1.755
22	3	6	1.123
42	18	2	1.988
21	1	7	1.656
104	58	9	4.318
124	74	6	0.972
128	77	3	0.221
163	198	1	6.633
123	72	3	0.191
40	18	3	1.142

$$(X'X)=\begin{pmatrix} 10 & 575 & 41 \\ 575 & 62951 & 1782 \\ 41 & 1782 & 235 \end{pmatrix} \quad (X'X)^{-1}=\begin{pmatrix} 0.69638 & -0.00372 & -0.09329 \\ -0.00372 & 0.00004 & 0.00034 \\ -0.09329 & 0.00034 & 0.01792 \end{pmatrix}$$

donde d_i es la distancia de Mahalanobis.

(a) Determine qué observaciones tienen influencia potencial.

(b) Si una vez eliminada de la muestra la observación i con mayor influencia potencial, la estimación del vector de coeficientes es:

$$\hat{\beta}^{(i)}=\begin{pmatrix} 14.02 \\ 1.46 \\ 0.61 \end{pmatrix}$$

indique si esta observación tiene influencia real, calculando para ello la distancia de Cook.

Solución

(a) Para determinar qué observación tiene influencia potencia tenemos que calcular, en primer lugar, el límite aceptable de apalancamiento o *leverage* establecido por Hoaglin y Welsch. En este caso es

$$\frac{2k}{N}=2\cdot\frac{3}{10}=0.6.$$

A continuación obtenemos el *leverage* a través de la expresión siguiente

$$h_{ii} = \frac{1}{N}(1 + d_i)$$

Los resultados, recogidos en la Tabla 3.25, muestran que la sexta observación presenta un *leverage* superior al límite establecido, por lo que podemos hablar de la presencia de una observación con influencia potencial.

Tabla 3.25

d_i	h_{ii}
1.755	0.2755
1.123	0.2123
1.988	0.2988
1.656	0.2656
4.318	0.5318
0.972	0.1972
0.221	0.1221
6.633	**0.7633**
0.191	0.1191
1.142	0.2142

(b) El contraste a realizar para determinar si una observación tiene influencia real es el siguiente:

$$\left.\begin{aligned} H_0&: \beta = \hat{\beta}^{(i)} \\ H_1&: \beta \neq \hat{\beta}^{(i)} \end{aligned}\right\}$$

El estadístico de contraste que nos permite decidirnos por una hipótesis o la otra es el siguiente:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}^{(i)}\right)' (X'X)\left(\hat{\beta} - \hat{\beta}^{(i)}\right)}{k\hat{\sigma}_u^2}$$

A continuación calculamos cada uno de los elementos de dicho estadístico

$$X'Y = \begin{pmatrix} 863 \\ 73133 \\ 3175 \end{pmatrix} \Rightarrow \hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 32.73 \\ 0.79 \\ 1.25 \end{pmatrix} \Rightarrow \hat{\beta} - \hat{\beta}^{(i)} = \begin{pmatrix} 18.71 \\ -0.67 \\ 0.64 \end{pmatrix}$$

$$Y'Y = 97779 \quad \hat{\beta}'X'Y = 90318.23 \quad \Rightarrow \quad e'e = Y'Y - \hat{\beta}'X'Y = 7460.77$$

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{7460.77}{10-3} = 1065.82$$

Por tanto, la distancia de Cook valdrá:

$$D_i = \frac{\begin{pmatrix} 18.71 & -0.67 & 0.64 \end{pmatrix} \begin{pmatrix} 10 & 575 & 41 \\ 575 & 62951 & 1782 \\ 41 & 1782 & 235 \end{pmatrix} \begin{pmatrix} 18.71 \\ -0.67 \\ 0.64 \end{pmatrix}}{3 \cdot 1065.82} =$$

$$= \frac{16624.87}{3 \cdot 1065.82} = 5.20$$

Como este valor es superior al valor tabulado para una F de Fisher-Snedecor con 3 y 17 grados de libertad (que es 3.1968), no tenemos evidencia que nos permita aceptar la hipótesis nula y, por tanto, consideramos que la observación eliminada sí tenía influencia real.

EJERCICIO 3.33

Se estima el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ obteniendo los siguientes resultados:

$$\hat{Y}_i = \underset{(32.04631)}{75.47912} + \underset{(0.375947)}{1.272025} X_{2i} - \underset{(1.742295)}{1.969836} X_{3i}$$

(a) Si conoce los valores de la distancia de Cook para cada observación, recogidos en la Tabla 3.26, ¿podría decir que existe alguna observación con influencia real? ¿Cómo se distribuye el estadístico de prueba?

Tabla 3.26

Observación	Distancia Cook
1	0.004
2	0.173
3	0.029
4	0.017
5	0.036
6	6.155
7	0.004
8	0.065
9	1.361

(b) Si además se le facilita la matriz de correlaciones del Cuadro 3.9, para cuyo cálculo se han tenido en cuenta todas las observaciones,

Cuadro 3.9

Matriz Correlaciones

	Y	*K*	*L*
Y	1	0.886	0.698
K	0.886	1	0.888
L	0.698	0.888	1

¿puede afirmar que existe algún otro problema con los datos? Asimismo, ¿cuál sería el coeficiente de determinación de las regresiones auxiliares? Calcule el *FIV* y comente el resultado.

(c) Al estimar el modelo nuevamente incluyendo una variable dicotómica (*IREAL*) que recoge el efecto de la observación con influencia real, los resultados del modelo estimado son los que siguen:

$$\underset{}{\hat{Y}_i} = \underset{(13.50680)}{84.63318} + \underset{(0.301655)}{2.665914}\, X_{2i} - \underset{(1.467502)}{8.866817}\, X_{3i} - \underset{(32.10475)}{173.8205}\, IREAL_i$$

Comente los resultados.

Solución

(a) Las hipótesis nula y alternativa del contraste a realizar son respectivamente:

$$\left.\begin{array}{l} H_0: \beta = \hat{\beta}^{(i)} \\ H_1: \beta \neq \hat{\beta}^{(i)} \end{array}\right\}$$

El estadístico de prueba se distribuye bajo la hipótesis nula como una F de Fisher-Snedecor con k y $N-k$ grados de libertad. Dado que el valor crítico en tablas es $F_{3,9-3}^{0.05} = 4.76$, la única observación que presenta influencia real es la número 6.

(b) La matriz de correlaciones indica que existe una correlación lineal elevada entre las variables exógenas del modelo, ya que $r_{X_2,X_3} = 0.888$, lo que indica la posible existencia de multicolinealidad aproximada entre las variables.

Como en este caso sólo hay dos variables explicativas, la regresión auxiliar posible sería la del siguiente modelo de regresión lineal simple (MRLS)

$$X_{2i} = \alpha_1 + \alpha_2 X_{3i} + \nu_i$$

y, por lo tanto, podemos calcular el *FIV* a partir del coeficiente de determinación de esta regresión auxiliar como sigue:

$$R_j^2 = \left(r_{X_2,X_3}\right)^2 = 0.888^2 = 0.788$$

El *FIV* se obtendrá a partir de la expresión

$$FIV\left(X_j\right)=\frac{1}{1-R_j^2}=\frac{1}{1-0.788}=4.71$$

lo que implica que la varianza del estimador se ve multiplicada por 4.71, en comparación con la obtenida en ausencia de multicolinealidad.

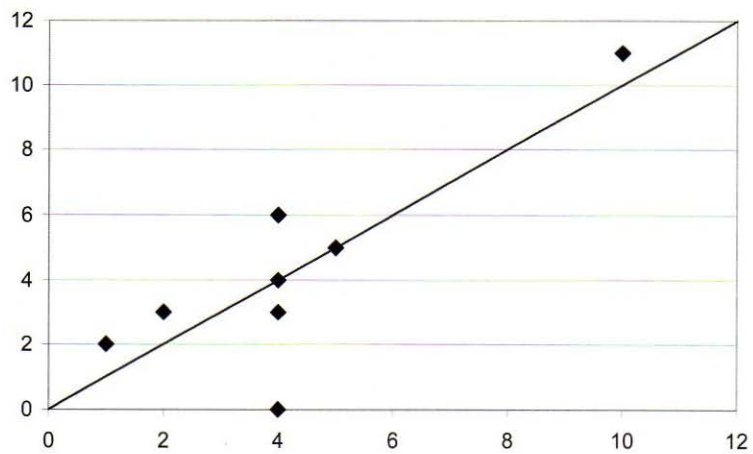
(c) Los resultados muestran que los coeficientes estimados varían sustancialmente antes y después de introducir la variable *IREAL*, que recoge el efecto de una observación que ejerce influencia real en la estimación.

El efecto es tan fuerte que, si dicha dicotómica no es incluida, dicha observación "escora" la recta de regresión, de modo que es capaz de alterar el punto de corte de la recta e incluso de hacer pasar la variable X_3 de no significativa a significativa.

EJERCICIO 3.34

A partir de los datos contenidos en el Gráfico 3.2, donde todos los valores de X e Y son números enteros, se realiza una regresión de la variable Y contra X.

Gráfico 3.2



(a) A la vista del Gráfico 3.2, ¿cree que alguna observación tendrá influencia potencial? ¿Cuál de ellas y por qué? Verifique cuantitativamente la existencia de influencia potencial y/o real de dicha observación sabiendo que la distancia de Mahalanobis es 4.67551 y la de Cook 1.80725.

(b) Calcule el valor del mayor error cometido en la estimación, así como el valor de dicho residuo estandarizado ¿Se puede considerar un *outlier*?

Solución

(a) La observación (10,11) puede tener influencia potencial, porque está muy alejada del valor medio de X. Esta observación, sin embargo, no distorsiona los resultados de la regresión.

Para determinar cuantitativamente si tiene influencia potencial calculamos su *leverage*:

$$h_{ii} = \frac{1}{N}(1+d_i) = \frac{1}{8}(1+4.67551) = 0.709$$

Como el valor obtenido supera el límite establecido por Hoaglin y Welsch de $2k/N = 2 \cdot 2/8 = 0.5$, podemos considerar que la observación presenta influencia potencial.

En cuanto a la influencia real, habrá que decidir en función de la distancia de Cook. Dado que ésta es de 1.80725 y el valor crítico al 95% de una distribución F de Fisher-Snedecor con 2 y 6 grados de libertad es 5.14, no podemos rechazar la hipótesis nula, por lo que, en cambio, consideramos que la observación (10,11) no ejerce influencia real.

(b) La observación con mayor error es la coordenada (4,0). Para ésta, su error es de

$$e_{máx} = 0 - 4 = -4$$

Para comprobar si, dado este error, podemos considerar a esta observación como un *outlier*, tendremos que calcular su residuo estandarizado. De esta forma podremos contrastar la hipótesis nula que, de forma general, considera que el residuo *i*-ésimo no es *outlier*. La información necesaria para el cálculo está contenida en la Tabla 3.27, de donde se deriva que $e'e = 24$.

Tabla 3.27

Observación	Y	X	e	e^2
1	0	4	–4	16
2	2	1	1	1
3	3	2	1	1
4	3	4	–1	1
5	4	4	0	0
6	5	5	0	0
7	6	4	2	4
8	11	10	1	1

El estadístico de contraste, que bajo la hipótesis nula se distribuye como una $N(0,1)$, valdrá

$$\frac{e_i}{S_e} = \frac{-4}{\sqrt{\frac{e'e}{N}}} = \frac{-4}{\sqrt{\frac{24}{8}}} = -2.31$$

Como el estadístico cae en zona de rechazo, ya que el valor crítico correspondiente a una $N(0,1)$ al 95% de nivel de confianza es ± 1.96, rechazamos la hipótesis nula, por lo que la observación (4,0) se considera *outlier*.

EJERCICIO 3.35

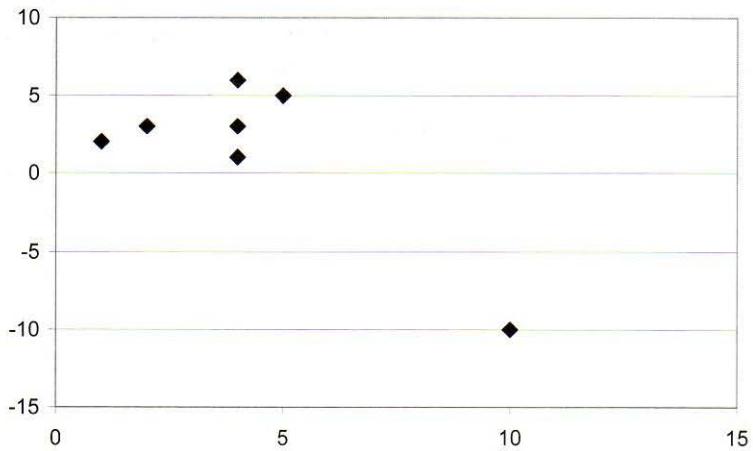
A partir de los datos de la Tabla 3.28:

Tabla 3.28

Y	X
1	4
2	1
3	2
3	4
5	5
6	4
–10	10

que generan el siguiente gráfico de dispersión:

Gráfico 3.3



se obtienen las siguientes estimaciones:

Cuadro 3.10

Dependent Variable: Y
Method: Least Squares
Sample: 1 7
Included observations: 7

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	−1.4	0.535934	−2.637071	0.0461
C	7.5	2.702539	2.769821	0.0394
R-squared	0.581735	Mean dependent var		1.428571
Adjusted R-squared	0.498082	S.D. dependent var		5.318432
S.E. of regression	3.767905	Akaike info criterion		5.725872
Sum squared resid	70.985550	Schwarz criterion		5.710418

Cuadro 3.11

Dependent Variable: Y
Method: Least Squares
Sample: 1 6
Included observations: 6

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.5	0.551618	1.013063	0.3683
C	1.5	1.988886	0.739403	0.5007
R-squared	0.204186	Mean dependent var		3.333333
Adjusted R-squared	0.005232	S.D. dependent var		1.861899
S.E. of regression	1.857022	Akaike info criterion		4.337026
Sum squared resid	13.794120	Schwarz criterion		4.267613

(a) A partir del Gráfico 3.3, ¿cree que la observación nº 7 de la muestra tiene influencia potencial? ¿Y real? Argumente su respuesta sin realizar cálculos.

(b) Calcule el estadístico de contraste que permita determinar si la observación nº 7 de la muestra tiene o no influencia real.

(c) Posteriormente se sabe que el *leverage* correspondiente a la observación 7 es de 0.66. ¿Tiene esta observación influencia potencial?

(d) ¿Es la observación 7 un *outlier*? Realice el cálculo necesario para responder a esta pregunta. A la vista de los resultados, comente si esta observación es o no problemática y, si existiese problema, indique cómo podría resolverlo.

Solución

(a) Según el Gráfico 3.3, la observación (10, −10) debe presentar influencia potencial y real. Presenta influencia potencial, ya que el valor de $X = 10$ está muy alejado de la media. En cuanto a la influencia real, esta observación distorsiona el modelo, empeorando su ajuste.

(b) Para determinar si la observación séptima tiene influencia real, tendremos que calcular la distancia de Cook mediante la expresión

$$D_7 = \frac{\sum_{i=1}^{N}\left(\hat{Y}_i - \hat{Y}_i^{(7)}\right)^2}{k\hat{\sigma}_u^2}$$

Calculando los valores necesarios, que están recogidos en la Tabla 3.29,

Tabla 3.29

Y	$\hat{Y}$	$\hat{Y}^{(7)}$	$\left(\hat{Y} - \hat{Y}^{(7)}\right)^2$
4	1.9	3.5	2.56
1	6.1	2.0	16.81
2	4.7	2.5	4.84
4	1.9	3.5	2.56
5	0.5	4.0	12.25
4	1.9	3.5	2.56
10	−6.5	6.5	169.00
	$\sum_{i=1}^{7}\left(\hat{Y}_i - \hat{Y}_i^{(7)}\right)^2 =$		210.58

obtenemos el valor del estadístico

$$D_7 = \frac{\sum_{i=1}^{N}\left(\hat{Y}_i - \hat{Y}_i^{(7)}\right)^2}{k\hat{\sigma}_u^2} = \frac{210.58}{2 \cdot 3.7679^2} = 7.416$$

Como el valor crítico de la distribución F de Fisher-Snedecor con 2 y 5 grados de libertad al 95% de nivel de confianza es 5.79, rechazamos la hipótesis nula y consideramos que la observación (10, −10) tiene influencia real.

(c) El cálculo del límite de apalancamiento es $2k/N = 2 \cdot 2/7 = 0.57$. Como el *leverage* de la observación séptima es superior a este límite (según el enunciado el *leverage* es de 0.66), tenemos que considerar que la observación (10, −10) también tiene influencia potencial.

(d) Por último, para determinar si la observación (10, –10) puede ser considerada un *outlier*, realizamos el contraste correspondiente

$$\left.\begin{array}{l} H_0\text{: La observación no es } outlier \\ H_1\text{: La observación sí es } outlier \end{array}\right\}$$

el estadístico de prueba es

$$\frac{e_7 - \overline{e}}{S_e} = \frac{\left(Y_7 - \hat{Y}_7\right) - \overline{e}}{S_e} = \frac{\left(-10 - \left(-6.5\right)\right) - 0}{\sqrt{\frac{70.985}{7}}} = \frac{-3.5}{\sqrt{10.14}} = -1.099$$

El valor crítico correspondiente a una N(0,1) en un contraste bilateral al 95% de nivel de confianza es ± 1.96. Como el valor del estadístico cae en la región de aceptación del contraste, no podemos rechazar la hipótesis nula. Por tanto, la observación séptima no debe ser considerada como *outlier*, debido precisamente a esa capacidad que tiene de "escorar" la recta de regresión hacia ella al tener influencia real.

Como conclusión, al haber comprobado que la observación séptima tiene influencia real, ésta está distorsionando el modelo. Por tanto, si es posible, debemos eliminar la misma del modelo, o bien debemos introducir en el modelo una variable ficticia que recoja el comportamiento particular de esta observación.

EJERCICIO 3.36

Para un conjunto de diez familias se ha estimado el consumo de un determinado artículo en función del precio del artículo, así como de la renta de la familia. Los valores obtenidos en la muestra fueron los de la Tabla 3.30.

Tabla 3.30

Observaciones	*CONSUMO*	*PRECIO*	*RENTA*
1	97.160	14.990	118.20
2	96.510	11.470	112.40
3	38.830	9.542	37.30
4	16.000	9.298	155.50
5	194.300	10.540	246.40
6	5.198	14.360	8.30
7	97.260	10.080	112.60
8	123.100	8.873	154.30
9	29.260	9.518	33.65
10	86.030	15.940	106.40

Se estima el modelo econométrico por MCO

$$CONSUMO_i = \beta_1 + \beta_2 PRECIO_i + \beta_3 RENTA_i + u_i$$

obteniendo los resultados del Cuadro 3.12:

Cuadro 3.12

Dependent Variable: *CONSUMO*
Method: Least Squares
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–28.110280	63.946840	–0.439588	0.6735
PRECIO	2.818749	4.862355	0.579709	0.5803
RENTA	0.683555	0.182926	3.736776	0.0073
R-squared	0.666605	Mean dependent var		78.364800
Adjusted R-squared	0.571349	S.D. dependent var		57.501240
S.E. of regression	37.646880	Akaike info criterion		10.337700
Sum squared resid	9921.012000	Schwarz criterion		10.428480
Log likelihood	–48.688510	F-statistic		6.998058
Durbin-Watson stat	2.642649	Prob(F-statistic)		0.021397

(a) Obtenga la influencia potencial de cada familia.

(b) Estudie los residuos de la regresión, obteniendo los valores de los errores estandarizados y estudentizados.

(c) Obtenga la influencia real de aquellos valores calificados como *outliers*.

(d) Obtenga los errores estudentizados con omisión teniendo en cuenta cada uno de los *outliers* detectados.

Solución

(a) Para obtener la influencia potencial de cada individuo es necesario obtener la matriz H a través de la expresión:

$$H = X\left(X'X\right)^{-1}X'$$

Calculándola obtenemos los siguientes valores:

$$H = \begin{vmatrix} 0.32 & 0.10 & -0.06 & 0.00 & 0.14 & 0.21 & 0.02 & -0.02 & -0.07 & 0.37 \\ 0.10 & 0.10 & 0.09 & 0.10 & 0.11 & 0.09 & 0.10 & 0.10 & 0.09 & 0.10 \\ -0.06 & 0.09 & 0.32 & 0.10 & -0.13 & 0.17 & 0.15 & 0.12 & 0.32 & -0.08 \\ 0.00 & 0.10 & 0.10 & 0.20 & 0.24 & -0.07 & 0.15 & 0.22 & 0.10 & -0.04 \\ 0.14 & 0.11 & -0.13 & 0.24 & 0.53 & -0.21 & 0.11 & 0.24 & -0.14 & 0.10 \\ 0.21 & 0.09 & 0.17 & -0.07 & -0.21 & 0.40 & 0.04 & -0.08 & 0.18 & 0.26 \\ 0.02 & 0.10 & 0.15 & 0.15 & 0.11 & 0.04 & 0.13 & 0.15 & 0.15 & 0.00 \\ -0.02 & 0.10 & 0.12 & 0.22 & 0.24 & -0.08 & 0.15 & 0.23 & 0.12 & -0.07 \\ -0.07 & 0.09 & 0.32 & 0.10 & -0.14 & 0.18 & 0.15 & 0.12 & 0.33 & -0.08 \\ 0.37 & 0.10 & -0.08 & -0.04 & 0.10 & 0.26 & 0.00 & -0.07 & -0.08 & 0.43 \end{vmatrix}$$

Los elementos de la diagonal principal de esta matriz H (h_{ii}) son los denominados *leverage* o apalancamiento, y reflejan la influencia potencial del individuo *i*. Siguiendo la regla de decisión de Hoaglin y Welsch (1978) y utilizando el doble del valor medio, éste nos queda:

$$\frac{2k}{N} = \frac{2 \cdot 3}{10} = 0.6$$

Si hubiese algún h_{ii} con un valor superior a 0.6, la observación *i*-ésima sería potencialmente influyente. Podemos observar que no se da este caso para ninguna observación, por lo que diríamos que ningún elemento tendría influencia potencial elevada individualmente.

Si hubiésemos seguido el criterio de Huber (1981), habríamos concluido que tenemos influencia potencial asegurada, ya que el valor máximo del *leverage* es 0.53 (superior al 0.5 que establece el límite inferior para asegurar la existencia de influencia potencial).

(b)

Tabla 3.31

Observaciones	Consumo	Consumo estimado	Errores	Errores estandarizados	Errores estudentizados
1	97.160	94.94	2.22	0.07	0.07
2	96.510	81.05	15.46	0.49	0.43
3	38.830	24.28	14.55	0.46	0.47
4	16.000	104.39	–88.39	–2.81	–2.63
5	194.300	170.03	24.27	0.77	0.94
6	5.198	18.04	–12.84	–0.41	–0.44
7	97.260	77.27	19.99	0.63	0.57
8	123.100	102.37	20.73	0.66	0.63
9	29.260	21.72	7.54	0.24	0.25
10	86.030	89.55	–3.52	–0.11	–0.12

Para obtener los residuos estandarizados hay que utilizar la expresión:

$$\frac{e_i}{S_e} = \frac{\left(Y_i - \hat{Y}_i\right)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}} \quad \text{donde} \quad S_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} = \sqrt{\frac{9921.01}{10}} = 31.49763723$$

Estos residuos se distribuyen como una normal estandarizada. El valor crítico para la misma, para un nivel de significación del 5% en un contraste bilateral es 1.96. Vemos, por tanto, que para la observación cuarta (familia 4) tiene un residuo estandarizado superior a este límite $\left(\left|-2.81\right|\right)$, con lo cual éste cae en la región crítica, y se considera que esta familia representa un *outlier*.

Para obtener los residuos estudentizados hay que utilizar la expresión:

$$\frac{e_i}{S_{e_i}} = \frac{\left(Y_i - \hat{Y}_i\right)}{\sqrt{\hat{\sigma}_u^2\left(1 - h_{ii}\right)}} \quad \text{donde} \quad \hat{\sigma}_u^2 = \frac{1}{N-k}\sum_{i=1}^{N} e_i^2 = \frac{9921.01}{7} = 1417.28714$$

Estos residuos se distribuyen como una *t*-Student con 7 grados de libertad. El valor crítico entonces para un nivel de significación del 5% en un contraste bilateral será de 1.895. Se vuelve a cumplir entonces que la observación cuarta (familia 4) representa un *outlier*, ya que presenta un residuo estudentizado superior a este límite $\left(\left|-2.63\right|\right)$.

(c) Para obtener la influencia real del individuo 4 es necesario calcular el estadístico de la distancia de Cook.

$$D_i = \frac{\left(\hat{Y} - \hat{Y}^{(i)}\right)'\left(\hat{Y} - \hat{Y}^{(i)}\right)}{k\sigma_u^2}$$

Para obtener éste, es necesario obtener los valores estimados para toda la muestra, $\hat{Y}^{(i)}$, a través de los estimadores del modelo obtenido una vez eliminado el elemento 4. La estimación del modelo sin el elemento 4 nos da el resultado del Cuadro 3.13.

Cuadro 3.13

Dependent Variable: *CONSUMO*
Method: Least Squares
Sample: 1 3 5 10
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.278670	7.164445	1.574256	0.1665
PRECIO	–0.526583	0.548256	–0.960468	0.3739
RENTA	0.776285	0.020322	38.198860	0.0000
R-squared	0.996021	Mean dependent var		85.294220
Adjusted R-squared	0.994695	S.D. dependent var		56.387070
S.E. of regression	4.106913	Akaike info criterion		5.924422
Sum squared resid	101.200400	Schwarz criterion		5.990164
Log likelihood	–23.659900	F-statistic		751.028800
Durbin-Watson stat	1.693361	Prob(F-statistic)		0.000000

En función de estos nuevos estimadores se obtienen de nuevo los valores del consumo estimado $\hat{Y}^{(i)}$

Tabla 3.32

Observaciones	Consumo	Consumo Estimado	Consumo Estimado$^{(i)}$
1	97.160	94.94	95.14
2	96.510	81.05	92.49
3	38.830	24.28	35.21
4	16.000	104.39	127.09
5	194.300	170.03	197.01
6	5.198	18.04	10.16
7	97.260	77.27	93.38
8	123.100	102.37	126.39
9	29.260	21.72	32.39
10	86.030	89.55	85.48

con lo que el valor del estadístico distancia de Cook queda como:

$$D_i = \frac{\left(\hat{Y} - \hat{Y}^{(i)}\right)' \left(\hat{Y} - \hat{Y}^{(i)}\right)}{k\sigma_u^2} = \frac{2522.26}{3 \cdot 37.65^2} = 0.593212971$$

El valor crítico para decidir es el de una $F_{3,7}$ para un nivel de significación del 5%, o sea el valor 4.35. Con lo cual, nuestro estadístico cae en la región de aceptación y, por tanto, diremos que la familia cuarta no tiene una influencia real significativa sobre los parámetros estimados.

(d) Para obtener los valores de los errores estudentizados por omisión habría que calcular:

$$\frac{e_i^{(i)}}{S_{e_i}^{(i)}} = \frac{\left(Y_i - \hat{Y}_i^{(i)}\right)}{\sqrt{\sigma_u^{2(i)}(1-h_{ii})}}$$

donde

$$\hat{\sigma}_u^{(i)} = \sqrt{\left(\frac{1}{N-k}\sum_{i=1}^{N} e_i^2\right)^{(i)}} = \sqrt{\frac{12443.28}{6}} = 45.53986761$$

Tabla 3.33

Observaciones	Consumo	Consumo Estimado$^{(i)}$	Errores$^{(i)}$	$S_e^{(i)}$	Errores Estudentizados con omisión
1	97.160	95.14	2.02	37.59	0.05
2	96.510	92.49	4.02	43.19	0.09
3	38.830	35.21	3.62	37.67	0.10
4	16.000	127.09	−111.09	40.62	−2.73
5	194.300	197.01	−2.71	31.19	−0.09
6	5.198	10.16	−4.96	35.17	−0.14
7	97.260	93.38	3.88	42.46	0.09
8	123.100	126.39	−3.29	39.93	−0.08
9	29.260	32.39	−3.13	37.21	−0.08
10	86.030	85.48	0.55	34.31	0.02

EJERCICIO 3.37

Con los datos del Ejercicio 3.12 estudie la normalidad de los errores del modelo (3.26) y con los del Ejercicio 3.25 estudie la de los errores del modelo (3.27).

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{3.26}$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 D22_i + u_i \tag{3.27}$$

Solución

La normalidad la estudiamos mediante el contraste de Jarque-Bera (*JB*). La hipótesis nula del contraste es que la perturbación aleatoria es normal frente a una hipótesis alternativa que mantiene la no normalidad de dicha perturbación. Si la hipótesis nula se cumple, el estadístico *JB* se distribuye como una chi-cuadrado con dos grados de libertad, y viene dado por

$$JB = \frac{N-k}{6}\left(\gamma_1^2 + \frac{(\gamma_2 - 3)^2}{4}\right)$$

siendo γ_1 y γ_2 los coeficientes de asimetría y curtosis respectivamente.

$$\gamma_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}\right)^3} \quad \text{y} \quad \gamma_2 = \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^4}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}\right)^2}$$

Los datos intermedios para calcular el estadístico *JB* se muestran en la Tabla 3.34.

Tabla 3.34

$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$				$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 D22_i + u_i$			
e	e^2	e^3	e^4	e	e^2	e^3	e^4
−0.2978	0.0887	−0.0264	0.0079	−0.2812	0.0791	−0.0222	0.0063
−0.1781	0.0317	−0.0057	0.0010	−0.1547	0.0239	−0.0037	0.0006
0.2407	0.0579	0.0139	0.0034	0.2645	0.0700	0.0185	0.0049
0.4576	0.2094	0.0958	0.0438	0.4359	0.1900	0.0828	0.0361
0.1336	0.0179	0.0024	0.0003	0.1272	0.0162	0.0021	0.0003
−0.1279	0.0164	−0.0021	0.0003	−0.1832	0.0336	−0.0062	0.0011
−0.1431	0.0205	−0.0029	0.0004	−0.1856	0.0344	−0.0064	0.0012
−0.1178	0.0139	−0.0016	0.0002	−0.0820	0.0067	−0.0006	0.0000
0.0969	0.0094	0.0009	0.0001	0.1027	0.0105	0.0011	0.0001
0.2389	0.0571	0.0136	0.0033	0.1786	0.0319	0.0057	0.0010
0.1262	0.0159	0.0020	0.0003	0.1064	0.0113	0.0012	0.0001
−0.0795	0.0063	−0.0005	0.0000	−0.0979	0.0096	−0.0009	0.0001
0.2167	0.0470	0.0102	0.0022	0.1455	0.0212	0.0031	0.0004
−0.3258	0.1061	−0.0346	0.0113	−0.3034	0.0921	−0.0279	0.0085
0.1750	0.0306	0.0054	0.0009	0.1067	0.0114	0.0012	0.0001
−0.0773	0.0060	−0.0005	0.0000	−0.0797	0.0063	−0.0005	0.0000
0.0647	0.0042	0.0003	0.0000	0.0380	0.0014	0.0001	0.0000
0.4619	0.2133	0.0985	0.0455	0.4474	0.2002	0.0896	0.0401
−0.4871	0.2372	−0.1155	0.0563	−0.5148	0.2651	−0.1365	0.0703
−0.0388	0.0015	−0.0001	0.0000	−0.0688	0.0047	−0.0003	0.0000
0.0611	0.0037	0.0002	0.0000	0.0095	0.0001	0.0000	0.0000
−0.6469	0.4185	−0.2707	0.1751	0.0000	0.0000	0.0000	0.0000
−0.3305	0.1092	−0.0361	0.0119	−0.3492	0.1219	−0.0426	0.0149
0.4964	0.2464	0.1223	0.0607	0.4878	0.2380	0.1161	0.0566

(*continúa en la página siguiente*)

Tabla 3.34 (continuación)

	$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$				$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 D22_i + u_i$			
	e	e^2	e^3	e^4	e	e^2	e^3	e^4
	0.0465	0.0022	0.0001	0.0000	−0.0178	0.0003	0.0000	0.0000
	0.0637	0.0041	0.0003	0.0000	0.0540	0.0029	0.0002	0.0000
	0.4259	0.1814	0.0773	0.0329	0.3984	0.1587	0.0632	0.0252
	−0.1469	0.0216	−0.0032	0.0005	−0.1213	0.0147	−0.0018	0.0002
	−0.1471	0.0216	−0.0032	0.0005	−0.2056	0.0423	−0.0087	0.0018
	0.4088	0.1671	0.0683	0.0279	0.3833	0.1469	0.0563	0.0216
	−0.2033	0.0413	−0.0084	0.0017	−0.2215	0.0491	−0.0109	0.0024
	0.5178	0.2681	0.1388	0.0719	0.5178	0.2681	0.1388	0.0719
	−0.2801	0.0785	−0.0220	0.0062	−0.2566	0.0658	−0.0169	0.0043
	−0.1632	0.0266	−0.0043	0.0007	−0.1762	0.0310	−0.0055	0.0010
	−0.4410	0.1945	−0.0858	0.0378	−0.5044	0.2545	−0.1284	0.0648
Sumatorios	0.0000	2.9757	0.0267	0.6050	0.0000	2.5140	0.1601	0.4359

	$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 D22_i + u_i$
Media	0.0000	0.0000
Varianza	0.0850	0.0718
Desviación típica	0.2916	0.2680
γ_1	0.0308	0.2376
γ_2	2.3914	2.4138

Con los datos de la Tabla 3.34 y la fórmula del estadístico *JB* es inmediato calcular el valor numérico del estadístico para ambas series de errores.

El cálculo de *JB* para el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ es:

$$JB = \frac{N-k}{6}\left(\gamma_1^2 + \frac{(\gamma_2 - 3)^2}{4}\right) = \frac{35-2}{6}\left(0.0308^2 + \frac{(2.3914-3)^2}{4}\right) = 0.52$$

En el caso del modelo $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 D22_i + u_i$, el estadístico *JB* da

$$JB = \frac{N-k}{6}\left(\gamma_1^2 + \frac{(\gamma_2 - 3)^2}{4}\right) = \frac{35-3}{6}\left(0.2376^2 + \frac{(2.4138-3)^2}{4}\right) = 0.76$$

Para decidir sobre estos contrastes únicamente nos falta conocer el punto crítico para el nivel de significación con el que deseemos trabajar. Para un 5%, las tablas de la chi-cuadrado de 2 grados de libertad nos informan de que el valor crítico es 5.99. En consecuencia, en ambos casos no podemos rechazar

la hipótesis de normalidad de la perturbación aleatoria. Por tanto, tampoco sería imprescindible incluir la variable dicotómica en el modelo, debido a que la existencia de la observación 22 como *outlier* no genera problemas de falta de normalidad.

EJERCICIO 3.38

Sea la siguiente estimación por Mínimos Cuadrados Ordinarios:

$$\hat{Y}_i = 0.465121 + 0.521412\, X_i \qquad R^2 = 0.719744 \qquad N = 50 \qquad (3.28)$$

de la cual además sabemos que

$$\sum_{i=1}^{N} e_i^3 = -0.5752378 \qquad \sum_{i=1}^{N} e_i^4 = 1.332327169 \qquad S_Y^2 = 0.325331953$$

Con la información suministrada, contraste la normalidad de los residuos de la estimación (3.28).

Solución

Para contrastar la normalidad de los errores del modelo utilizamos el contraste de Jarque Bera, cuya hipótesis nula es la existencia de normalidad, y su estadístico de contraste se obtiene de la expresión siguiente:

$$JB = \frac{N-k}{6}\left(\gamma_1^2 + \frac{(\gamma_2 - 3)^2}{4}\right)$$

donde γ_1 y γ_2 son el coeficiente de asimetría y curtosis respectivamente.

A continuación procedemos a calcular cada uno de los valores necesarios:

- Cálculo de la desviación típica de los errores:

$$R^2 = 1 - \frac{S_e^2}{S_Y^2} \Rightarrow$$

$$\Rightarrow S_e^2 = \left(1 - R^2\right) S_Y^2 = \left(1 - 0.719744\right) \cdot 0.325331953 \Rightarrow S_e = 0.301954094$$

- Cálculo del coeficiente de asimetría:

$$\gamma_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^3}{S_e^3} = \frac{-0.5752373}{50 \cdot \left(0.301954094\right)^3} = -0.417882942$$

- Cálculo del coeficiente de curtosis:

$$\gamma_2 = \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^4}{S_e^4} = \frac{1.33237169}{50 \cdot (0.301954094)^4} = 3.205362771$$

Finalmente se obtiene el valor del estadístico de prueba propuesto por Jarque Bera:

$$JB = \frac{50-2}{6}\left[(-0.417882942)^2 + \frac{(3.205362771-3)^2}{4}\right] = 1.481356961$$

Teniendo en cuenta que, para un nivel de significación del 5%, una χ_2^2 presenta un valor crítico tabulado en 5.99, el estadístico cae en la región de aceptación y no tenemos evidencia para rechazar la hipótesis nula de normalidad.

EJERCICIO 3.39

La relación empírica existente entre los ingresos que obtienen los individuos en el mercado de trabajo y su nivel educativo se apoya en la ecuación de ingresos de Mincer (1974) que tiene la siguiente especificación:

$$Log(SALARIO_i) = \beta_1 + \beta_2 EDUCACIÓN_i + \beta_3 EXPERIENCIA_i +$$
$$+ \beta_4 EXPERIENCIA_i^2 + u_i$$

En la estimación del Cuadro 3.14 se ha incorporado a dicha especificación una dicotómica referida al lugar de residencia del individuo, que toma valor 1 cuando el individuo reside en una zona urbana y 0 cuando reside en una zona rural.

Cuadro 3.14

Dependent Variable: *LSALARIO*
Method: Least Squares
Sample: 1 120
Included observations: 120

Variable	Coefficient	Std. Error	t-Statistic	Prob.
EDUCA	0.073290	0.016103	4.551218	0.0000
EXPER	0.065069	0.019822	3.282737	0.0014
EXPER2	–0.001628	0.000506	–3.219022	0.0017
RURALURB	0.240332	0.067729	3.548443	0.2006
C	8.258771	0.272031	30.359620	0.0000
R-squared	0.295652	Mean dependent var		9.837799
Adjusted R-squared	0.271153	S.D. dependent var		0.418795
S.E. of regression	0.357536	Akaike info criterion		0.821613
Sum squared resid	14.700680	Schwarz criterion		0.937758
Log likelihood	–44.296760	F-statistic		12.067890
Durbin-Watson stat	1.955671	Prob(F-statistic)		0.085000

El análisis de los residuos de esta regresión aporta los resultados del Cuadro 3.15:

Cuadro 3.15

Series: Residuals
Sample 1 120
Observations 120

Mean	7.55E–16
Median	0.035190
Maximum	0.756713
Minimum	–1.167407
Std. Dev.	0.351476
Skewness	–0.526577
Kurtosis	3.551141
Jarque-Bera	
Probability	

(a) ¿Siguen las perturbaciones de este modelo una distribución Normal?

(b) ¿Puede considerar, sin temor a equivocarse, que la dicotómica introducida en el modelo es significativa? ¿Y que el modelo es globalmente significativo?

Solución

(a) Para comprobar si los residuos se distribuyen como una Normal habrá que realizar el contraste de Jarque-Bera.

Las hipótesis a contrastar son:

$$\left.\begin{array}{l} H_0\text{: Perturbaciones normales} \\ H_1\text{: Perturbaciones no normales} \end{array}\right\}$$

Y el estadístico de contraste es:

$$JB = \frac{N-k}{6}\cdot\left(\gamma_1^{\,2} + \frac{(\gamma_2 - 3)^2}{4}\right) \quad \text{con} \quad \left\{\begin{array}{l} \gamma_1 = \dfrac{m_3}{S_e^3} = \dfrac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} e_i^3}{\left(\sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} e_i^2}\right)^3} \\ \gamma_2 = \dfrac{m_4}{S_e^4} = \dfrac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} e_i^4}{\left(\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} e_i^2\right)^2} \end{array}\right. \tag{3.29}$$

En el Cuadro 3.15 tenemos los datos de N (Observations), γ_1 (Skewness) y γ_2 (Kurtosis), mientras que en el Cuadro 3.14 podemos ver el número de regresores k. Sustituyendo estos valores en la expresión (3.29) obtenemos el valor del estadístico:

$$JB = \frac{120-5}{6}\cdot\left((-0.526577)^2 + \frac{(3.551141-3)^2}{4}\right) = 6.77$$

Comparando este resultado con el valor tabulado para una distribución chi-cuadrado con 2 grados de libertad, que vale 5.99, comprobamos que el valor del estadístico cae en la región de rechazo, por lo que no podemos aceptar la hipótesis nula de normalidad de los residuos.

(b) Aunque los contrastes de significación individual y global nos lleven a rechazar ambas hipótesis —por mostrar probabilidades asociadas a sus estadísticos superiores al 5% —, hay que tener en cuenta que la distribución de los errores no es normal, por lo que dichos contrastes deben ser interpretados con cautela.

EJERCICIO 3.40

Dispone de la información del Cuadro 3.16 acerca de los residuos de la estimación de un Modelo de Regresión Lineal Simple.

Cuadro 3.16

Series: Residuals
Sample 1 50
Observations 50

Mean	1.85E–16
Median	–0.046000
Maximum	1.561000
Minimum	–0.877000
Std. Dev.	0.366000
Skewness	2.004917
Kurtosis	10.303300
Jarque-Bera	
Probability	0.000000

(a) Calcule el valor del estadístico Jarque-Bera. ¿Qué puedes decir acerca de la normalidad de la perturbación aleatoria?

(b) Si dispone de información de los valores observados y estimados de las 10 primeras observaciones en la Tabla 3.35, ¿detecta algún *outlier*?

Tabla 3.35

Y	$\hat{Y}$
0.33	0.19
0.30	0.32
1.85	1.68
0.66	0.43
0.96	1.08
0.08	0.09
0.31	0.63
1.57	1.52
7.77	6.58
0.31	0.23

Solución

(a) Para calcular el estadístico de Jarque-Bera es necesario utilizar los valores de los coeficientes de asimetría y curtosis que nos proporciona el Cuadro 3.16.

$$JB = \frac{50-2}{6} \cdot \left((2.004927)^2 + \frac{(10.30330-3)^2}{4} \right) = 138.834$$

Dado este resultado no podemos considerar que el término de perturbación aleatoria sea normal, ya que el valor del estadístico de contraste de Jarque-Bera es 138.834, valor muy superior al tabulado para un nivel de significación del 5% ($\chi^2_2 = 5.99$).

También podemos llegar a esta conclusión sin más que mirar la probabilidad asociada al estadístico de prueba, que en este caso es 0.0000; lo que indica que dicho estadístico se sitúa en la región de rechazo de la hipótesis nula.

(b) Para detectar la presencia de *outliers* calculamos el valor de los residuos mínimo cuadrático ordinarios para las 10 primeras observaciones, obteniendo los resultados recogidos en la Tabla 3.36

Tabla 3.36

Y	$\hat{Y}$	Residuo	Residuo tipificado
0.33	0.19	0.14	0.38
0.30	0.32	–0.05	–0.05
1.85	1.68	0.17	0.46
0.66	0.43	0.23	0,63
0.96	1.08	–0.12	–0.33
0.08	0.09	–0.01	–0.03
0.31	0.63	–0.32	–0.87
1.57	1.52	0.05	0.14
7.77	6.58	1.19	3.25
0.31	0.23	0.08	0.22

El estadístico de prueba del contraste se distribuye bajo la hipótesis nula como una normal estándar, luego la única observación que puede ser considerada un *outlier* es la novena, puesto que su residuo tipificado vale 3.25, superior al 1.96 que adopta la $N(0,1)$ en las tablas, para un nivel de significación del 5%.

4

Especificación de la forma funcional. Aspectos cualitativos y cambio estructural. Selección de regresores.

EJERCICIO 4.1

Linealice los siguientes modelos e interprete los coeficientes.

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i} \tag{4.1}$$

$$Y_i = \beta_1^{X_{1i}} X_{2i}^{\beta_2} e^{u_i} \tag{4.2}$$

Solución

Para linealizar el modelo (4.1) tomamos logaritmo neperiano en ambos lados de la ecuación y obtenemos

$$\ln(Y_i) = \ln \beta_1 + \beta_2 \ln(X_i) + u_i$$

Cambiando la notación de la siguiente forma:

$$Y_i^* = \ln(Y_i) \quad \alpha_1 = \ln \beta_1 \quad X_i^* = \ln(X_i) \tag{4.3}$$

el modelo lineal a estimar es el siguiente:

$$Y_i^* = \alpha_1 + \beta_2 X_i^* + u_i$$

En este modelo un incremento de un 1% en X provoca un incremento de un β_2% en Y.

En el modelo (4.2), tomando logaritmo neperiano, es inmediato llegar a la siguiente expresión:

$$\ln(Y_i) = \ln(\beta_1) X_{1i} + \beta_2 \ln(X_{2i}) + u_i$$

Realizando los cambios de variable de forma similar al modelo (4.1), el modelo a estimar viene dado por

$$Y_i^* = \alpha_1 X_{1i} + \beta_2 X_{2i}^* + u_i$$

en donde las variables Y_i^* y X_i^* coinciden con las definidas en (4.3) y α_1 es igual a $\ln(\beta_1)$ y, por tanto, $\beta_1 = e^{\alpha_1}$. En consecuencia, el incremento de una unidad de X_1 producirá un cambio del $\ln(\beta_1)\cdot 100\%$ en Y, mientras que un incremento de un 1% en la variable X_2 produce un cambio del β_2% en Y.

EJERCICIO 4.2

Dada la especificación del modelo $e^{Y_i} = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$, demuestre analíticamente cómo debe interpretarse β_2.

Solución

Si se aplica logaritmo neperiano al modelo se obtiene el siguiente modelo econométrico:

$$Y_i = \ln(\beta_1) + \beta_2 \ln(X_{2i}) + \beta_3 \ln(X_{3i}) + u_i$$

La elasticidad de Y con respecto a X_2 se puede escribir como

$$\varepsilon_Y^{X_2} = \frac{\partial Y}{\partial X_2} \cdot \frac{X_2}{Y} = \beta_2 \frac{1}{X_2} \cdot \frac{X_2}{Y}, \quad \text{de donde} \quad \partial Y = \frac{\partial X_2}{X_2} \cdot \beta_2$$

Discretizando esta última expresión podemos escribir

$$\Delta Y = \frac{\Delta X_2}{X_2} \cdot \beta_2$$

La expresión $\Delta X_2/X_2$ mide el cambio relativo de la variable X_2 y ΔY mide el cambio absoluto de Y, por tanto, β_2 mide el cambio absoluto que sufre la variable Y ante un cambio relativo de la variable X_2. Es decir, si X_2 se incrementa en un 1% (0.01), la variable Y se incrementa en $0.01 \cdot \beta_2$ unidades.

EJERCICIO 4.3

Dado el modelo

$$\ln(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 \ln(X_{3i}) + \beta_4 \frac{1}{X_{4i}} + u_i$$

(a) Calcule las expresiones de los multiplicadores de la variable endógena con respecto a cada una de las variables explicativas.

(b) Calcule las expresiones de las elasticidades de la variable endógena con respecto a cada una de las variables explicativas.

Solución

El modelo planteado se puede escribir como

$$Y_i = e^{\beta_1 + \beta_2 X_{2i} + \beta_3 \ln(X_{3i}) + \beta_4 \frac{1}{X_{4i}} + u_i}$$

(a) La expresión que nos permite calcular los multiplicadores se corresponde con la derivada parcial de la variable endógena con respecto a cada una de las variables explicativas. Por tanto, derivando el modelo anterior con respecto a cada una de las variables explicativas se obtienen los siguientes resultados:

$$\text{Multiplicador de } X_2\text{:} \quad \frac{\partial Y}{\partial X_2} = \beta_2 Y$$

$$\text{Multiplicador de } X_3\text{:} \quad \frac{\partial Y}{\partial X_3} = \beta_3 \frac{1}{X_3} Y$$

$$\text{Multiplicador de } X_4\text{:} \quad \frac{\partial Y}{\partial X_4} = \beta_4 \frac{-1}{X_4^2} Y$$

(b) Para calcular las correspondientes elasticidades únicamente se debe multiplicar el multiplicador obtenido en el apartado anterior por la ratio formada por la variable correspondiente y la endógena. Los resultados son los siguientes:

$$\text{Elasticidad debido a } X_2: \quad \varepsilon_Y^{X_2} = \frac{\partial Y}{\partial X_2} \cdot \frac{X_2}{Y} = \beta_2 Y \cdot \frac{X_2}{Y} = \beta_2 X_2$$

$$\text{Elasticidad debido a } X_3: \quad \varepsilon_Y^{X_3} = \frac{\partial Y}{\partial X_3} \cdot \frac{X_3}{Y} = \beta_3 \frac{1}{X_3} Y \cdot \frac{X_3}{Y} = \beta_3$$

$$\text{Elasticidad debido a } X_4: \quad \varepsilon_Y^{X_4} = \frac{\partial Y}{\partial X_4} \cdot \frac{X_4}{Y} = \beta_4 \frac{-1}{X_4^2} Y \cdot \frac{X_4}{Y} = -\frac{\beta_4}{X_4}$$

EJERCICIO 4.4

Demuestre que, en un modelo de elasticidad constante, como es el caso de la función de producción Cobb-Douglas $Y_i = AX_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{u_i}$, donde X_1 se corresponde con el factor capital y X_2 con el factor trabajo, los coeficientes pueden ser interpretados como elasticidades. Compruébelo sólo para el caso de uno de los factores de producción, el factor capital.

Solución

Si calculamos la derivada parcial de la producción respecto del factor capital X_1, tenemos

$$\frac{\partial Y}{\partial X_1} = AX_2^{\beta_2} e^U \beta_1 X_1^{\beta_1 - 1}$$

Utilizando la expresión para el cálculo de la elasticidad, nos lleva a

$$\varepsilon_Y^{X_1} = \frac{\Delta Y/Y}{\Delta X_1/X_1} = \frac{\partial Y/Y}{\partial X_1/X_1} = \frac{\partial Y}{\partial X_1} \cdot \frac{X_1}{Y} = AX_2^{\beta_2} e^U \beta_1 X_1^{\beta_1 - 1} \frac{X_1}{AX_2^{\beta_2} e^U X_1^{\beta_1}} = \beta_1$$

Por tanto, el coeficiente β_1 se interpreta en términos de elasticidad. Es decir, un incremento del factor capital de un 1% se traduce en un incremento en la producción en un β_1%.

EJERCICIO 4.5

Se desea representar la tasa de variación salarial (Y) en función de la tasa de desempleo (X) mediante un modelo con forma funcional recíproca como

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i$$

Si los coeficientes estimados son: $\hat{\beta}_1 = 0.5$ y $\hat{\beta}_2 = 0.4$,

(a) ¿Son los signos los esperados?

(b) Represente gráficamente la función.

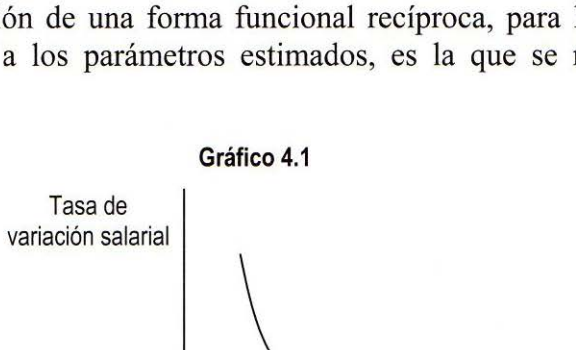
(c) Calcule la pendiente y la elasticidad en el punto $(X_{2i} = 0.2; Y_i = 2.5)$ e interprételos.

Solución

(a) Se trata de una curva de Phillips. En este caso, la estimación de β_2 tiene el signo esperado, mientras que no ocurre lo mismo con la estimación de β_1, coeficiente para el cual esperábamos un signo negativo, dada la relación inversa entre la tasa de variación salarial y la tasa de desempleo.

(b) La representación de una forma funcional recíproca, para los signos correspondientes a los parámetros estimados, es la que se muestra en el Gráfico 4.1:

Gráfico 4.1



(c) El cálculo de la pendiente se obtiene como

$$\frac{\partial Y}{\partial X_2} = \frac{-\hat{\beta}_1}{X_2^2} = \frac{-0.4}{0.2^2} = -10$$

Cuando $X_2 = 0.2$ (tasa de desempleo), un aumento de dicha tasa en una unidad supone una reducción aproximada de la tasa de variación salarial de 10 unidades.

Por su parte, la elasticidad es igual a

$$\frac{\partial Y}{\partial X_2}\cdot\frac{X_2}{Y_i}=\frac{-\hat{\beta}_1}{X_2^2}=\frac{-0.4}{0.2\cdot 2.5}=-0.8$$

En el punto (0.2, 2.5) un incremento de la tasa de desempleo de un 1% produce una reducción aproximada de la tasa de variación salarial del 80%.

EJERCICIO 4.6

Sea el modelo

$$Y_i=\beta_1+\beta_2 X_{2i}^*+u_i$$

donde Y es el consumo de carne mensual de familias con igual número de miembros, medido en kilogramos, y $X_2^*=1/X_2$, siendo X_2 los ingresos mensuales de las familias, en euros.

Para ese modelo se sabe que

$$(X'X)^{-1}=\begin{pmatrix}0.0317 & -8.7937\\ & 5147.5185\end{pmatrix}\quad X'Y=\begin{pmatrix}206\\0.1735\end{pmatrix}$$

(a) ¿Cuál es el consumo de carne mensual que satura a las familias?

(b) ¿Por debajo de qué nivel de ingresos mensuales no se consume carne?

(c) Represente gráficamente el consumo estimado (eje Y) contra los ingresos mensuales (eje X). Señale en el gráfico los puntos que representan la solución de los apartados (a) y (b).

Solución

(a) El modelo que se plantea es el referido a la curva de gasto de Engel, en el que la constante del modelo indica el umbral de saturación y el punto de corte con el eje de abscisas, que viene dado por $-\hat{\beta}_2/\hat{\beta}_1$, indica el umbral de ingresos por debajo del cual no se consume ese bien. Teniendo en cuenta la estimación del modelo, el consumo de carne mensual que satura a las familias es de 5 Kg, que coincide con el coeficiente estimado para la constante.

$$\hat{\beta}=(X'X)^{-1}X'Y=\begin{pmatrix}0.0317 & -8.7937\\ -8.7937 & 5147.5185\end{pmatrix}\begin{pmatrix}206\\0.1735\end{pmatrix}=\begin{pmatrix}5\\-918.40\end{pmatrix}$$

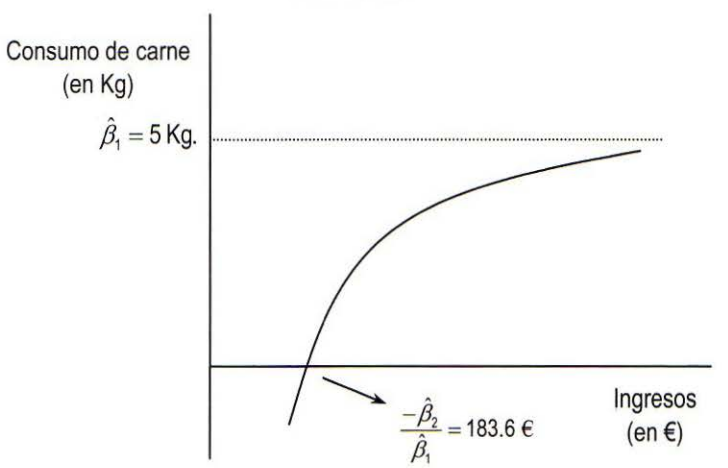
(b) El umbral de renta por debajo del cual no se consume carne se obtiene a partir de

$$\frac{-\hat{\beta}_2}{\hat{\beta}_1} = \frac{-(-918.40)}{5} = 183.5\ €$$

Éste es el nivel de ingresos mínimo a partir del cual el consumo de carne empieza a ser habitual.

(c) La representación gráfica tras la estimación quedaría como se recoge en el Gráfico 4.2.

Gráfico 4.2



EJERCICIO 4.7

Se tienen las siguientes ecuaciones y gráficos:

$$\hat{Y}_i = 0.5 - 2X_{2i} + X_{2i}^2 \tag{4.4}$$

$$\hat{Y}_i = 0.5 + 2X_{2i} - X_{2i}^2 \tag{4.5}$$

$$\hat{Y}_i = 2 - 4\frac{1}{X_{2i}} \tag{4.6}$$

$$\hat{Y}_i = 2 + 4\frac{1}{X_{2i}} \tag{4.7}$$

Gráfico 4.3

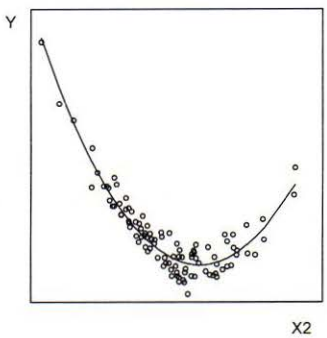


Gráfico 4.4

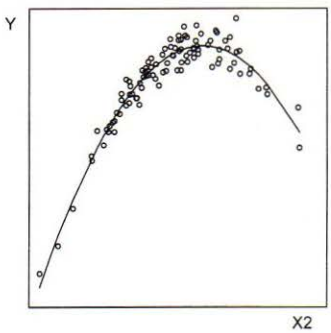


Gráfico 4.5

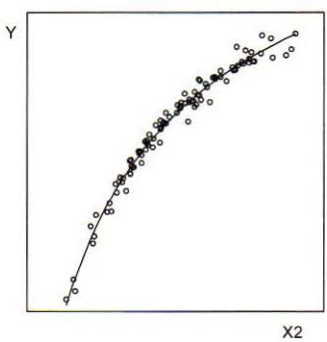
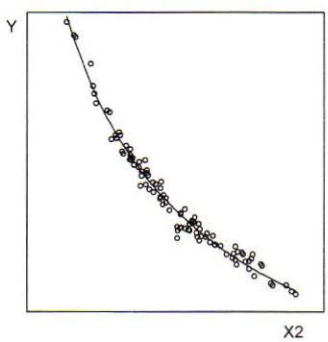


Gráfico 4.6



Identifique cada uno de los gráficos con su ecuación correspondiente y responda a las siguientes cuestiones:

(a) ¿Qué valor tiene X_2 en el mínimo de la curva de regresión dibujada en el Gráfico 4.3? ¿Y para el Gráfico 4.4, cuál sería el valor de X_2 en el máximo de la curva de regresión?

(b) De entre todos los valores que $\hat{Y}$ no puede alcanzar en la función correspondiente al Gráfico 4.5, ¿cuál es el más pequeño? Justifique su respuesta.

(c) De entre todos los valores que $\hat{Y}$ no puede alcanzar en la función correspondiente al Gráfico 4.6, ¿cuál es el más grande? Justifique su respuesta.

Solución

La relación entre gráficos y ecuaciones se recoge en la Tabla 4.1:

Tabla 4.1

Gráfico	Ecuación
4.3	(4.4)
4.4	(4.5)
4.5	(4.6)
4.6	(4.7)

(a) En el caso del Gráfico 4.3, el resultado lo obtenemos a partir de

$$\frac{d\hat{Y}}{dX_2} = -2 + 2X_2 = 0 \quad \Rightarrow \quad X_2 = \frac{2}{2} = 1$$

Por tanto, el valor mínimo es 1.

En el caso del Gráfico 4.4, la solución es análoga, obteniéndose incluso el mismo valor.

$$\frac{d\hat{Y}}{dX_2} = 2 - 2X_2 = 0 \quad \Rightarrow \quad X_2 = \frac{2}{2} = 1$$

(b) Este apartado pregunta por la asíntota del Gráfico 4.5. Ésta se calcula a partir del valor de Y cuando X_2 tiende a ∞. Dicho valor, según la expresión (4.6), es de 2.

(c) De la misma forma, en el Gráfico 4.6, la asíntota se encuentra también en el valor 2.

EJERCICIO 4.8

Dado el siguiente modelo estimado:

$$\ln\left(\hat{Y}_i\right) = -2.55 + 0.68 \ln\left(X_{2i}\right)$$

donde Y es el consumo de leche y X_2 la renta disponible, obtenga la expresión de la elasticidad consumo-renta y de la propensión marginal al consumo, sabiendo que el consumo medio de leche es de 11.49 litros y que la renta disponible media es de 1413.13 euros. Calcule sus valores e interprete el resultado.

Solución

Teniendo en cuenta que podemos aproximar el logaritmo de una variable de la siguiente forma:

$$\ln(Y) \sim \frac{\partial Y}{Y} \quad \text{y} \quad \ln(X_2) \sim \frac{\partial X_2}{X_2}$$

y que en los modelos *doble-log*, como el del enunciado, los coeficientes de los regresores miden elasticidades, la elasticidad de Y respecto a X_2 es:

$$\varepsilon_Y^{X_2} = \frac{\partial \ln(Y)}{\partial \ln(X_2)} \sim \frac{\dfrac{\partial Y}{Y}}{\dfrac{\partial X_2}{X_2}} = 0.68$$

Dicha elasticidad implica que por cada 1% que aumenta la renta, el consumo lácteo aumenta un 0.68%. Se trata pues de un bien normal.

La propensión marginal de una variable Y respecto a X_2 viene dada por $\partial Y/\partial X_2$. Conociendo la elasticidad, y despejando, obtenemos para un individuo promedio:

$$\frac{\dfrac{\partial Y}{Y}}{\dfrac{\partial X_2}{X_2}} = 0.68 \quad \Rightarrow \quad \frac{\partial Y}{\partial X_2} = \frac{\overline{Y}}{\overline{X}_2} \cdot 0.68 = \frac{11.49}{1413.13} \cdot 0.68 = 0.0055$$

lo que indica que, por cada euro que aumenta la renta disponible, el consumo de leche se incrementa en promedio en 0.0055 litros.

EJERCICIO 4.9

La estimación de un modelo de regresión para una determinada fábrica proporciona el siguiente resultado:

$$\hat{Y}_i = 2.5X_i - 0.01X_i^2$$

donde Y es la producción total de un determinado producto y X es el *input* necesario para su fabricación.

(a) Si en estos momentos la cantidad de *input* que está utilizando la fábrica es de 200 unidades, ¿cuál es la estimación del incremento de la producción si se incrementa en una unidad la cantidad de *input*? Comente el resultado.

(b) ¿Cuál es la capacidad máxima de producción estimada de esta fábrica?

Solución

(a) Si se incrementa en una unidad la cantidad de *input*, el incremento de producción estimado, para un nivel de *input* genérico, será de

$$\frac{\partial \hat{Y}}{\partial X} = 2.5 - 0.02X$$

En el caso concreto que plantea el enunciado, en donde se está utilizando un *input* de 200 unidades, la variación marginal en la producción total será de $2.5 - 0.02 \cdot 200 = -1.5$ unidades.

El haber obtenido un resultado negativo indica que nos encontramos en la parte de rendimientos decrecientes de la función de producción y que se ha rebasado el óptimo de producción de la fábrica.

(b) Para conocer la capacidad máxima de producción de la fábrica basta con calcular el máximo de su función de producción. Para ello habrá que calcular su primera derivada e igualarla a cero y comprobar que su segunda derivada es menor a cero (de esta forma comprobamos que se trata de un máximo y no de un mínimo).

La primera derivada se obtiene como

$$\frac{\partial \hat{Y}}{\partial X} = 2.5 - 0.02X = 0 \quad \Rightarrow \quad X = \frac{2.5}{0.02} = 125$$

y con la segunda derivada comprobamos la condición de máximo

$$\frac{\partial \hat{Y}}{\partial^2 X} = -0.02 < 0$$

Por tanto, la capacidad máxima estimada de producción de esta fábrica es la que se consigue utilizando 125 unidades de *input*, es decir, cuando se producen $\hat{Y} = 2.5 \cdot 125 - 0.01 \cdot 125^2 = 156.25$ unidades.

EJERCICIO 4.10

Dado el siguiente modelo estimado:

$$\ln\left(\hat{Y}_i\right) = 10.53892 + 0.00000226X_{2i} + 0.042878\ln\left(X_{3i}\right) - 27153.75\frac{1}{X_{4i}}$$

y teniendo en cuenta que los valores medios de Y, X_2, X_3 y X_4 son respectivamente 117108.4, 235017, 17291.82 y 223981.2:

(a) interprete los parámetros estimados del modelo,

(b) calcule los multiplicadores y las elasticidades con respecto a un individuo promedio.

Solución

(a) Como se puede observar, la variable endógena está en logaritmos, la variable X_2 está incluida en la especificación del modelo en forma lineal, X_3 en logaritmos y X_4 como inversa o recíproco de una variable. Teniendo en cuenta esto, y dado que el estimador de X_2 es igual a 0.00000226, un incremento unitario en la variable X_2 produce un incremento del $0.00000226 \cdot 100\% = 0.000226\%$ en la variable Y. En cuanto a la variable X_3, el parámetro estimado se corresponde con la elasticidad estimada. Por tanto, un incremento del 1% en dicha variable produce un incremento del 0.043% en la variable Y. Del parámetro de la inversa de X_4 únicamente podemos decir que, dado que su signo es negativo, al incrementar el valor de dicha variable se incrementa el valor de Y.

(b) Teniendo en cuenta que el modelo estimado se corresponde con

$$Y_i = e^{\beta_1 + \beta_2 X_{2i} + \beta_3 \ln(X_{3i}) + \beta_4 \frac{1}{X_{4i}} + u_i}$$

el cálculo de los multiplicadores y las elasticidades para un individuo promedio es inmediato.

Usando las expresiones obtenidas en el ejercicio 4.3 y refiriéndolas a los valores medios obtenemos:

Multiplicador de X_2 para un individuo promedio: $\dfrac{\partial Y}{\partial X_2} = \beta_2 \overline{Y}$

Multiplicador de X_3 para un individuo promedio: $\dfrac{\partial Y}{\partial X_3} = \beta_3 \dfrac{1}{\overline{X}_3} \overline{Y}$

Multiplicador de X_4 para un individuo promedio: $\dfrac{\partial Y}{\partial X_4} = \beta_4 \dfrac{-1}{\overline{X}_4^2} \overline{Y}$

Elasticidad debida a X_2 para un individuo promedio: $\varepsilon_Y^{X_2} = \beta_2 \overline{X}_2$

Elasticidad debida a X_3 para un individuo promedio: $\varepsilon_Y^{X_3} = \beta_3$

Elasticidad debida a X_4 para un individuo promedio: $\varepsilon_Y^{X_4} = -\dfrac{\beta_4}{\overline{X}_4}$

Sustituyendo los parámetros por sus estimadores y las medias de cada variable por los valores proporcionados se obtienen las siguientes estimaciones:

- Estimación del multiplicador de X_2 para un individuo promedio:

$$\frac{\partial Y}{\partial X_2} = \hat{\beta}_2 \overline{Y} = 0.00000226 \cdot 117\,108.4 = 0.265$$

Es decir, para un individuo promedio, el incremento unitario en X_2 produce un incremento de 0.265 unidades en Y.

- Estimación del multiplicador de X_3 para un individuo promedio:

$$\frac{\partial Y}{\partial X_3} = \hat{\beta}_3 \frac{1}{\overline{X}_3} \overline{Y} = 0.042879 \frac{117\,108.4}{17\,291.82} = 0.29$$

Es decir, para un individuo promedio, el incremento unitario en X_3 produce un incremento de 0.29 unidades en Y.

- Estimación del multiplicador de X_4 para un individuo promedio:

$$\frac{\partial Y}{\partial X_4} = \hat{\beta}_4 \frac{-1}{\overline{X}_4^2} \overline{Y} = -27\,153.75 \frac{-1}{223\,981.2} 117\,108.4 = 14\,197.3$$

Es decir, para un individuo promedio, el incremento unitario en X_4 produce un incremento de 14197.3 unidades en Y.

- De la misma forma, las elasticidades debidas a X_2, X_3 y X_4 valdrán 0.5311, 0.042878 y –0.1212 respectivamente.

EJERCICIO 4.11

A partir de una muestra de 10 familias procedentes de 3 provincias distintas se ha obtenido la siguiente estimación:

$$\ln\left(\hat{C}_i\right) = 3.11 - 1.56D1_i - 0.89D2_i + 0.017R_i$$

donde:

*D*1 toma valor 1 para las familias correspondientes a la provincia 1 y 0 para el resto.

*D*2 toma valor 1 para las familias correspondientes a la provincia 2 y 0 para el resto.

Se sabe además que el consumo medio es 25.48 euros diarios y que la renta media es 31.30 euros diarios.

(a) Obtenga la propensión marginal al consumo e interprete el resultado.

(b) Calcule la elasticidad renta e interprete el resultado.

(c) Interprete el coeficiente de $D1$.

Solución

(a) La propensión marginal al consumo respecto de la renta es: $\frac{\partial C}{\partial R}$. Esta expresión podemos despejarla de la siguiente manera, evaluándola en el consumo medio para obtener un valor promedio de la misma:

$$\frac{\partial \ln C}{\partial \mathrm{R}} = \frac{\partial C / C}{\partial \mathrm{R}} = 0.017$$

$$\frac{\partial C}{\partial \mathrm{R}} = \overline{C} \cdot 0.017 = 25.48 \cdot 0.017 = 0.433$$

Por cada unidad que aumente la renta, el consumo de un individuo promedio aumenta en 0.433 unidades.

(b) La elasticidad del consumo respecto de la renta se calcula como $\varepsilon_C^R = \frac{\partial C}{\partial R} \cdot \frac{R}{C}$. Si la evaluamos para los valores medios obtenemos

$$\frac{\partial C}{\partial R} \cdot \frac{\overline{R}}{\overline{C}} = \overline{C} \cdot 0.017 \cdot \frac{\overline{R}}{\overline{C}} = 0.017 \cdot 31.3 = 0.5321$$

Es decir, por cada uno por ciento que aumenta la renta, el consumo aumenta un 0.5321%, siempre respecto a un individuo promedio.

(c) Al estar la variable endógena medida el logaritmos neperianos y estar refiriéndonos al coeficiente estimado de una variable ficticia, éste se interpreta a partir del cálculo de $e^{-1.56} - 1 = -0.79$, lo que indica que el consumo medio de la provincia 1 es inferior en un 79% al de la provincia 3, que actúa como referencia.

EJERCICIO 4.12

Suponga que se desea estudiar los efectos de la discriminación en un mercado laboral de un determinado país y sus efectos sobre el consumo de un determinado grupo de familias. Para ello se considera el siguiente modelo:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 P_i + \beta_4 T_i + u_i$$

donde C_i es el consumo de la familia i, Y_i es su renta, P es una variable dicotómica que toma valor 1 para familias donde el sustentador principal trabaja en el sector primario y 0 en los demás casos, y T es otra variable ficticia que toma valor 1 para familias en las que el sustentador principal de la familia trabaja en el sector terciario y 0 en los demás casos. Se dispone de una muestra de familias en las que el trabajador principal pertenece a los sectores primario, secundario o terciario.

(a) ¿Cuál es el consumo medio para los distintos sectores?

(b) ¿Cómo verificaría la hipótesis de que el pertenecer a un sector no tiene efecto sobre el consumo?

(c) Construya un modelo para el caso en el que el efecto renta pueda depender del sector y determine el consumo medio para cada una de ellos.

Solución

(a) El consumo medio para cada sector es:

$$\text{Sector Primario: } E(C_i / P_i = 1) = (\beta_1 + \beta_3) + \beta_2 Y_i$$

$$\text{Sector Secundario: } E(C_i / P_i = T_i = 0) = \beta_1 + \beta_2 Y_i$$

$$\text{Sector Terciario: } E(C_i / T_i = 1) = (\beta_1 + \beta_4) + \beta_2 Y_i$$

(b) Las hipótesis a contrastar para verificar si el pertenecer a un sector no tiene efecto sobre el consumo serían:

$$\left.\begin{array}{l} H_0 : \beta_3 = \beta_4 = 0 \\ H_1 : \beta_3 \quad \text{y/o} \quad \beta_4 \neq 0 \end{array}\right\}$$

(c) Si queremos incorporar que el efecto renta dependa del sector, habría que introducir una dicotómica multiplicativa de la siguiente forma:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 P_i + \beta_4 T_i + \beta_5 Y_i P_i + \beta_6 Y_i T_i + u_i$$

Los consumos medios para cada uno de ellos vendrían dados por:

Sector Primario: $$E(C_i / P_i = 1) = (\beta_1 + \beta_3) + \beta_2 E(Y_i) + \beta_5 E(Y_i) = \\ = \beta_1 + \beta_3 + (\beta_2 + \beta_5) E(Y_i)$$

Sector Secundario: $$E(C_i / P_i = T_i = 0) = \beta_1 + \beta_2 E(Y_i)$$

Sector Terciario: $$E(C_i / T_i = 1) = \beta_1 + \beta_4 + \beta_2 E(Y_i) + \beta_6 E(Y_i) = \\ = (\beta_1 + \beta_4) + (\beta_2 + \beta_6) E(Y_i)$$

EJERCICIO 4.13

Dada la siguiente definición de las variables ficticias:

$$Z_{1i} = \begin{cases} 1, & i \in \text{zona costera} \\ 0, & i \notin \text{zona costera} \end{cases} \quad Z_{2i} = \begin{cases} 1, & i \notin \text{zona costera} \\ 0, & i \in \text{zona costera} \end{cases}$$

de forma que ambas variables ficticias están definidas de modo que el municipio i puede pertenecer a una zona costera o no, calcule el valor esperado de la variable endógena condicionado a la pertenencia a cada grupo. Indique si la especificación es o no correcta para los modelos (4.8) y (4.9) sugeridos a continuación, así como qué problemas generaría la estimación de los mismos.

$$Y_i = \beta_1 + \beta_2 Z_{1i} + \beta_3 Z_{2i} + \beta_4 X_i + u_i \tag{4.8}$$

$$Y_i = \beta_1 + \beta_2 Z_{1i} + \beta_3 X_i + \beta_4 Z_{1i} X_i + \beta_5 Z_{2i} X_i + u_i \tag{4.9}$$

Solución

En el caso del modelo (4.8) los valores esperados condicionados a la pertenencia de un municipio a una zona costera y no costera vienen dados por:

$$E(Y_i / Z_{1i} = 1, Z_{2i} = 0) = (\beta_1 + \beta_2) + \beta_4 E(X_i)$$

$$E(Y_i / Z_{1i} = 0, Z_{2i} = 1) = (\beta_1 + \beta_3) + \beta_4 E(X_i)$$

Los parámetros β_2 y β_3 se interpretan como el cambio que se produce en el término independiente debido al hecho de que el municipio pertenezca o no a una zona costera. Este modelo no se puede estimar por MCO ya que, al incluir las dos variables ficticias, se incurre en un problema de multicolinealidad exacta con el término independiente. La solución aquí pasaría por especificar un nuevo modelo que incluyera una única dicotómica, o bien por especificar otro que contuviera las dos ficticias, pero eliminando el término constante.

En el caso del modelo (4.9) se procede de manera semejante para la obtención de los valores esperados.

$$E(Y_i / Z_{1i} = 1, Z_{2i} = 0) = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) E(X_i)$$

$$E(Y_i / Z_{1i} = 0, Z_{2i} = 1) = \beta_1 + (\beta_3 + \beta_5) E(X_i)$$

Este modelo tampoco se puede estimar, ya que plantea también un problema de multicolinealidad exacta en la forma $Z_{1i} X_i + Z_{2i} X_i = X_i$

EJERCICIO 4.14

Se ha planteado el siguiente modelo para explicar las ventas trimestrales Y de una empresa:

$$Y_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \qquad t = 1, 2, ..., T \tag{4.10}$$

donde

$$E(u_t) = 0 \quad E(u_t^2) = \sigma_u^2 \quad E(u_t u_{t-s}) = 0, \quad \forall s \neq 0$$

$$D_{jt} = \begin{cases} 1 & \text{si } t \text{ pertenece al trimestre } j\text{-ésimo} \\ 0 & \text{en otro caso} \end{cases} \qquad j = 1, 2, 3, 4$$

(a) Interprete los coeficientes del modelo (4.10).

(b) ¿Existe algún problema con la estimación del siguiente modelo alternativo?

$$Y_t = \beta_0 + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + v_t \text{ con } t = 1, 2, ..., T \tag{4.11}$$

donde

$$E(v_t) = 0 \quad E(v_t^2) = \sigma_v^2 \quad E(v_t v_{t-s}) = 0, \quad \forall s \neq 0$$

(c) Dado el siguiente modelo alternativo

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + w_t \text{ con } t = 1, 2, ..., T \tag{4.12}$$

donde

$$E(w_t) = 0 \quad E(w_t^2) = \sigma_w^2 \quad E(w_t w_{t-s}) = 0 \quad \forall s \neq 0$$

¿Existe algún tipo de relación entre los coeficientes de los modelos (4.10) y (4.12)?

Solución

(a) Al tratarse de un modelo lineal en donde todas las variables son dicotómicas y haberse mantenido una de ellas como referencia (la que hace referencia al primer trimestre), la interpretación de los coeficientes es la siguiente:

β_1: Ventas medias del primer trimestre

β_2: Diferencia entre las ventas medias del segundo trimestre y las del primero

β_3: Diferencia entre las ventas medias del tercer trimestre y las del primero

β_4: Diferencia entre las ventas medias del cuarto trimestre y las del primero

(b) Al introducir variables ficticias de forma aditiva se debe tener la precaución de no incluir la constante junto con todas las modalidades de dicha ficticia. El modelo (4.11) comete precisamente ese error. El problema al hacer algo así radica en que se incurre en un problema de multicolinealidad exacta, estando ante el siguiente caso de combinación lineal exacta: $cte_t = D_{1t} + D_{2t} + D_{3t} + D_{4t}$, con lo cual, el rango de la matriz X es inferior al número de parámetros a estimar, $\rho(X) < 5$, el determinante $|X'X|$ vale cero, no se puede calcular $(X'X)^{-1}$ y, en consecuencia, no se puede obtener una única solución para $\hat{\beta}_{MCO}$.

(c) Los valores esperados para las ventas en cada uno de los trimestres según el modelo (4.10) son

$$E(Y/D_1 = 1) = \beta_1$$
$$E(Y/D_2 = 1) = \beta_1 + \beta_2$$
$$E(Y/D_3 = 1) = \beta_1 + \beta_3$$
$$E(Y/D_4 = 1) = \beta_1 + \beta_4$$

Por su parte, los valores esperados según el modelo (4.12) son

$$E(Y/D_1 = 1) = \alpha_1$$
$$E(Y/D_2 = 1) = \alpha_2$$
$$E(Y/D_3 = 1) = \alpha_3$$
$$E(Y/D_4 = 1) = \alpha_4$$

Por tanto, la relación existente entre los coeficientes de ambos modelos es

$$\alpha_1 = \beta_1$$
$$\alpha_2 = \beta_1 + \beta_2$$
$$\alpha_3 = \beta_1 + \beta_3$$
$$\alpha_4 = \beta_1 + \beta_4$$

EJERCICIO 4.15

Se ha estimado el siguiente modelo de ventas diarias de una tienda que abre de lunes a viernes, siendo D_i variables ficticias que toman valor 1 según el i-ésimo día de la semana al que corresponda, siendo el viernes la variable de referencia. Las ventas dependen del día de la semana y de la publicidad que se realice diariamente repartiendo tarjetas a la salida del centro comercial:

$$\ln\left(\hat{V}_t\right) = 5 - 0.5D_{1t} - 0.8D_{2t} - 0.4D_{3t} - 0.2D_{4t} + 0.5\ln\left(Pub_t\right)$$

(a) ¿Cuál es la diferencia de ventas del lunes con respecto al martes? Interprete el resultado.

(b) ¿Cuáles serían los valores de los coeficientes estimados si tomamos como referencia el lunes en lugar del viernes?

(c) ¿Cuáles serían los coeficientes estimados si planteamos el modelo sin constante?

Solución

(a) Las ventas estimadas para el lunes y el martes se expresan como

$$\hat{V}_l = e^{5-0.5+0.5\ln Pub} = e^{4.5}e^{0.5\ln Pub} = 90\, e^{0.5\ln Pub}$$
$$\hat{V}_m = e^{5-0.8+0.5\ln Pub} = e^{4.2}e^{0.5\ln Pub} = 66.6\, e^{0.5\ln Pub}$$

y de aquí, $\hat{V}_l/\hat{V}_m = 1.35$, lo que nos indica que las ventas de los lunes son un 35% superior a las de los martes.

(b) Para contestar a esta pregunta únicamente debemos recordar que los valores esperados de las ventas para cada día deben coincidir en ambos modelos.

En el caso del modelo con referencia el viernes:

$$\ln\left(\hat{V}_t\right) = 5 - 0.5D_{1t} - 0.8D_{2t} - 0.4D_{3t} - 0.2D_{4t} + 0.5\left(\ln Pub_t\right)$$

los valores estimados para cada modalidad se obtienen a partir de:

$$\ln\left(\hat{V}_t\right)/lunes = 5 - 0.5 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/martes = 5 - 0.8 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/miércoles = 5 - 0.4 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/jueves = 5 - 0.2 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/viernes = 5 + 0.5\ln\left(Pub_t\right)$$

En el caso del modelo con el lunes como referencia:

$$\ln V_t = \alpha + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \beta_6 \ln Pub_t + u_t$$

los valores estimados para cada modalidad se derivan de:

$$\ln\left(\hat{V}_t\right)/lunes = \hat{\alpha} + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/martes = \hat{\alpha} + \hat{\beta}_2 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/miércoles = \hat{\alpha} + \hat{\beta}_3 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/jueves = \hat{\alpha} + \hat{\beta}_4 + 0.5\ln\left(Pub_t\right)$$
$$\ln\left(\hat{V}_t\right)/viernes = \hat{\alpha} + \hat{\beta}_5 + 0.5\ln\left(Pub_t\right)$$

Sabiendo que los valores estimados en ambos modelos deben coincidir:

$$\hat{\alpha} = 5 - 0.5 = 4.5$$
$$\hat{\alpha} + \hat{\beta}_2 = 5 - 0.8 = 4.2 \Rightarrow \hat{\beta}_2 = 4.2 - 4.5 = -0.3$$
$$\hat{\alpha} + \hat{\beta}_3 = 5 - 0.4 = 4.6 \Rightarrow \hat{\beta}_3 = 4.6 - 4.5 = 0.1$$
$$\hat{\alpha} + \hat{\beta}_4 = 5 - 0.2 = 4.8 \Rightarrow \hat{\beta}_4 = 4.8 - 4.5 = 0.3$$
$$\hat{\alpha} + \hat{\beta}_5 = 5 \Rightarrow \hat{\beta}_5 = 5 - 4.5 = 0.5$$

Por tanto, tomando el lunes como día de referencia, el modelo estimado es:

$$\ln\left(\hat{V}_t\right) = 4.5 - 0.3D_{2t} + 0.1D_{3t} + 0.3D_{4t} + 0.5D_{5t} + 0.5\ln\left(Pub_t\right)$$

(c) El modelo sin constante será:

$$\ln\left(V_t\right) = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \beta_6 \ln\left(Pub_t\right) + u_t$$

Sus valores estimados para cada día valdrán:

$$\ln\left(\hat{V}_t\right)/lunes = \hat{\beta}_1 + \hat{\beta}_6 \ln\left(Pub_t\right)$$

$$\ln\left(\hat{V}_t\right)/martes = \hat{\beta}_2 + \hat{\beta}_6 \ln\left(Pub_t\right)$$

$$\ln\left(\hat{V}_t\right)/miércoles = \hat{\beta}_3 + \hat{\beta}_6 \ln\left(Pub_t\right)$$

$$\ln\left(\hat{V}_t\right)/jueves = \hat{\beta}_4 + \hat{\beta}_6 \ln\left(Pub_t\right)$$

$$\ln\left(\hat{V}_t\right)/viernes = \hat{\beta}_5 + \hat{\beta}_6 \ln\left(Pub_t\right)$$

Análogamente al apartado anterior, igualando términos para cada día, se obtiene:

$$\hat{\beta}_1 = 5 - 0.5 = 4.5$$

$$\hat{\beta}_2 = 5 - 0.8 = 4.2$$

$$\hat{\beta}_3 = 5 - 0.4 = 4.6$$

$$\hat{\beta}_4 = 5 - 0.2 = 4.8$$

$$\hat{\beta}_5 = 5$$

Por lo que el modelo estimado sin constante vendrá dado por:

$$\ln\left(\hat{V}_t\right) = 4.5D_{1t} + 4.2D_{2t} + 4.6D_{3t} + 4.8D_{4t} + 5D_{5t} + 0.5\ln\left(Pub_t\right)$$

EJERCICIO 4.16

Se dispone de un conjunto de datos correspondiente a N familias de doce países de la Unión Europea y de las siguientes variables: gastos familiares, nivel de renta y número de miembros de la familia. Se cree que, por razones socio-culturales, el comportamiento de las familias respecto al gasto puede ser distinto al del resto para los países mediterráneos de la UE.

(a) Proponga un modelo que permita tener en cuenta que el gasto autónomo varía según los individuos sean de un país mediterráneo o no. Interprete el coeficiente y especifique el contraste que realizaría para constatar si el gasto autónomo varía o no según el país sea o no mediterráneo (indicando claramente cuál sería la hipótesis nula y el estadístico de prueba que emplearía).

(b) Especifique un modelo en el que incluya, además de la variación del gasto autónomo según el tipo de país, el distinto comportamiento de la variable renta sobre el gasto, es decir la propensión marginal a consumir según el país

sea mediterráneo o no. Interprete el coeficiente y especifique claramente cuáles son las hipótesis que formularía para realizar los siguientes contrastes:

- La zona geográfica no influye en la propensión marginal a consumir. (Especifique el estadístico que emplearía para realizar este contraste).
- La zona geográfica no influye en la determinación del gasto familiar en la UE. (Especifique el estadístico que emplearía para realizar este contraste).

Solución

(a) El modelo propuesto es:

$$Gto_i = \beta_1 + \beta_2 D_i + \beta_3 Rta_i + \beta_4 Nfam_i + u_i$$

donde

D es una variable dicotómica que toma valor 1 si el país es mediterráneo y 0 en caso contrario.

β_2 se interpreta como el diferencial de gasto que se produce en los países mediterráneos con respecto a los no mediterráneos.

El contraste de hipótesis para comprobar si el gasto autónomo varía según el país sea o no mediterráneo será:

$$\left.\begin{array}{l} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{array}\right\}$$

Se trata de un contraste de significación individual que se resuelve con el estadístico de contraste:

$$t = \frac{\hat{\beta}_2}{S\left(\hat{\beta}_2\right)}$$

que se distribuye como un t-Student $\left(t_{N-k}\right)$ bajo la hipótesis nula como cierta.

(b) Para incluir la existencia de distintas propensiones marginales con respecto a la renta, hay que añadir una dicotómica multiplicativa al modelo:

$$Gto_i = \beta_1 + \beta_2 D_i + \beta_3 Rta_i + \beta_4 Nfam_i + \beta_5 D_i Rta_i + u_i$$

El contraste de hipótesis para determinar que la zona geográfica no influye en la propensión marginal a consumir del modelo se realiza a partir de:

$$\left.\begin{array}{l} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{array}\right\}$$

Para contrastar la no influencia de la zona geográfica en ningún caso, el contraste será:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_5 = 0 \\ H_1: \beta_2 \quad \text{y/o} \quad \beta_5 \neq 0 \end{array}\right\}$$

Se trata de un contraste de nulidad de subconjunto de parámetros. El estadístico de contraste a utilizar podría ser:

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)},$$

que bajo el cumplimiento de la hipótesis nula se distribuye como una F de Fisher-Snedecor con $m = 2$ y $N - k$ grados de libertad.

EJERCICIO 4.17

Se desea estudiar el número de veces que una persona va al médico de familia en atención privada al año (*médico*). Para ello se supone que las variables relevantes son, entre otras, la renta (*renta*), el estado civil (*casado*, *divorciado*, *viudo*, *soltero*), el sexo (*mujer*, *hombre*) y el tener o no seguro médico privado (*seguro_sí*, *seguro_no*). Las variables estado civil, sexo y seguro médico se incorporan al modelo como variables ficticias.

(a) Especifique la ecuación más sencilla que permita estudiar el número de veces que una persona va al médico, tomando en consideración todas las variables mencionadas. Las modalidades de referencia de las variables cualitativas son: estar soltero, ser hombre y tener seguro médico privado. Considere tan solo el caso de la hipótesis aditiva.

(b) Interprete los coeficientes de dicha ecuación.

(c) ¿Qué signos esperaría para los coeficientes de las variables renta y seguro? Argumente la respuesta.

(d) ¿Cuál sería la hipótesis nula a plantear para comprobar si existen diferencias significativas entre viudos y divorciados en cuanto a su asistencia al médico?

(e) ¿Y si se quisiera comprobar si hay diferencias significativas entre viudos, solteros y divorciados?

(f) Si se plantea la posibilidad de que el estado civil no afecte al número de veces que se asiste el médico al año, ¿cómo plantearía la hipótesis nula del contraste?

(g) Plantee el modelo del apartado (a) considerando la hipótesis mixta en relación a la variable seguro al interactuar con la renta.

Solución

(a) La ecuación más sencilla es la siguiente:

$$médico_i = \beta_1 + \beta_2 casado_i + \beta_3 divorciado_i + \beta_4 viudo_i + \\ + \beta_5 mujer_i + \beta_6 seguro\_no_i + \beta_7 renta_i + u_i$$

(b) La interpretación de los coeficientes es la siguiente:

β_1: Es el número medio de veces que un hombre soltero, con seguro médico privado y sin renta iría al médico.

β_2: Es la diferencia entre el número de veces que va al médico un hombre soltero con seguro médico privado y un hombre casado con seguro médico privado. El hombre soltero irá en promedio β_1 veces y el hombre casado $\beta_1 + \beta_2$ veces también en promedio.

β_3: Es la diferencia entre el número de veces que va al médico un hombre soltero con seguro médico privado y un hombre divorciado con seguro médico privado. El hombre soltero irá un promedio de β_1 veces y el hombre divorciado un promedio de $\beta_1 + \beta_3$ veces.

β_4: Es la diferencia entre el número de veces que va al médico un hombre soltero con seguro médico privado y un hombre viudo con seguro médico privado. El hombre soltero irá un promedio de β_1 veces y el hombre viudo un promedio de $\beta_1 + \beta_4$ veces.

β_5: Es la diferencia entre el número de veces que va al médico privado un hombre soltero con seguro médico privado y una mujer soltera con seguro médico privado. El hombre irá un promedio de β_1 veces y la mujer un promedio de $\beta_1 + \beta_5$ veces.

β_6: Es la diferencia entre el número de veces que va al médico un hombre soltero con seguro médico privado y un hombre soltero sin seguro

médico privado. El hombre con seguro médico privado irá un promedio de β_1 veces y el hombre sin seguro médico privado irá $\beta_1 + \beta_6$ veces.

β_7: Es la variación que se produce en las veces que una persona va al médico al año ante una variación unitaria en su renta.

(c) Se esperaría que el signo de la renta fuera positivo y el del seguro negativo. En el caso de la renta, cuanto mayor es el nivel de renta mayor será la proclividad a ir al médico privado, aún cuando la dolencia se pudiera considerar a priori leve. En el caso de la variable seguro, dado que recoge la no tenencia de seguro privado, es de esperar que las personas que no tengan seguro tiendan a ir menos al médico en comparación con aquellos que sí tienen seguro médico privado.

(d) La hipótesis nula a plantear sería:

$$H_0: \beta_3 = \beta_4$$

(e) En ese caso, la hipótesis nula a plantear sería:

$$H_0: \beta_3 = \beta_4 = 0$$

(f) En este caso, la hipótesis nula a plantear sería:

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$

(g) El modelo, considerando la hipótesis mixta en relación a la variable seguro, será:

$$médico_i = \beta_1 + \beta_2 casado_i + \beta_3 divorciado_i + \beta_4 viudo_i + \beta_5 mujer_i + \\ + \beta_6 seguro\_no_i + \beta_7 renta_i + \beta_8 renta_i \cdot seguro\_no_i + u_i$$

EJERCICIO 4.18

Se dispone del siguiente modelo estimado, donde se explica el logaritmo de los precios aplicados por una empresa de decoración, $\ln(PY)$, en función del logaritmo del precio de la competencia, $\ln(PX)$, y del número de horas de trabajo, *horas*:

$$\ln\left(\widehat{PY}_i\right) = -2\ln(PX_i) + 0.05\, horas_i$$

Se piensa que el nivel de precisión requerido para realizar el trabajo también influye en la determinación del precio.

(a) Introduzca esta nueva variable en el modelo de forma aditiva considerando 3 categorías: precisión baja, media y alta, y tomando como categoría de referencia el nivel de precisión bajo.

(b) ¿Cómo contrastaría el hecho de que el precio del servicio realizado depende del nivel de precisión requerido?

(c) ¿Cómo contrastaría que existen diferencias en el nivel medio de los precios según el nivel de precisión sea alto o medio?

(d) Interprete el coeficiente de la variable *horas*.

(e) ¿Cómo introduciría el hecho de que el número de horas de trabajo pueda influir más sobre el precio en los trabajos de precisión alta que sobre los de precisión baja?

Solución

(a) La nueva variable es el nivel de precisión. Definimos:

PM como variable dicotómica que toma el valor 1 si la precisión es media y 0 en otro caso.

PA como variable dicotómica que toma el valor 1 si la precisión es alta y 0 en otro caso.

PB como la modalidad de referencia (precisión baja).

El modelo que incluye esta nueva variable es:

$$\ln(PY)_i = \beta_1 + \beta_2 \ln(PX_i) + \beta_3 horas_i + \beta_4 PM_i + \beta_5 PA_i + u_i$$

(b) El hecho de que el precio del servicio realizado dependa del nivel de precisión se analizará con el siguiente contraste, que recoge en su hipótesis nula la inexistencia de diferencias entre los niveles de precisión:

$$\left.\begin{array}{l} H_0: \beta_4 = \beta_5 = 0 \\ H_1: \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0 \end{array}\right\}$$

(c) Para contrastar si existen diferencias en el nivel medio de los precios según el nivel de precisión sea alto o medio el contraste a plantear es:

$$\left.\begin{array}{l} H_0: \beta_4 = \beta_5 \\ H_1: \beta_4 \neq \beta_5 \end{array}\right\}$$

(d) Por cada hora de trabajo, el precio del servicio aumenta un 5%.

(e) Introduciendo la variable dicotómica PA de manera multiplicativa con la variable horas, el modelo quedaría especificado como sigue:

$$\ln(PY_i) = \beta_1 + \beta_2 \ln(PX_i) + \beta_3 horas_i + \beta_4 PM_i + \beta_5 PA_i + \beta_6 horas_i \cdot PA_i + u_i$$

EJERCICIO 4.19

Se dispone de información correspondiente a N familias de los 15 países de la UE sobre las variables gasto turístico (Y), nivel de renta (X_2), número de miembros de la familia (X_3) y zona geográfica de residencia (X_4). Esta última variable hace referencia a la residencia en zona mediterránea o zona no mediterránea.

(a) Especifique dos modelos distintos a través de los cuales se explique la variabilidad en el gasto familiar en función de todos estos factores. Interprete los parámetros de ambos modelos.

(b) ¿Cómo contrastaría en ambos modelos si la variable cualitativa Zona geográfica de residencia (X_4) es relevante a la hora de determinar el gasto familiar? Indique las hipótesis nula y alternativa así como el tipo de contraste a realizar.

(c) Escriba las ecuaciones correspondientes a los valores esperados de gasto, condicionado a cada uno de los posibles valores de la variable Zona de residencia, y comente los resultados.

Solución

(a) A partir de la variable X_4 planteamos la creación de dos variables dicotómicas alternativas con el fin de poder especificar dos modelos distintos:

DM_i tomará valor 1 si la familia i-ésima vive en zona mediterránea y valor 0 en caso contrario.

DNM_i tomará valor 1 si la familia i-ésima vive en zona no mediterránea y valor 0 en caso contrario.

Los dos posibles modelos podrían especificarse como sigue:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 DM_i + u_i \tag{4.13}$$

$$Y_i = \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 DM_i + \beta_5 DNM_i + v_i \tag{4.14}$$

La interpretación de los coeficientes es la siguiente:

β_2 es la variación esperada del gasto turístico ante un incremento unitario de la renta (propensión marginal al gasto) *caeteris paribus.*

β_3 es la variación esperada del gasto turístico por cada miembro adicional de la familia que viaja *caeteris paribus.*

La diferencia entre los modelos (4.13) y (4.14), viene dada por cómo se introduce la variable cualitativa "Zona de procedencia". La interpretación de los parámetros cambia según consideremos el (4.13) o el (4.14) en los que se introducen una o dos dicotómicas, evitando en ambos casos incurrir en un problema de multicolinealidad exacta.

β_4 en el modelo (4.13), mide el diferencial promedio de gasto turístico entre los turistas que proceden de un país de la zona mediterránea y los que proceden de cualquier otra zona de la UE, mientras que en este mismo modelo, β_1 representa el gasto promedio de un turista residente en zona no mediterránea.

En el modelo (4.14) la interpretación de β_4 es análoga a la realizada para el modelo (4.13), aunque en este caso es el valor del coeficiente β_5 el que coincidirá con el de β_1 del modelo (4.13).

(b) Para contrastar si la variable cualitativa "Zona geográfica" es relevante a la hora de explicar el gasto turístico familiar, es decir, si podemos esperar gastos diferentes de las familias según que procedan de unas zonas de la UE u otras, hemos de referirnos al modelo con el que trabajamos.

En el modelo (4.13) se haría como sigue:

$$\left.\begin{array}{l} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{array}\right\}$$

Se trata de un contraste de significación individual cuyo estadístico de prueba es

$$t = \frac{\hat{\beta}_3}{S\left(\hat{\beta}_3\right)}$$

y que, bajo la hipótesis nula, se distribuye como una t-Student con $N-k$ grados de libertad.

En el modelo (4.14), si queremos contrastar que la Zona geográfica de residencia tiene efecto sobre el gasto turístico, se haría con las hipótesis nula y alternativa siguientes:

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \beta_3 \neq 0 \ \text{ y/o } \ \beta_4 \neq 0 \end{array}\right\}$$

Éste podría ser un contraste de restricciones lineales homogéneas, donde las matrices necesarias para la obtención del estadístico de prueba serían

$$R=[0 \ \ 0 \ \ 1 \ \ -1] \quad \text{y} \quad r=0.$$

El estadístico de prueba, que se distribuye bajo la hipótesis nula como una F de Fisher-Snedecor, viene dado por

$$F = \frac{\left(R\hat{\beta}-r\right)'\left(R\left(X'X\right)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)/2}{\hat{\sigma}_u^2}.$$

(c) En el modelo (4.13), dados un número de miembros y un nivel de renta determinados, los valores esperados del gasto, condicionados a la zona geográfica de residencia, son:

$$E\left(Y_i / DM_i = 1\right) = \left(\beta_1 + \beta_4\right) + \beta_2 E\left(X_{2i}\right) + \beta_3 E\left(X_{3i}\right)$$
$$E\left(Y_i / DM_i = 0\right) = \beta_1 + \beta_2 E\left(X_{2i}\right) + \beta_3 E\left(X_{3i}\right)$$

donde:

β_1 recoge el gasto autónomo de las familias que residen en zonas no mediterráneas.

β_4 recoge el diferencial de gasto esperado en las familias de zonas mediterráneas con respecto al resto, *caeteris paribus*.

En el modelo (4.14) los valores esperados del gasto, condicionados a la zona geográfica de residencia son:

$$E\left(Y_i / DM_i = 1\right) = \beta_4 + \beta_2 E\left(X_{2i}\right) + \beta_3 E\left(X_{3i}\right)$$
$$E\left(Y_i / DNM_i = 1\right) = \beta_5 + \beta_2 E\left(X_{2i}\right) + \beta_3 E\left(X_{3i}\right)$$

donde:

β_4 mide el gasto autónomo para las familias de la zona mediterránea.

β_5 mide el correspondiente a las familias de zonas no mediterráneas.

EJERCICIO 4.20

Para estudiar si existen diferencias salariales entre hombres y mujeres en una industria, se toman 100 individuos al azar y se considera una función de regresión que relaciona a la vez el salario con los años de estudio y experiencia laboral para hombres y mujeres utilizando una variable ficticia.

$$W_i = \beta_1 + \beta_2 M_i + \beta_3 S_i + \beta_4 S_i \cdot M_i + \beta_5 E_i + \beta_6 E_i \cdot M_i + u_i \qquad (4.15)$$

donde:

W_i = salario hora del individuo i en cientos de euros

M_i = variable dicotómica que toma valor 1 si el individuo i es mujer y 0 si es hombre

S_i = años de estudio del individuo i

E_i = años de experiencia laboral del individuo i

La estimación de dicho modelo es la siguiente, donde los valores entre paréntesis muestran los errores estándar de los coeficientes estimados:

$$\widehat{W}_i = \underset{(0.10)}{2} - \underset{(0.20)}{0.5} M_i + \underset{(0.01)}{0.09} S_i - \underset{(0.10)}{0.02} S_i \cdot M_i + \underset{(0.05)}{0.1} E_i - \underset{(0.02)}{0.05} E_i \cdot M_i; \quad e'e = 240$$

Por otro lado, se estima el modelo imponiendo la restricción $\beta_2 = \beta_4 = \beta_6 = 0$, obteniendo una suma cuadrática de los errores de 360.

(a) Interprete cada uno de los coeficientes del modelo y escriba las funciones de regresión estimadas para hombres y para mujeres.

(b) Contraste, al 5% de significación, que los modelos de salarios son idénticos para hombres y para mujeres.

Solución

(a) La función estimada para los hombres es:

$$\hat{W}_i = \hat{\beta}_1 + \hat{\beta}_3 S_i + \hat{\beta}_5 E_i$$

La función estimada para las mujeres es:

$$\hat{W}_i = \left(\hat{\beta}_1 + \hat{\beta}_2\right) + \left(\hat{\beta}_3 + \hat{\beta}_4\right) S_i + \left(\hat{\beta}_5 + \hat{\beta}_6\right) E_i$$

La interpretación de los coeficientes del modelo (4.15) es, *caeteris paribus*, la siguiente:

β_1: Salario medio de un hombre con 0 años de educación y 0 años de experiencia.

β_2: Diferencia entre el salario medio de un hombre y una mujer con 0 años de educación y experiencia.

β_3: Variación del salario de los hombres al aumentar en una unidad los años de estudio, manteniendo constante la experiencia.

β_4: Diferencia entre el salario de hombres y mujeres ante un aumento unitario de los años de estudio, manteniendo constante la experiencia.

β_5: Variación en el salario de los hombres al aumentar en una unidad los años de experiencia.

β_6: Diferencia entre el salario de hombres y mujeres ante un aumento unitario de los años de experiencia.

(b) El contraste de hipótesis a plantear es:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_4 = \beta_6 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

y el estadístico de prueba se calcula a partir de la expresión

$$F = \frac{(e'e_r - e'e_{nr})/q}{e'e_{nr}/(N-k)}$$

Sustituyendo los valores obtenemos el siguiente valor para el estadístico de contraste:

$$F = \frac{(360-240)/3}{240/(100-6)} = 15.67$$

Comparando este valor con el tabulado al 5% de significación para una $F^{0.05}_{q,N-k} = F^{0.05}_{3,100-6} = 2.61$, comprobamos que el estadístico de contraste cae en la región de rechazo, por lo que no podemos afirmar que las ecuaciones salariales de hombres y mujeres sean idénticas.

EJERCICIO 4.21

Interprete en términos precisos —y utilizando las unidades de medida indicadas—, los coeficientes de cada uno de los siguientes modelos, sabiendo que:

W mide el salario mensual en euros

AE mide los años de educación

EL mide los años de experiencia laboral

DM es una dicotómica que toma el valor 1 cuando es una mujer y 0 si es hombre

$$\hat{W}_i = 477.12 - 0.46 \cdot AE_i + 30.12 \cdot EL_i - 34.14 \cdot DM_i \quad (4.16)$$

$$\ln\left(\hat{W}_i\right) = 6.19 + 0.0003 \cdot AE_i + 0.046 \cdot EL_i - 0.046 \cdot DM_i \quad (4.17)$$

$$\hat{W}_i = 357.09 + 4.20 \ln\left(AE_i\right) + 167.38 \ln\left(EL_i\right) - 44.33 \cdot DM_i \quad (4.18)$$

$$\ln\left(\hat{W}_i\right) = 5.98 + 0.02 \ln\left(AE_i\right) + 0.26 \ln\left(EL_i\right) - 0.06 \cdot DM_i \quad (4.19)$$

Solución

La interpretación de los coeficientes del modelo (4.16), al tratarse de un modelo con especificación lineal, será la siguiente:

$\beta_1 = 477.12$: El salario esperado de un hombre sin ninguna experiencia y con 0 años de educación es de 477.12 €.

$\beta_2 = -0.46$: El aumento de un año en los años de educación implica, *caeteris paribus*, una disminución de 0.46 € en el salario mensual.

$\beta_3 = 30.12$: El aumento de un año en los años de experiencia implica, *caeteris paribus*, un aumento de 30.12 € en el salario mensual.

$\beta_4 = -34.14$: El diferencial de salarios entre los hombres y las mujeres, *caeteris paribus*, es de 34.14 € mensuales a favor de los hombres.

En el modelo (4.17), al tratarse de un modelo con especificación *log-lin*, la interpretación de los coeficientes es la siguiente:

$\beta_1 = 6.19$: El valor esperado del logaritmo del salario de un hombre sin ninguna experiencia y con 0 años de educación es 6.19; por tanto, el valor esperado de dicho salario es de $e^{6.19} = 487.85$€.

$\beta_2 = 0.0003$: El aumento de un año en los años de educación implica, *caeteris paribus*, un aumento del 0.03% en el salario mensual.

$\beta_3 = 0.046$: El aumento de un año en los años de experiencia implica, *caeteris paribus*, un aumento del 4.6% en el salario mensual.

$\beta_4 = -0.046$: El diferencial del "logaritmo de los salarios" entre los hombres y las mujeres, *caeteris paribus*, es del 4.6% a favor de los hombres. Si queremos ver la diferencia en términos de "salario", ésta es de $\left(e^{-0.046} - 1\right) \cdot 100 = -4.496\,\%$, lo que implica que el salario de las mujeres es un 4.5% inferior al de los hombres, o que éstas ganan el $e^{-0.046} \cdot 100 = 95.5\%$ del salario de los hombres.

El modelo (4.18) presenta una especificación *lin-log*, por lo que la interpretación de sus coeficientes vuelve a ser diferente, en este caso es la siguiente:

$\beta_1 = 357.09$: El salario esperado de un hombre sin ninguna experiencia y con 0 años de educación es de 357.09 €.

$\beta_2 = 4.20$: El aumento de un 1% en los años de educación implica, *caeteris paribus*, un aumento de 0.042 € en el salario mensual.

$\beta_3 = 167.38$: El aumento de un 1% en los años de experiencia implica, *caeteris paribus*, un aumento de 1.67 € en el salario mensual.

$\beta_4 = -44.33$: El diferencial de salarios entre los hombres y las mujeres, *caeteris paribus*, es de 44.33 € mensuales a favor de los hombres.

Por último, el modelo (4.19) presenta una especificación *doble-log*, por lo que la interpretación de sus coeficientes sería la siguiente:

$\beta_1 = 5.98$: El valor esperado del logaritmo del salario de un hombre sin ninguna experiencia y con 0 años de educación es 5.98; por tanto, el valor esperado de dicho salario es de $e^{5.98} = 395.44\,€$.

$\beta_2 = 0.02$: Un aumento de un 1% en los años de educación implica, *caeteris paribus*, un aumento del 0.02% en el salario mensual.

$\beta_3 = 0.26$: Un aumento de un 1% en los años de experiencia implica, *caeteris paribus*, un aumento del 0.26% en el salario mensual.

$\beta_4 = -0.06$: El diferencial del "logaritmo de los salarios" entre los hombres y las mujeres, *caeteris paribus*, es del 6% a favor de los hombres. Si queremos ver la diferencia en términos de "salario", ésta es de $\left(e^{-0.06} - 1\right) \cdot 100 = -5.823\,\%$, lo que implica que el salario de las mujeres es un 5.8% inferior al de los hombres, o que éstas ganan el $e^{-0.06} \cdot 100 = 94.2\%$ del salario de los hombres.

EJERCICIO 4.22

Se estudian los efectos que, sobre las pernoctaciones hoteleras por habitante en diferentes provincias (P), tienen la renta per capita (R) así como tres variables dicotómicas relativas a si la provincia está ubicada en la costa (ZC) o si pertenece a un archipiélago (ZA). Si la provincia pertenece al interior (ZI) queda como modalidad de referencia. Se obtiene la estimación $\hat{P}_i = -0.24 + 0.24R_i + 0.60ZC_i + 1.6ZA_i$.

Si dispone de la siguiente información:

$$\hat{\sigma}_u^2 = 1.58 \quad t_{51-4}^{0.05} = 2.01 \quad t_{51-4}^{0.10} = 1.67$$

$$(X'X)^{-1} = \begin{pmatrix} 0.884 & -0.0850 & -0.056 & -0.0010 \\ -0.085 & 0.0090 & 0.002 & 0.0001 \\ -0.056 & 0.0020 & 0.086 & -0.0500 \\ -0.001 & 0.0001 & -0.050 & 0.3830 \end{pmatrix}$$

(a) ¿Cómo verificaría la hipótesis de que la zona geográfica en general (de costa o archipiélago) no tiene efecto sobre las pernoctaciones (P)? Explique claramente cuales serían las hipótesis nula y alternativa y realice el contraste oportuno que le permitiría concluir la existencia de un efecto "zona geográfica".

(b) Si se especificase un modelo alternativo al anterior, que incluyese una variable que recogiese el efecto "Ciudad Patrimonio de la Humanidad" (PH), cuando dicha provincia tiene una o más ciudades con esta denominación, ¿cuál sería el contraste adecuado para decidir cuál de los dos modelos —(4.20) o (4.21)— sería el válido? Especifique cuáles serían la hipótesis nula y alternativa, así como la expresión del estadístico de prueba. ¿Qué implicaría el rechazo de la hipótesis nula?

$$P_i = \beta_1 + \beta_2 R_i + \beta_3 ZC_i + \beta_4 ZA_i + \beta_5 PH_i + u_i \tag{4.20}$$

$$P_i = \beta_1 + \beta_2 R_i + \beta_3 ZC_i + \beta_4 ZA_i + u_i \tag{4.21}$$

Solución

(a) La hipótesis a contrastar la podemos expresar como

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \text{No } H_0 \text{ (existe efecto "zona geográfica")} \end{array}\right\}$$

lo que nos permite obtener

$$R\beta = r \Rightarrow \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

El estadístico de contraste vendría dado por

$$F = \frac{\left(\left(R\hat{\beta} - r \right)' \left(R\left(X'X \right)^{-1} R' \right)^{-1} \left(R\hat{\beta} - r \right) \right) / q}{e'e / \left(N - k \right)}$$

para cuyo cálculo hemos de obtener previamente

$$R\beta = r \Rightarrow \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -0.24 \\ 0.24 \\ 0.60 \\ 1.60 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow$$

$$\Rightarrow R\beta - r = \begin{pmatrix} 0.60 \\ 1.60 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.60 \\ 1.60 \end{pmatrix}$$

$$\Rightarrow \left(R\beta - r \right)' = \begin{pmatrix} 0.60 & 1.60 \end{pmatrix}$$

además de

$$\left(R\left(X'X \right)^{-1} R' \right) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.884 & -0.085 & -0.056 & -0.001 \\ -0.085 & 0.009 & 0.002 & 0.000 \\ -0.056 & 0.002 & 0.086 & -0.050 \\ -0.001 & 0.000 & -0.050 & 0.383 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.09 & -0.05 \\ -0.05 & 0.38 \end{pmatrix} \Rightarrow \left(R\left(X'X \right)^{-1} R' \right)^{-1} = \begin{pmatrix} 12.55 & 1.64 \\ 1.64 & 2.82 \end{pmatrix}$$

Por tanto, el numerador del estadístico de contraste valdrá

$$\begin{pmatrix} 0.60 & 1.60 \end{pmatrix} \begin{pmatrix} 12.55 & 1.64 \\ 1.64 & 2.82 \end{pmatrix} \begin{pmatrix} 0.60 \\ 1.60 \end{pmatrix} = \begin{pmatrix} 10.17 & 5.5 \end{pmatrix} \begin{pmatrix} 0.60 \\ 1.60 \end{pmatrix} = 14.92$$

y, sin más que sustituir en la expresión del estadístico de contraste, obtenemos

$$F = \frac{\left(\left(R\hat{\beta} - r\right)'\left(R\left(X'X\right)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right)\right)/q}{e'e/(N-k)} = \frac{14.92/2}{1.58} = 4.72$$

Como $F_{2,47}^{0.05} \approx 3.2$ rechazamos la hipótesis nula, por lo que existe un "efecto zona geográfica".

(b) Una posibilidad para decidir sería la siguiente:

$$\left.\begin{array}{l} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{array}\right\}$$

El estadístico de prueba para resolver el contraste es:

$$F = \frac{(e'e_R - e'e_{NR})/m}{e'e_{NR}/(N-k)}$$

que, si se cumple la hipótesis nula, se distribuye como una F de Fisher-Snedecor con 1 y $N-k$ grados de libertad.

Otra posibilidad consistiría en contrastar la nulidad del coeficiente usando como estadístico de prueba $\left(\hat{\beta}_j - \beta_0\right)/S\left(\hat{\beta}_j\right)$ que, bajo la hipótesis nula como cierta, se distribuye como una t-Student con $N-k$ grados de libertad.

Si se rechaza la Hipótesis nula implicaría que el modelo correcto sería el modelo (4.20), es decir, que existe un efecto significativo de la declaración de ciudades como Patrimonio de la Humanidad sobre las pernoctaciones hoteleras.

EJERCICIO 4.23

La Teoría del Capital Humano establece que los salarios de los trabajadores son función del número de años de educación ($EDUC$) y de los años de experiencia laboral (EXP). Con una muestra de 1993 individuos se obtiene el siguiente modelo estimado, en donde, entre paréntesis, figura la desviación típica de los estimadores:

$$\ln\left(\widehat{W}_i\right) = \underset{(0.124)}{11.2} + \underset{(0.005)}{0.014}\,EDUC_i + \underset{(0.003)}{0.009}\,EXP_i; \quad e'e = 483.058$$

(a) ¿Cómo se interpreta el coeficiente que acompaña a la variable relativa al número de años de estudios?

(b) ¿Es estadísticamente significativa la variable *EDUC*?

(c) Si se supone que el efecto sobre los salarios de cada año adicional de educación es el doble que el de los años adicionales de experiencia, especifique el modelo correcto a estimar. Si se obtienen los siguientes resultados del modelo restringido en los que la variable *NUEVA* se refiere a la variable antes mencionada, contraste al 5% esta hipótesis.

$$\ln\left(\widehat{W}_i\right) = \underset{(0.071)}{11.17} + \underset{(0.002)}{0.009}\, NUEVA_i; \qquad e'^{*}e^{*} = 483.092$$

(d) La Teoría del Capital Humano establece que la experiencia laboral no es lineal sino que presenta rendimientos decrecientes, por lo que sugiere estimar especificaciones cuadráticas. Especifique el modelo correcto bajo esta hipótesis. Indique cuál sería la hipótesis nula si se quiere contrastar que dicho efecto cuadrático existe y señale, asimismo, cuál sería la hipótesis a contrastar si se quiere conocer si la experiencia tiene efecto o no sobre los salarios.

Solución

(a) Al ser un modelo *log-lin* el coeficiente que acompaña a la variable relativa al número de años de educación (*EDUC*) se interpreta como el cambio relativo que produce en Y un cambio absoluto en *EDUC*. Éste que viene dado por $\left(e^{\beta}-1\right)\cdot 100 = \left(e^{0.014}-1\right)\cdot 100 = 1.409\,\%$, lo que implica que el aumento en un año en los años de educación implica, *caeteris paribus*, un aumento del 1.409% en el salario mensual.

(b) Contrastamos la hipótesis nula de significatividad individual de la variable *EDUC* obteniendo 2.8 como valor del estadístico de contraste, que se distribuye bajo la hipótesis nula como una *t*-Student con $N-k$ grados de libertad.

$$t = \frac{\hat{\beta}_2 - \beta_2}{S\left(\hat{\beta}_2\right)} = \frac{0.014 - 0}{0.005} = 2.8$$

Dado que el tamaño muestral es elevado, el valor crítico coincide con el de una Normal tipificada, que, para el 5% de nivel de significación, vale 1.96. Ello nos lleva a afirmar que la variable es estadísticamente significativa a niveles estándar.

(c) El modelo que habría que estimar sería

$$\ln(W_i) = \beta_1 + \beta_2 (2 \cdot EDUC + EXP)_i + u_i$$

Para contrastar la hipótesis planteada tenemos:

$$\left.\begin{array}{l} H_0: \beta_2 = 2\beta_3 \\ H_1: \beta_2 \neq 2\beta_3 \end{array}\right\}$$

y utilizaremos como estadístico para el contraste el que viene dado por la expresión siguiente:

$$\frac{\dfrac{(e'e_R - e'e_{NR})}{q}}{\dfrac{e'e_{NR}}{N-k}} = \frac{\dfrac{(483.092 - 483.058)}{1}}{\dfrac{483.058}{1993-3}} = 0.14$$

Bajo la hipótesis nula como cierta, el estadístico de prueba se distribuye como una F de Fisher-Snedecor con 1 y 1990 grados de libertad. Al tomar el estadístico de prueba un valor muy cercano al cero, caerá en la región de aceptación de la hipótesis nula, por lo que la evidencia empírica no nos permite rechazar esta hipótesis, que indica que el efecto de cada año educación sobre los salarios duplica al de los años de experiencia.

(d) La especificación cuadrática sería la siguiente:

$$\ln(W_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXP_i + \beta_4 EXP_i^2 + u_i$$

Para contrastar si es o no estadísticamente significativa la variable EXP^2, realizaríamos el contraste siguiente:

$$\left.\begin{array}{l} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{array}\right\}$$

mientras que, para contrastar si la experiencia ejerce algún efecto sobre los salarios, debemos usar como hipótesis nula y alternativa las que siguen.

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

EJERCICIO 4.24

Una empresa pretende explicar el número de horas extraordinarias que realizan los trabajadores al mes, en función del número de hijos que tienen (X) y

de una variable ficticia con tres modalidades, según los trabajadores sean casados, solteros o separados. Sabiendo que $D1$ toma valor 1 si los trabajadores son solteros y $D2$ si son casados, y dada la siguiente estimación:

$$\ln \hat{Y}_i = 1.9 + 0.27 \ln(X_i) - 1.207 D1_i + 0.256 D2_i \qquad (4.22)$$

donde el tamaño de la muestra es $N = 9$ y la $SCE = 0.06449$ y sabiendo además que

$$(X'X)^{-1} = \begin{pmatrix} 0.61 & -0.140 & -0.58 & -0.41 \\ -0.14 & 0.069 & 0.12 & 0.04 \\ -0.58 & 0.120 & 0.75 & 0.40 \\ -0.41 & 0.040 & 0.40 & 0.49 \end{pmatrix}$$

(a) indique en cuánto aumentarán las horas extraordinarias los trabajadores si estos duplican el número de hijos. Interprete además el coeficiente de $D1$.

(b) A partir del modelo (4.22), construya la expresión análoga si ahora se define como modalidad de referencia a los casados.

Solución

(a) Si el número de hijos se incrementa en un 100%, el número de horas extras de los trabajadores aumentará en un 27%.

Los trabajadores solteros trabajan un $e^{-1.207} \cdot 100 = 29\,\%$ de las horas extras que trabajan los separados.

(b) El modelo cuya modalidad de referencia son los separados es el siguiente:

$$\ln \hat{Y}_i = 1.9 + 0.27 \ln(X_i) - 1.207 D1_i + 0.256 D2_i \qquad (4.22)$$

Por su parte, el modelo cuya modalidad de referencia son los casados es

$$\ln \hat{Y}_i = \hat{\alpha} + 0.27 \ln X_i + \hat{\gamma}_1 D1_i + \hat{\gamma}_2 D3_i \qquad (4.23)$$

donde $D3$ toma valor 1 para los separados y 0 para el resto.

Para expresar el mismo modelo definiendo como modalidad de referencia a los casados, calcularemos las estimaciones para cada modalidad tanto para el modelo (4.22) como para el nuevo modelo propuesto (4.23). Teniendo en cuenta que los valores esperados deben coincidir en ambos casos, despejando, obtendremos los coeficientes pedidos.

Las estimaciones obtenidas para cada modalidad en el modelo (4.22) son:

$$\ln \hat{Y}_i / solteros = 1.9 + 0.27 \ln X_i - 1.207$$

$$\ln \hat{Y}_i / casados = 1.9 + 0.27 \ln X_i + 0.256$$

$$\ln \hat{Y}_i / separados = 1.9 + 0.27 \ln X_i$$

Las estimaciones correspondientes a cada modalidad para el modelo (4.23) son:

$$\ln \hat{Y}_i / solteros = \hat{\alpha} + 0.27 \ln X_i + \hat{\gamma}_1$$

$$\ln \hat{Y}_i / casados = \hat{\alpha} + 0.27 \ln X_i$$

$$\ln \hat{Y}_i / separados = \hat{\alpha} + 0.27 \ln X_i + \hat{\gamma}_2$$

Igualando términos:

$$\left.\begin{array}{l} 1.9 - 1.207 = \hat{\alpha} + \hat{\gamma}_1 \Rightarrow \hat{\gamma}_1 = -1.463 \\ 1.9 + 0.256 = \hat{\alpha} \Rightarrow \hat{\alpha} = 2.156 \\ 1.9 = \hat{\alpha} + \hat{\gamma}_2 \Rightarrow \hat{\gamma}_2 = -0.256 \end{array}\right\}$$

El modelo final estimado sería

$$\ln \hat{Y}_i = 2.156 + 0.271 X_i - 1.463 D1_i - 0.256 D3_i$$

EJERCICIO 4.25

Se desea analizar la relación existente entre el nivel de empleo X y el nivel de ventas Y en un sector industrial dado. Para ello se han utilizado datos trimestrales, desde el primer trimestre de 1984 hasta el tercero de 1989. Si se sospecha la existencia de estacionalidad:

(a) ¿qué modelo especificaría como más apropiado tomando el primer trimestre como referencia?

(b) Contraste la existencia de estacionalidad trimestral si, estimando el modelo restringido, se ha obtenido:

$$\hat{Y}_t = \underset{(13.41)}{53.35} + \underset{(0.05)}{0.438} X_t \quad \hat{\sigma}_u^2 = 242.11$$

Entre paréntesis figuran las desviaciones típicas estimadas de los coeficientes. Se sabe además que la varianza estimada de la perturbación del modelo no restringido especificado en el apartado (a) es $\hat{\sigma}_u^2 = 227.2$.

(c) ¿Qué hipótesis nula plantearía para contrastar la existencia de diferencias entre el nivel medio de ventas del segundo trimestre con respecto al tercero? ¿Qué estadístico emplearía para el contraste?

Solución

(a) Para recoger los efectos mencionados se introducirán variables dicotómicas aditivas que recojan el comportamiento propio de cada trimestre según la siguiente especificación:

$$Y_t = \alpha + \beta X_t + \gamma_1 T_{2t} + \gamma_2 T_{3t} + \gamma_3 T_{4t} + u_t$$

donde T_2, T_3 y T_4 toman valor 1 si se refieren al segundo, tercer o cuarto trimestre, respectivamente, y cero en caso contrario.

(b) Para analizar la existencia de estacionalidad, el contraste de hipótesis a plantear será:

$$\left.\begin{array}{l} H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0 \\ H_1: \text{Al menos un } \gamma_i \neq 0 \end{array}\right\}$$

cuyo estadístico de contraste, que bajo la hipótesis nula se distribuye como $F_{q,N-k}$, es:

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)} = \frac{(242.11(23-2) - 227.2(23-5))/3}{227.2} = 1.5$$

El valor crítico $F_{3,18}$ al 95% de nivel de confianza es 3.16. Ello implica que no podemos rechazar la hipótesis nula, con lo cual la evidencia empírica favorece la hipótesis de que no existe un efecto estacional.

(c) Para contrastar la existencia de diferencias entre el nivel medio de ventas del segundo trimestre con respecto al tercero, las hipótesis serán:

$$\left.\begin{array}{l} H_0: \gamma_1 = \gamma_2 \\ H_1: \gamma_1 \neq \gamma_2 \end{array}\right\}$$

El estadístico de contraste vendría dado en este caso por

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)}$$

que, bajo la hipótesis nula, se distribuye como una $F_{q,N-k}$. También podría utilizarse, para realizar el contraste mencionado, la expresión general del contraste de restricciones lineales.

EJERCICIO 4.26

La función de consumo keynesiana estimada en el período 1964-83 es la siguiente:

$$\hat{C}_t = 1416.21 + 0.36643R_t \tag{4.24}$$

En el período 1984-1990 ésta es:

$$\hat{C}_t = -52321.29 + 1.6069R_t \tag{4.25}$$

Finalmente, la estimación para los años 1964-1990, es:

$$\hat{C}_t = 202.23 + 0.895R_t \quad R^2 = 0.9974 \tag{4.26}$$

Como información adicional se obtiene:

$$\sum_{t=1964}^{1990} (C_t - \bar{C})^2 = 6403270 \quad \sum_{t=1964}^{1983} e_t^2 = 2379.315 \quad \sum_{t=1984}^{1990} e_t^2 = 1295.861$$

(a) Explique el concepto de cambio estructural.

(b) Contraste si es admisible la hipótesis de constancia de los parámetros del modelo en los tres períodos considerados. Especifique claramente cuál es la hipótesis nula del contraste.

(c) En caso de que exista cambio estructural, ¿qué propondría para mejorar el modelo?

Solución

(a) Dentro de las hipótesis del modelo de regresión básico suponemos que existe constancia en los parámetros. Suponemos que la relación existente entre las variables se mantiene constante en toda la muestra. Cuando esto no es así, existe cambio estructural, y los coeficientes estimados son medias ponderadas de los verdaderos valores de los parámetros. En estos casos, se debe estimar por separado las submuestras que manifiestan diferente estructura, o bien, introducir ficticias que permitan incorporar distintas constantes y/o diferentes pendientes para las submuestras con distinta estructura.

(b) Dados los modelos $C_t = \alpha_1 + \beta_1 R_t + u_t$ (para 1964-73) y $C_t = \alpha_2 + \beta_2 R_t + u_t$ (para 1974-90) el contraste de hipótesis de cambio estructural es:

$$\left.\begin{array}{l} H_0: \alpha_1 = \alpha_2 \quad \text{y} \quad \beta_1 = \beta_2 \\ H_1: \alpha_1 \neq \alpha_2 \quad \text{y/o} \quad \beta_1 \neq \beta_2 \end{array}\right\}$$

El estadístico de contraste es:

$$F = \frac{\left(e'e - \left(e_1'e_1 + e_2'e_2\right)\right)/k}{\left(e_1'e_1 + e_2'e_2\right)/\left(N_1 + N_2 - 2k\right)}$$

que, bajo la hipótesis nula como cierta, se distribuye como una F_{k,N_1+N_2-2k}.

A partir del coeficiente de determinación, obtenemos la suma de cuadrados de los errores del modelo para toda la muestra como:

$$R^2 = 1 - \frac{e'e}{SCT} \Rightarrow e'e = \left(1 - R^2\right)SCT = (1 - 0.997)6\,403\,270 = 16\,648.502$$

El estadístico de contraste es:

$$F = \frac{(16648.502 - (2379.315 + 1295.861))/2}{(2379.315 + 1295.861)/(27 - 4)} = \frac{6486.663}{159.79} = 40.59$$

El valor crítico correspondiente para una $F_{2,27-4}$ para un nivel de significación del 5% es 3.42. En consecuencia, existe evidencia de cambio estructural.

(c) En caso de cambio estructural debemos estimar por separado ambas submuestras, o bien introducir variables dicotómicas aditivas y multiplicativas que permitan distintas constantes y pendientes para cada subperíodo.

EJERCICIO 4.27

A partir de una serie de datos trimestrales se han estimado las ventas de una empresa en función de la publicidad. Además, se ha introducido una variable ficticia aditiva y/o multiplicativa que toma valor 1 si la observación corresponde al 3º trimestre.

Cuadro 4.1

Dependent Variable: *Y*
Sample(adjusted): 1995:1 1998:4
Included observations: 16 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PUB	1.267069	0.217289	5.831277	0.0000
C	−3.462236	2.176278	−1.590898	0.1340
R-squared	0.708356	Mean dependent var		8.812500
Adjusted R-squared	0.687525	S.D. dependent var		3.953374
S.E. of regression	2.209916	Akaike info criterion		4.540255
Sum squared resid	68.372210	Schwarz criterion		4.636828
Log likelihood	−34.322040	F-statistic		34.003790
Durbin-Watson stat	2.062382	Prob(F-statistic)		0.000044

Cuadro 4.2

Dependent Variable: *Y*
Sample(adjusted): 1995:1 1998:4
Included observations: 16 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
*PUB*T3*	0.493771	0.320936	1.538532	0.1499
PUB	0.956229	0.147840	6.467979	0.0000
T3	−1.188215	3.497060	−0.339776	0.7399
C	−1.511785	1.416108	−1.067563	0.3067
R-squared	0.916931	Mean dependent var		8.812500
Adjusted R-squared	0.896164	S.D. dependent var		3.953374
S.E. of regression	1.273918	Akaike info criterion		3.534390
Sum squared resid	19.474410	Schwarz criterion		3.727537
Log likelihood	−24.275120	F-statistic		44.152930
Durbin-Watson stat	0.665893	Prob(F-statistic)		0.000001

Cuadro 4.3

Dependent Variable: *Y*
Sample(adjusted): 1995:1 1998:4
Included observations: 16 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PUB	1.061008	0.137947	7.691396	0.0000
C	−2.480990	1.333293	−1.860799	0.0855
T3	4.059903	0.810013	5.012147	0.0002
R-squared	0.900545			
Adjusted R-squared	0.885245			
Log likelihood	−25.715380	F-statistic		58.856500
Durbin-Watson stat	0.775750	Prob(F-statistic)		0.000000

(a) Contraste si el tercer trimestre afecta de forma significativa a las ventas (contraste de ruptura total).

(b) Realice los contrastes pertinentes para identificar si el tercer trimestre afecta a la constante y/o a la pendiente del modelo (contrastes de ruptura parcial), y determine cuál será el modelo final.

Solución

(a) Para realizar el contraste de ruptura total, es decir, para determinar si el tercer trimestre afecta o no de alguna forma significativa al modelo, partimos del siguiente modelo no restringido:

$$Y_t = \beta_1 + \beta_2 T3_t + \beta_3 Pub_t + \beta_4 Pub_i \cdot T3_i + u_i$$

y realizamos el siguiente contraste:

$$\left.\begin{array}{l} H_0: \beta_2 = \beta_4 = 0 \\ H_1: \beta_2 \quad \text{y/o} \quad \beta_4 \neq 0 \end{array}\right\}$$

El estadístico de contraste es:

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)} = \frac{(68.372 - 19.474)/2}{19.474/(16-4)} = \frac{24.449}{1.622} = 15.07$$

El valor crítico de una $F_{2,12}$ al 95% de nivel de confianza es 6.93 y, por tanto, rechazamos la H_0, lo que implica que existe cambio estructural.

(b) Sabiendo que existe cambio estructural, hay que determinar si éste afecta a las pendientes y/o a la constante del modelo:

Pasamos a contrastar la estabilidad de pendientes (suponiendo distinta constante)

$$\left.\begin{array}{l} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{array}\right\}$$

El valor del estadístico de contraste es $0.49377/0.32 = 1.53$. El valor crítico correspondiente a una t_{12} para un nivel de significación de 5% es 2.17, lo que supone que no podemos rechazar H_0, es decir, no existe diferencia significativa en la pendiente del modelo.

A continuación contrastamos la estabilidad de constante (suponiendo pendientes iguales) en el modelo $Y_t = \beta_1 + \beta_2 T3_t + \beta_3 Pub_t + u_t$

$$\left.\begin{array}{l} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{array}\right\}$$

El valor del estadístico de contraste en este caso es $4.059/0.81 = 5$. El valor crítico $t_{13}^{0.05}$ es 2.16 y, en consecuencia, rechazamos la H_0. Existen, por tanto, distintas ordenadas en el origen para el modelo según estemos o no en el tercer trimestre, pero la pendiente no varía.

El modelo adecuado es entonces

$$Y_t = \beta_1 + \beta_2 T3_t + \beta_3 Pub_t + u_t.$$

EJERCICIO 4.28

Se ha estimado el siguiente modelo:

$$\hat{Y}_i = 4.96 - 0.088X_i + 1.039 \cdot D1_i + 0.4736 D1_i \cdot X_i; \quad e'e = 17.17 \qquad (4.27)$$

donde:

Y es el número de horas de fútbol que un individuo ve en televisión a la semana.

X es el número de canales de televisión al que se tiene acceso.

$D1$ es una ficticia que toma valor 1 si es hombre y 0 si es mujer.

La muestra es de 10 individuos, siendo los 5 primeros hombres y los 5 restantes mujeres.

Sabiendo que:

$$Y'Y = 580 \quad X'X = \begin{pmatrix} 10 & 133 & 5 & 44 \\ 133 & 3931 & 44 & 728 \\ 5 & 44 & 5 & 44 \\ 44 & 728 & 44 & 728 \end{pmatrix} \quad X'Y = \begin{pmatrix} 64 \\ 705 \\ 47 \\ 545 \end{pmatrix}$$

(a) Estime los coeficientes del modelo restringido, no considerando diferencias entre sexos.

(b) Contraste si existe cambio estructural según el sexo.

Solución

(a) El modelo que no considera diferencias entre sexos (modelo restringido) es:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Las correspondientes matrices de datos necesarias para obtener los coeficientes estimados son:

$$X'X = \begin{pmatrix} 10 & 133 \\ 133 & 3931 \end{pmatrix} \quad (X'X)^{-1} = \begin{pmatrix} 0.1818 & -0.00610 \\ -0.0061 & 0.00046 \end{pmatrix} \quad X'Y = \begin{pmatrix} 64 \\ 705 \end{pmatrix}$$

Con esta información podemos estimar los coeficientes del modelo como:

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 7.290 \\ -0.067 \end{pmatrix}$$

(b) El contraste a realizar para determinar si existe cambio estructural según el sexo es:

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \beta_3 \neq 0 \quad \text{y/o} \quad \beta_4 \neq 0 \end{array}\right\}$$

El estadístico de contraste, que se distribuye bajo la H_0 como una $F_{2,10-4}$, es:

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)}$$

Calculando la *SCE* para el modelo restringido obtenemos:

$$SCE = Y'Y - \hat{\beta}'X'Y = 580 - (7.29 \quad -0.067)\begin{pmatrix} 64 \\ 705 \end{pmatrix} = 160.51$$

Sustituyendo, el estadístico de contraste es

$$F = \frac{(160.51 - 17.17)/2}{17.17/(10-4)} = 25.04$$

El valor crítico correspondiente a una $F_{2,6}$ al nivel de confianza del 95% es 5.14. Por tanto, rechazamos la hipótesis nula, es decir, existe cambio estructural.

EJERCICIO 4.29

Se pretende modelizar las ventas de una empresa y para ello se dispone de datos trimestrales desde el primer trimestre de 1980 hasta el cuarto de 1995, siendo, por tanto, el número de observaciones disponibles 64. Se estima un primer modelo:

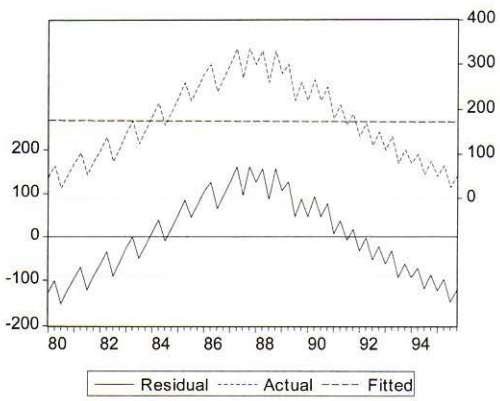
$$Vtas_t = \alpha + \beta t + u_t$$

donde t es la variable tiempo y toma valores $t = 1, 2, 3, ..., 64$.

Esta primera estimación resulta pésima y se pretende mejorar el modelo.

El Gráfico 4.7 corresponde a las ventas reales, ventas estimadas y errores de esta primera regresión:

Gráfico 4.7



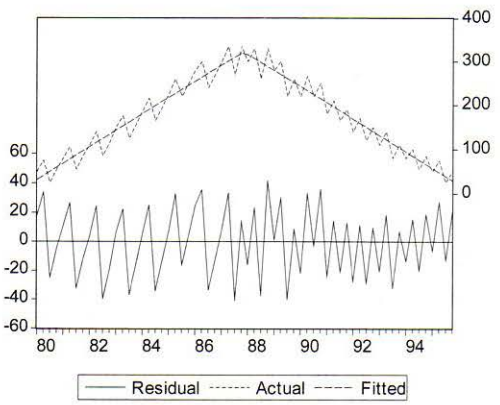
(a) ¿Cree que existe cambio estructural en el modelo? Realice el contraste correspondiente, sabiendo que los resultados de la *SCE* al estimar el modelo anterior para las submuestras 1980-1987 y 1988-1985 son:

$$SCE_{80-87} = 18612.45 \quad SCE_{88-95} = 17267.72$$

y que la *SCE* para la muestra total vale 505628.4.

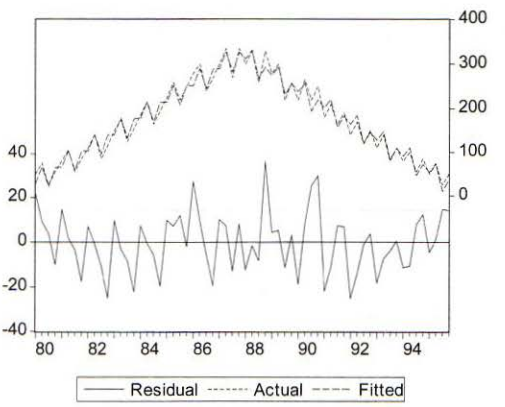
(b) ¿Qué variable/s cualitativa/s se ha/n añadido al modelo propuesto en el apartado anterior si el gráfico correspondiente al nuevo modelo es ahora el Gráfico 4.8? Escriba el nuevo modelo completo. ($SCE_{80-95} = 35880.17$)

Gráfico 4.8



(c) ¿Qué nuevos elementos se han añadido al modelo propuesto en (b) para obtener el Gráfico 4.9? Escriba el modelo completo. ($SCE_{80-95} = 11612.46$)

Gráfico 4.9



(d) Contraste que el componente estacional no es significativo en el modelo propuesto en el apartado (c).

Solución

(a) Existe claramente cambio estructural, ya que cambia la tendencia a partir de 1988, pasando de ser creciente a decreciente.

El contraste de cambio estructural se realiza a partir de los siguientes modelos:

$$Y_{1t} = \alpha_1 + \beta_1 t + u_{1t} \quad t = 1,...,32$$

$$Y_{2t} = \alpha_2 + \beta_2 t + u_{2t} \quad t = 33,...,64$$

El contraste de hipótesis de que no existe cambio estructural entre ambos subperiodos, formalmente se expresa como:

$$\left.\begin{array}{lll} H_0: \alpha_1 = \alpha_2 & \text{y} & \beta_1 = \beta_2 \\ H_1: \alpha_1 \neq \alpha_2 & \text{y/o} & \beta_1 \neq \beta_2 \end{array}\right\}$$

El estadístico de contraste es:

$$F = \frac{\left(e'e - \left(e_1'e_1 + e_2'e_2\right)\right)/k}{\left(e_1'e_1 + e_2'e_2\right)/\left(N_1 + N_2 - 2k\right)} =$$

$$= \frac{\left(505628.4 - \left(18612.45 + 17267.72\right)\right)/2}{\left(18612.45 + 17267.72\right)/\left(32 + 32 - 2\cdot 2\right)} = \frac{234874.115}{598} = 392.76$$

El valor crítico correspondiente a una $F_{2,60}$ al 95% de nivel de confianza es 3.15. Como el estadístico se encuentra en zona de rechazo de la hipótesis nula, rechazamos que los parámetros sean constantes en ambas submuestras, existiendo, por tanto, cambio estructural.

(b) En el modelo correspondiente al Gráfico 4.8, se ha incluido una variable cualitativa que toma valor 0 para una submuestra y valor 1 para la otra, permitiendo que las constantes y las tendencias varíen antes y después de 1988. El modelo estimado es:

$$Y_t = \alpha_1 + \alpha_2 D1_t + \beta_1 t + \beta_{2t} tD1_t + \varepsilon_t$$

donde $D1$ es la variable dicotómica incorporada.

(c) En este caso, en el modelo correspondiente al Gráfico 4.9, se han añadido además variables ficticias que recogen el efecto estacional, permitiendo diferente constante según que el dato corresponda a un trimestre u otro. Estas ficticias serán: $E1$ (que tomará valor 1 si la observación corresponde al primer trimestre), $E2$ (que tomará valor 1 si la observación corresponde al segundo trimestre) y $E3$ (que tomará valor 1 si corresponde al tercer trimestre). El cuarto trimestre queda definido como el de referencia. El modelo estimado es:

$$Y_t = \alpha_1 + \alpha_2 D1_t + \beta_1 t + \beta_{2t} tD1_t + \gamma_1 E1_t + \gamma_2 E2_t + \gamma_3 E3_t + \varepsilon_t$$

(d) Para contrastar que el componente estacional no es significativo, tenemos que realizar el siguiente contraste a partir de la ecuación anterior:

$$\left.\begin{array}{l} H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0 \\ H_1: \text{No } H_0 \end{array}\right\}$$

El estadístico de contraste, bajo la hipótesis nula, valdrá:

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)} = \frac{(35880.17 - 11612.46)/3}{11612.46/(64-7)} = \frac{8089.23}{203.72} = 39.7$$

El valor crítico es $F_{3,57}^{0.95} = 2.76$. Rechazamos, por tanto, la hipótesis nula de que no existe efecto estacional.

EJERCICIO 4.30

En un trabajo sobre el rendimiento académico se ha considerado el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (4.28)$$

donde Y es la puntuación media del expediente académico y X son las horas semanales dedicadas al estudio de un determinado alumno.

Para llevar adelante esta investigación se ha encuestado a 50 alumnos (20 hombre y 30 mujeres). Los resultados obtenidos para cada una de los sexos y la muestra en general son:

$$\begin{aligned}
&\text{Para los hombres} \quad \overline{X} = 9.02 \quad \overline{Y} = 4.60 \quad S_X^2 = 10.35 \quad S_Y^2 = 5.09 \\
&\text{Para las mujeres} \quad \overline{X} = 12.86 \quad \overline{Y} = 6.14 \quad S_X^2 = 9.54 \quad S_Y^2 = 3.28 \\
&\text{Para los 50 alumnos} \quad S_{X,Y} = 6.23
\end{aligned}$$

Ante la posibilidad de que exista cambio estructural en el modelo (4.28) debido al sexo del alumno, se ha estimado la siguiente regresión para los 50 individuos encuestados:

$$\hat{Y}_i = 0.791 - 1.465 D_i + 0.416 X_i + 0.169 \cdot D_i X_i \tag{4.29}$$

donde D es una variable dicotómica que toma el valor 1 si el individuo es hombre y cero si es mujer.

(a) Construya la matriz $X'X$ del modelo (4.29).

(b) Utilizando el resultado del apartado anterior contraste la existencia de cambio estructural en el modelo (4.28).

Solución

(a) La matriz $X'X$ se puede representar mediante la siguiente expresión:

$$X'X = \begin{pmatrix} N & \sum_{i=1}^{N} D_i & \sum_{i=1}^{N} X_i & \sum_{i=1}^{N} X_i D_i \\ & \sum_{i=1}^{N} D_i & \sum_{i=1}^{N} X_i D_i & \sum_{i=1}^{N} X_i D_i \\ & & \sum_{i=1}^{N} X_i^2 & \sum_{i=1}^{N} X_{2i}^2 D_i \\ & & & \sum_{i=1}^{N} X_i^2 D_i \end{pmatrix}$$

obteniéndose cada uno de los valores que la componen de la siguiente manera:

$$N = 50 \qquad \sum_{i=1}^{N} D_i = 20$$

$$\sum_{i=1}^{N} X_i = \bar{X}_H \cdot N_H + \bar{X}_M N_M = 9.02 \cdot 20 + 12.86 \cdot 30 = 566.2$$

$$\sum_{i=1}^{N} X_i D_i = \sum_{i=1}^{N} X_{Hi} = \bar{X}_H \cdot N_H = 9.02 \cdot 20 = 180.4$$

$$\sum_{i=1}^{N} X_i^2 = \left(S_{X_H}^2 + \bar{X}_H^2\right) N_H + \left(S_{X_M}^2 + \bar{X}_M^2\right) N_M$$

$$= \left(10.35 + 9.02^2\right) 20 + \left(9.54 + 12.86^2\right) 30 = 7081.796$$

$$\sum_{i=1}^{N} X_i^2 D_i = \sum_{i=1}^{N} X_{Hi}^2 = \left(S_{X_H}^2 + \bar{X}_H^2\right) N_H = \left(10.35 + 9.02\right) 20 = 1834.208$$

con lo que

$$\left(X'X\right) = \begin{pmatrix} 50 & 20 & 566.200 & 180.400 \\ & 20 & 180.400 & 180.400 \\ & & 7081.796 & 1834.208 \\ & & & 1834.208 \end{pmatrix}$$

(b) Para determinar la existencia de cambio estructural se utiliza el contraste de Chow, cuyo estadístico es

$$F = \frac{\left(SCE_R - SCE_{NR}\right)/q}{SCE_{NR}/\left(N-k\right)}$$

que, bajo la hipótesis nula de ausencia de cambio estructural como cierta, se distribuye como una $F_{q,N-k}$.

Para obtener la *SCE*, tanto del modelo restringido como del no restringido, utilizaremos la siguiente expresión:

$$SCE = Y'Y - \hat{\beta}'X'X\hat{\beta}$$

donde $Y'Y$ para ambos casos es igual a:

$$Y'Y = \left(S_{Y_H}^2 + \bar{Y}_H^2\right) N_H + \left(S_{Y_M}^2 + \bar{Y}_M^2\right) N_M =$$

$$= \left(5.09 + 4.60^2\right) 20 + \left(3.28 + 6.14^2\right) 30 = 1754.388$$

De esta forma, la *SCE* del modelo no restringido se obtendrá como sigue:

$$\hat{\beta}'(X'X) = (0.791 \quad -1.465 \quad 0.416 \quad 0.169)\begin{pmatrix} 50 & 20 & 566.200 & 180.400 \\ & 20 & 180.400 & 180.400 \\ & & 7081.796 & 1834.208 \\ & & & 1834.208 \end{pmatrix} =$$

$$= (276.2768 \quad 92.054 \quad 3439.5865 \quad 951.42208)$$

$$\hat{\beta}'(X'X)\beta = 1675.3341$$

$$SCE_{NR} = Y'Y - \hat{\beta}'X'X\hat{\beta} = 1754.388 - 1675.3341 = 79.0539$$

Mientras que para la *SCE* del modelo restringido habrá que calcular previamente los coeficientes del modelo.

$$S_X^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2 = \frac{7081.796}{50} - \left(\frac{566.2}{2}\right)^2 = 13.402944$$

$$\hat{\beta}_2 = \frac{S_{X,Y}}{S_X^2} = \frac{6.23}{13.402944} = 0.46482325$$

$$\hat{\beta}_1 = \overline{Y} - \overline{X}\hat{\beta}_2 = \frac{4.60 \cdot 20 + 6.14 \cdot 30}{50} = \frac{566.2}{50} \cdot 0.46482325 = 0.26034152$$

Por tanto, la *SCE* del modelo restringido quedará como sigue:

$$X'X = \begin{pmatrix} N & \sum_{i=1}^{N} X_i \\ & \sum_{i=1}^{N} X_i^2 \end{pmatrix} = \begin{pmatrix} 50 & 566.200 \\ & 7081.796 \end{pmatrix}$$

$$SCE_R = Y'Y - \hat{\beta}'X'X\hat{\beta} =$$

$$= 1754.388 - (0.260341512 \quad 0.46482325) \cdot$$

$$\cdot \begin{pmatrix} 50 & 566.200 \\ & 7081.796 \end{pmatrix} \begin{pmatrix} 0.260341512 \\ 0.464823250 \end{pmatrix} =$$

$$= 1754.388 - 1670.5212 = 83.8668$$

El valor del estadístico se obtiene tal que

$$F = \frac{(83.8668 - 79.0539)/2}{79.0539/(50-4)} = 1.4$$

mientras que el valor crítico para un nivel de significación del 5% es de 3.20. Por tanto, no se rechaza la hipótesis nula y no hay evidencia de cambio estructural.

EJERCICIO 4.31

A partir de una muestra de 36 familias se desea estimar la elasticidad precio e ingreso de la demanda de un producto determinado. Para ello se dispone de datos de cantidad y precio del producto así como del ingreso familiar. También se presume que podría existir una diferencia en el consumo de dicho producto según el grupo de edad al que pertenezca el consumidor (Joven o Adulto).

(a) Especifique el modelo econométrico adecuado que nos permita estimar dichas elasticidades recogiendo toda la información disponible.

(b) Si estimamos el modelo $Consumo = f(Precio, Ingreso)$ para cada submuestra se obtienen los resultados de la Tabla 4.2:

Tabla 4.2

	Coeficientes		$\hat{\sigma}_u^2$	N
	Precio	Ingreso		
Jóvenes	−1.238655 (0.303756)	1.220953 (0.112501)	0.09944355	24
Adultos	−1.051568 (0.578350)	1.057833 (0.210236)	0.14476080	12
Todos	−1.007670 (0.379428)	1.103880 (0.139650)	0.13801350	36

siendo la cifra entre paréntesis la desviación típica de cada estimador. Verifique la hipótesis de que las elasticidades precio e ingreso son las mismas para los dos grupos de edad.

Solución

(a) Para que los coeficientes tengan interpretación de elasticidades debemos especificar el modelo como *doble-log*. Por tanto, el modelo a especificar sería el siguiente:

$$\log(CONSUMO_i) = \beta_1 + \beta_2 \log(PRECIO_i) + \beta_3 \log(INGRESO_i) + \\ + \beta_4 JOVEN_i + u_i$$

En esta especificación hemos optado por dejar como referencia a los adultos, aunque también podríamos haber optado por dejar como referencia a los jóvenes.

(b) Para verificar la hipótesis de que las elasticidades son las mismas para todas las edades, debemos aplicar el test de Chow. El contraste de hipótesis y el estadístico de contraste son:

$$\left.\begin{array}{l} H_0: \beta_{jovenes} = \beta_{adultos} = \beta \\ H_1: \beta_{jovenes} \neq \beta_{adultos} \end{array}\right\} \Rightarrow F = \frac{\left(e'e - \left(e_1'e_1 + e_2'e_2\right)\right)/k}{\left(e_1'e_1 + e_2'e_2\right)/(N-2k)}$$

Las *SCE* de cada subgrupo son:

$$e'e = \hat{\sigma}_u^2 (N-k) = 0.1380135(36-3) = 4.5544455 \qquad \text{(Todos)}$$

$$e_1'e_1 = \hat{\sigma}_u^2 (N-k) = 0.09944355(24-3) = 2.08831455 \qquad \text{(Jóvenes)}$$

$$e_2'e_2 = \hat{\sigma}_u^2 (N-k) = 0.1447608(12-3) = 1.3028472 \qquad \text{(Adultos)}$$

Sustituyendo estos valores en el estadístico de contraste obtenemos:

$$F = \frac{(4.5544455 - (2.08831455 + 1.3028472))/3}{(2.08831455 + 1.3028472)/(36 - 2\cdot 3)} = 3.43$$

y, dado que el valor crítico es $F_{2,(36-2\cdot 3)} = 3.32$, como el estadístico cae en la región de rechazo, no podemos aceptar la hipótesis nula. Por tanto, las elasticidades precio e ingreso no van a ser iguales para los dos grupos de edad, puesto que nos encontramos ante un modelo con cambio estructural.

EJERCICIO 4.32

Se dispone del siguiente modelo temporal de periodicidad trimestral, donde se ha estimado el logaritmo de las ventas en función de las siguientes variables explicativas:

TIEMPO: variable tendencia

$T1$, $T2$, $T3$: variables dicotómicas estacionales, donde el 4º trimestre es el de referencia

$D1$: variable dicotómica que toma valor 1 para el subperíodo que va desde el primer trimestre de 1980 hasta el cuarto trimestre de 1987 y 0 para el resto

Cuadro 4.4

Dependent Variable: *LVENTAS*
Sample: 1980:1 1995:4
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIEMPO	–0.064063	0.003696	–17.331840	0.0000
T1	–0.050117	0.068457	–0.732092	0.4671
T2	0.144914	0.068206	2.124660	0.0380
T3	–0.294707	0.068054	–4.330481	0.0001
D1	–4.139808	0.194956	–21.234570	0.0000
D1TIEMPO	0.127960	0.005208	24.570000	0.0000
C	8.136376	0.189854	42.855940	0.0000
R-squared	0.919069			
Sum squared resid	2.108767			

Sabiendo que se dispone, además, de las estimaciones que figuran en el Cuadro 4.5 y el Cuadro 4.6:

Cuadro 4.5

Dependent Variable: *LVENTAS*
Sample: 1980:1 1995:4
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIEMPO	–0.064487	0.004763	–13.53836	0.0000
D1	–4.153381	0.251826	–16.49309	0.0000
D1TIEMPO	0.127960	0.006736	18.99551	0.0000
C	8.106970	0.235170	34.47287	0.0000
R-squared	0.857473			
Sum squared resid	3.713755			

Cuadro 4.6

Dependent Variable: *LVENTAS*
Sample: 1980:1 1995:4
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIEMPO	–0.000528	0.004363	–0.120966	0.9041
T1	–0.051451	0.227946	–0.225716	0.8222
T2	0.144025	0.227738	0.632415	0.5296
T3	–0.295152	0.227612	–1.296732	0.1998
C	5.057909	0.218870	23.109180	0.0000
R-squared	0.061882			
Sum squared resid	24.444050			

(a) Contraste la existencia de diferencias estacionales.

(b) Contraste la existencia de cambio estructural entre el período 1980.1-1987.4 y 1988.1-1995.4.

Solución

(a) Para contrastar la existencia de diferencias estacionales, comparamos la suma de cuadrados de los errores del modelo no restringido (Cuadro 4.5) con la del restringido (Cuadro 4.6).

Las hipótesis a contrastar son

$$\left.\begin{array}{l} H_0\text{: no existe diferencia significativa en los componentes estacionales} \\ H_1\text{: existe diferencia estacional} \end{array}\right\}$$

Formalmente, a partir del modelo no restringido

$$\ln V_t = \beta_1 + \beta_2 TIEMPO_T + \beta_3 T1_t + \beta_4 T2_t + \\ + \beta_5 T3_t + \beta_6 D1_t + \beta_7 D1 \cdot TIEMPO_t + u_t$$

el contraste a plantear es

$$\left.\begin{array}{l} H_0\text{: } \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1\text{: Al menos un } \beta_i \neq 0 \quad \forall\, i = 3,4,5 \end{array}\right\}$$

El estadístico de contraste es

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)} = \frac{(3.713 - 2.1087)/3}{2.108767/(64-7)} = 14.46$$

El valor crítico correspondiente a una $F_{3,57}$ para un nivel de significación del 5% es 2.76. Rechazamos, por tanto, la hipótesis nula, es decir, el componente estacional es significativo.

(b) Para contrastar la existencia de cambio estructural comparamos la suma de cuadrados de los errores del modelo restringido a una sola muestra (Cuadro 4.6) con el modelo no restringido (Cuadro 4.4).

Partiendo del modelo no restringido

$$\ln V_t = \beta_1 + \beta_2 TIEMPO_T + \beta_3 T1_t + \beta_4 T2_t + \\ + \beta_5 T3_t + \beta_6 D1_t + \beta_7 D1 \cdot TIEMPO_t + u_t$$

el contraste de hipótesis a plantear es

$$\left.\begin{array}{l} H_0: \beta_6 = \beta_7 = 0 \\ H_1: \beta_6 \neq 0 \quad \text{y/o} \quad \beta_7 \neq 0 \end{array}\right\}$$

El estadístico de contraste es

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-k)} = \frac{(24.444 - 2.1087)/2}{2.108767/(64-7)} = 301.86$$

El valor crítico correspondiente a una $F_{2,57}$ al 95% de nivel de confianza es 3.15. Claramente rechazamos la hipótesis nula, existiendo, por tanto, evidencia a favor del cambio estructural.

EJERCICIO 4.33

Se dispone de la siguiente información:

$$(X'X)^{-1} = \begin{pmatrix} 79464.42162 & 16532.97071 & -18.57014 \\ & 9988.07801 & -7.92112 \\ & & 0.00709 \end{pmatrix}$$

y los siguientes sumatorios para valores centrados:

$$\sum_{i=1}^{20} x_{2i} y_i = 0.00045 \quad \sum_{i=1}^{20} x_{3i} y_i = 0.00279$$

$$\sum_{i=1}^{20} \hat{y}_i^2 y_i = 4.28212 \quad \sum_{i=1}^{20} y_i^2 = 0.0051334$$

Realice un contraste sobre la forma funcional del modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i,$$

donde la matriz X está compuesta por tres columnas, cada una de las cuales representa, en ese orden, a las variables centradas de X_2, X_3 e $\hat{Y}^2$, siendo $\hat{Y}$ el valor estimado de Y.

Solución

El contraste RESET de Ramsey usa la siguiente regresión auxiliar:

$$Y_i = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 \hat{Y}_i^2 + v_i$$

con hipótesis nula: $H_0: \gamma_4 = 0$. La estimación del vector γ se realiza por MCO, con lo que: $\hat{\gamma} = (X'X)^{-1} X'Y$, donde la matriz X contiene las variables explicativas de la regresión auxiliar. La matriz $(X'X)^{-1}$ se encuentra en el enunciado del ejercicio, mientras que del mismo enunciado se deduce que:

$$X'Y = \begin{pmatrix} 0.00045 \\ 0.00279 \\ 4.28212 \end{pmatrix}$$

De esta manera:

$$\hat{\gamma} = (X'X)^{-1} X'Y = \begin{pmatrix} 2.360000 \\ 1.387000 \\ -0.000096 \end{pmatrix}$$

Además:

$$e'e = Y'Y - \hat{\gamma}' X'Y = 0.0051334 - 0.0045235 = 0.0006099,$$

con lo que:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{0.0006099}{20-4} = 0.0000381$$

Por tanto:

$$\hat{\sigma}_{\hat{\gamma}_4} = \sqrt{0.0000381 \cdot 0.00709} = 0.0005199,$$

donde el valor 0.00709 se ha obtenido de la diagonal principal de la matriz $(X'X)^{-1}$ (último valor).

El estadístico de contraste de significación individual es

$$t = \frac{-0.000096}{0.0005199} = -0.185159$$

que, bajo la hipótesis nula como cierta, sigue una distribución t de Student de 16 grados de libertad. Para un nivel de significación del 5%, el valor crítico de esta distribución es –2.1199, con lo que no se rechaza la hipótesis nula y, por tanto, no existen evidencia para rechazar la forma funcional del modelo original.

EJERCICIO 4.34

De la estimación del siguiente modelo centrado

$$\hat{y} = 0.84x \tag{4.30}$$

se sabe que

Tabla 4.3

x	y
2	3
1	1
–3	–4
1	3
4	1
–2	–1
–3	–3

A partir de esta información determine si es correcta la forma funcional del modelo (4.30). Responda mediante la realización de un contraste.

Solución

El contraste que se utilizará para determinar la validez de la forma funcional de la expresión (4.30) es el RESET de Ramsey. Este contraste utiliza la siguiente regresión auxiliar:

$$y_i = \gamma_1 + \gamma_2 x_i + \alpha \hat{y}_i^2 + v_i$$

sobre la que se contrasta la hipótesis nula H_0: $\alpha = 0$, mediante el estadístico de contraste

$$t = \frac{\hat{\alpha}}{S(\hat{\alpha})}$$

que, bajo la hipótesis nula como cierta, se distribuye como una t_{N-k}.

La estimación de la regresión auxiliar se obtiene teniendo en cuenta que:

$$(X'X) = \begin{pmatrix} N & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} \hat{y}_i^2 \\ & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \hat{y}_i^2 \\ & & \sum_{i=1}^{N} \hat{y}_i^4 \end{pmatrix} \quad X'Y = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i x_i \\ \sum_{i=1}^{N} y_i \hat{y}_i^2 \end{pmatrix}$$

El desarrollo que se ha utilizado para la obtención de cada uno de los valores que forman las anteriores matrices se presenta en la Tabla 4.4.

Tabla 4.4

$\hat{y}_i$	$\hat{y}_i^2$	$\hat{y}_i^4$	x_i^2	$x_i\hat{y}_i^2$	x_iy_i	$y_i\hat{y}_i^2$	y_i^2
1.68	2.8224	7.9659	4	5.6448	6	8.4672	9
0.84	0.7056	0.4979	1	0.7056	1	0.7056	1
−2.52	6.3504	40.3276	9	−19.0512	12	−25.4016	16
0.84	0.7056	0.4979	1	0.7056	3	2.1168	9
3.36	11.2896	127.4551	16	45.1584	4	11.2896	1
−1.68	2.8224	7.9659	4	−5.6448	2	−2.8224	1
−2.52	6.3504	40.3276	9	−19.0512	9	−19.0512	9
	31.0464	225.0379	44	8.4672	37	−24.6960	46

con lo que

$$(X'X) = \begin{pmatrix} 7 & 0 & 31.04640 \\ & 44 & 8.46720 \\ & & 225.03785 \end{pmatrix} \quad X'Y = \begin{pmatrix} 0.000 \\ 37.000 \\ -24.696 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 0.3723597 & 0.0099578 & -0.051746 \\ & 0.0231593 & -0.022450 \\ & & 0.011667 \end{pmatrix}$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 1.6463890 \\ 0.9123416 \\ -0.3712010 \end{pmatrix}$$

siendo la *SCE* igual a

$$e'e = Y'Y - \hat{\beta}'X'Y = 46 - (1.646389 \quad 0.9123416 \quad -0.371201) \begin{pmatrix} 0.000 \\ 37.000 \\ -24.696 \end{pmatrix} =$$

$$= 46 - 42.923809 = 3.0761913$$

Por tanto,

$$S(\hat{\alpha}) = 3.0761913 \cdot 0.011667 = 0.0089725$$

y, finalmente, se obtiene el siguiente valor del estadístico de contraste:

$$t = \frac{-0.371201}{\sqrt{0.0089725}} = -3.918785$$

Para un nivel de significación del 5%, para un contraste unilateral de la cola derecha, el valor crítico es de -2.776. Por tanto, se rechaza la hipótesis nula, lo que implica que la especificación funcional es incorrecta.

EJERCICIO 4.35

Se quiere analizar la relación existente entre las ventas de una empresa determinada (Y) y las variables gasto en publicidad (X_1) y precio del producto (X_2) a partir de una muestra disponible de datos trimestrales desde 1995 a 2005.

Si el modelo estimado resultante viene dado por: $\hat{Y}_t = 19.3 + 7.1X_{1t} - 0,9X_{2t}$ y se considera que el precio de la competencia (X_3) es también una variable a tener en cuenta en el modelo, ¿qué propiedades tendrían los estimadores del modelo anterior y cuáles tendrían los de un modelo que incluyera dicha variable?

Solución

Si el precio de la competencia es una variable relevante, como no se ha tenido en cuenta en el modelo original estaríamos incurriendo en un error de especificación, concretamente estaríamos omitiendo una variable relevante.

Las implicaciones que dicha omisión tiene son las siguientes:

i. si X_3 no es ortogonal con X_1 y X_2, entonces los estimadores MCO serán lineales, insesgados, tendrán menor varianza que los del modelo verdadero y serán inconsistentes.

ii. si X_3 es ortogonal con X_1 y X_2, entonces los estimadores MCO serán lineales, insesgados, tendrán igual varianza que los del modelo verdadero y serán consistentes.

iii. además, si existe omisión de variables relevantes, el estimador de la varianza de las perturbaciones será sesgado.

EJERCICIO 4.36

Suponga que el modelo verdadero es el siguiente:

$$Y_i = \beta_2 X_{2i} + \beta_3 X_{2i}^2 + u_i$$

y que, sin embargo, se estima el siguiente modelo:

$$Y_i = \gamma_2 X_{2i} + v_i$$

¿Será el estimador de γ_2 sesgado o insesgado? Demuéstrelo.

Solución

Al omitir una variable relevante en el modelo, el estimador de γ_2 será sesgado.

Demostración:

$$E(\hat{\gamma}_2) = E\left(\left(X_2'X_2\right)^{-1} X_2'Y\right) = E\left(\left(X_2'X_2\right)^{-1} X_2'\left(\begin{pmatrix} X_2 & X_2^2 \end{pmatrix}\begin{pmatrix}\beta_2 \\ \beta_3\end{pmatrix} + u\right)\right) =$$

$$= E\left(\frac{1}{\sum_{i=1}^{N} X_{2i}^2}\begin{pmatrix}\sum_{i=1}^{N} X_{2i}^2 & \sum_{i=1}^{N} X_{2i}^3\end{pmatrix}\begin{pmatrix}\beta_2 \\ \beta_3\end{pmatrix} + \frac{X_2'}{\sum_{i=1}^{N} X_{2i}^2} u\right) =$$

$$= E\left(\begin{pmatrix} 1 & \frac{\sum_{i=1}^{N} X_{2i}^3}{\sum_{i=1}^{N} X_{2i}^2}\end{pmatrix}\begin{pmatrix}\beta_2 \\ \beta_3\end{pmatrix}\right) = \beta_2 + \frac{\sum_{i=1}^{N} X_{2i}^3}{\sum_{i=1}^{N} X_{2i}^2}\beta_3$$

En general, el estimador de γ_2 será sesgado. Sólo será insesgado en el caso particular en que $\sum_{i=1}^{N} X_{2i}^3 = 0$.

EJERCICIO 4.37

Dado el modelo verdadero expresado en desviaciones respecto a la media

$$y_i = \beta_2 x_{2i} + \beta_3 \frac{1}{x_{3i}} + u_i$$

se estima el siguiente modelo:

$$y_i = \beta_2 x_{2i} + u_i \tag{4.31}$$

¿Serán los estimadores del modelo (4.31) sesgados o insesgados? Demuéstrelo.

Solución

En general, al omitir una variable relevante, los estimadores serán sesgados.

Demostración:

$$\hat{\beta}_2 = (x_2'x_2)^{-1} x_2' \left(\begin{pmatrix} x_2 & \frac{1}{x_2} \end{pmatrix} \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} + u \right) =$$

$$= (x_2'x_2)^{-1} x_2' \begin{pmatrix} x_2 & \frac{1}{x_2} \end{pmatrix} \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} + (x_2'x_2)^{-1} x'u =$$

$$= (x_2'x_2)^{-1} x_2'x_2\beta_2 + (x_2'x_2)^{-1} x_2' \frac{1}{x_2} \beta_3 + (x_2'x_2)^{-1} x'u$$

$$E(\hat{\beta}_2) = E\left((x_2'x_2)^{-1} x_2'x_2\beta_2 + (x_2'x_2)^{-1} x_2' \frac{1}{x_2} \beta_3 + (x_2'x_2)^{-1} x'u \right) =$$

$$= \beta_2 + (x_2'x_2)^{-1} x_2' \frac{1}{x_2} \beta_3 = \beta_2 + N(x_2'x_2)^{-1} \beta_3$$

Sólo si se cumple que $N(x_2'x_2)^{-1}\beta_3 = 0$, el estimador sería insesgado.

EJERCICIO 4.38

Dado el modelo sin constante:

$$Y_i = X_{2i}\beta_2 + u_i$$

se estima el siguiente modelo:

$$Y_i = X_{2i}\beta_2 + X_{3i}\beta_3 + u_i \tag{4.32}$$

(a) ¿Son sesgados o insesgados los coeficientes del modelo (4.32)? Justifique su respuesta.

(b) ¿Qué interpretación le daría a la siguiente expresión: $(X_2'X_2)^{-1} X_2'X_3$. Razone su respuesta.

Solución

(a) En este caso estamos estimando un modelo donde hemos considerado una variable irrelevante por lo que los estimadores serán insesgados. Lo probamos a continuación:

El estimador obtenido tendrá la siguiente expresión:

$$\begin{pmatrix}\hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}=\left(\begin{pmatrix}X_2\\ X_3\end{pmatrix}\begin{pmatrix}X_2 & X_3\end{pmatrix}\right)^{-1}\begin{pmatrix}X_2\\ X_3\end{pmatrix}Y$$

Para estudiar su insesgadez debemos calcular la esperanza del estimador:

$$E\begin{pmatrix}\hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}=E\left(\left(\begin{pmatrix}X_2\\ X_3\end{pmatrix}\begin{pmatrix}X_2 & X_3\end{pmatrix}\right)^{-1}\begin{pmatrix}X_2\\ X_3\end{pmatrix}(X_2\beta_2+U)\right)=$$

$$=\left(\begin{pmatrix}X_2\\ X_3\end{pmatrix}\begin{pmatrix}X_2 & X_3\end{pmatrix}\right)^{-1}\begin{pmatrix}X_2\\ X_3\end{pmatrix}X_2\beta_2=\begin{pmatrix}\beta_2\\ 0\end{pmatrix}$$

(b) La expresión a interpretar es el estimador obtenido a partir de una regresión donde X_3 es endógena y X_2 la variable explicativa.

5

Perturbaciones no esféricas. Heterocedasticidad

EJERCICIO 5.1

Sea el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \text{ con } Var(u_i) = \sigma^2 \cdot X_{2i}^2 .$$

Este modelo, además, presenta un problema de multicolinealidad exacta de tal manera que: $X_{2i} = \delta X_{3i}$

(a) Transforme el modelo dividiendo cada variable entre X_{3i}. Compruebe si el modelo transformado tiene perturbaciones homocedásticas.

(b) Transforme asimismo el modelo dividiéndolo entre X_{2i}. Compruebe si el modelo transformado tiene perturbaciones homocedásticas.

(c) ¿Qué propiedades presentan los estimadores de los modelos transformados en relación al problema de la heterocedasticidad?

Solución

(a) El modelo transformado queda como sigue:

$$\frac{Y_i}{X_{3i}} = \beta_1 \frac{1}{X_{3i}} + \beta_2 \frac{X_{2i}}{X_{3i}} + \beta_3 + \frac{u_i}{X_{3i}}$$

Además, dado que $X_{2i} = \delta X_{3i}$ tendremos que

$$\frac{Y_i}{X_{3i}} = \beta_1 \frac{1}{X_{3i}} + \beta_2 \frac{\delta X_{3i}}{X_{3i}} + \beta_3 + \frac{u_i}{X_{3i}} = \beta_1 \frac{1}{X_{3i}} + (\beta_2 \delta + \beta_3) + \frac{u_i}{X_{3i}}$$

y la varianza del término de perturbación aleatoria de este modelo vendrá dada por:

$$Var\left(\frac{u_i}{X_{3i}}\right) = E\left[\frac{u_i}{X_{3i}} - E\left(\frac{u_i}{X_{3i}}\right)\right]^2 = E\left[\frac{u_i}{X_{3i}}\right]^2 =$$

$$= \frac{E(u_i)^2}{X_{3i}^2} = \frac{\sigma_u^2 X_{2i}^2}{X_{3i}^2} = \frac{\sigma_u^2 \delta^2 X_{3i}^2}{X_{3i}^2} = \sigma_u^2 \delta^2$$

lo que implica que la perturbación aleatoria del nuevo modelo ponderado sea ahora homocedástica.

(b) Si ponderamos el modelo por $\frac{1}{X_{2i}}$ quedaría como sigue:

$$\frac{Y_i}{X_{2i}} = \beta_1 \frac{1}{X_{2i}} + \beta_2 + \beta_3 \frac{X_{3i}}{X_{2i}} + \frac{u_i}{X_{2i}}$$

Y, si sustituimos $X_{2i} = \delta X_{3i}$, o lo que es lo mismo $\frac{X_{3i}}{X_{2i}} = \delta$, nos queda

$$\frac{Y_i}{X_{2i}} = \beta_1 \frac{1}{X_{2i}} + (\beta_2 + \beta_3 \delta) + \frac{u_i}{X_{2i}}$$

La varianza del término de perturbación aleatoria de este modelo vendrá dada por

$$Var\left(\frac{u_i}{X_{2i}}\right) = E\left[\frac{u_i}{X_{2i}} - E\left(\frac{u_i}{X_{2i}}\right)\right]^2 = E\left(\frac{u_i}{X_{2i}}\right)^2 = \frac{E(u_i)^2}{X_{2i}^2} = \frac{\sigma_u^2 X_{2i}^2}{X_{2i}^2} = \sigma_u^2$$

lo que implica que esta ponderación también genera perturbaciones homocedásticas.

(c) Al ser ambos modelos homocedásticos, los estimadores MCO son eficientes.

EJERCICIO 5.2

Sea el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad Var(u_i) = \frac{\sigma^2}{Z_i} \quad i = 1,2,...,N$$

(a) ¿Qué hipótesis del Modelo de Regresión Lineal Múltiple viola este modelo?

(b) ¿Qué expresión tendría el modelo ponderado que corrige este problema?

(c) Demuestre que el modelo ponderado en (b) es homocedástico.

Solución

(a) El modelo presenta perturbaciones no esféricas al ser heterocedásticas.

(b) El modelo ponderado que corrige la falta de heterocedasticidad vendría dado por

$$\sqrt{Z_i}Y_i = \beta_1\sqrt{Z_i} + \beta_2 X_{2i}\sqrt{Z_i} + \beta_3 X_{3i}\sqrt{Z_i} + u_i\sqrt{Z_i}$$

(c) Para que sea homocedástico, la varianza de la perturbación nueva debe ser constante. La nueva perturbación aleatoria vendrá dada por $v_i = u_i\sqrt{Z_i}$ y su varianza es homocedástica como se puede comprobar a continuación:

$$Var(v_i) = Var\left(u_i\sqrt{Z}\right) = \left(\sqrt{Z_i}\right)^2 Var(u_i) = Z_i\frac{\sigma^2}{Z_i} = \sigma^2$$

EJERCICIO 5.3

Se dispone de la siguiente información $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$, donde

$$E(u_t) = 0 \quad E\left(u_t^2\right) = \frac{t}{X_{2i}^2} \quad E(u_t u_s) = 0 \quad \forall t \neq s \quad t = 1,2,3,4$$

(a) Escriba el modelo ponderado en el que las perturbaciones sean homocedásticas y calcule la esperanza matemática y la varianza de la perturbación de este modelo.

(b) Si los datos originales vienen dados en la Tabla 5.1:

Tabla 5.1

X_2	X_3
2	5
1	3
1	1
2	1

escriba la matriz de regresores del modelo transformado.

Solución

(a) El modelo ponderado se obtiene utilizando como ponderación la inversa de $\sqrt{\frac{t}{X_{2t}^2}} = \frac{\sqrt{t}}{X_{2t}}$, resultando así

$$\frac{Y_t}{\frac{\sqrt{t}}{X_{2t}}} = \frac{\beta_1}{\frac{\sqrt{t}}{X_{2t}}} + \beta_2 \frac{X_{2t}}{\frac{\sqrt{t}}{X_{2t}}} + \beta_3 \frac{X_{3t}}{\frac{\sqrt{t}}{X_{2t}}} + \frac{u_t}{\frac{\sqrt{t}}{X_{2t}}}$$

Este modelo lo podemos renombrar como

$$Y_t^* = \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + u_t^*$$

donde

$$Y_t^* = \frac{Y_t X_{2t}}{\sqrt{t}} \quad X_{1t}^* = \frac{X_{2t}}{\sqrt{t}} \quad X_{2t}^* = \frac{X_{2t}^2}{\sqrt{t}} \quad X_{3t}^* = \frac{X_{2t} X_{3t}}{\sqrt{t}} \quad u_t^* = \frac{u_t X_{2t}}{\sqrt{t}}$$

La esperanza de la perturbación aleatoria del modelo así transformado vendrá dada por

$$E\left(u_t^*\right) = E\left(\frac{u_t}{\sqrt{t}/X_{2t}}\right) = \frac{E\left(u_t\right)}{E\left(\sqrt{t}/X_{2t}\right)} = 0$$

y su varianza valdrá:

$$Var\left(u_t^*\right) = Var\left(\frac{u_t}{\sqrt{t}/X_{2t}}\right) = \frac{1}{t/X_{2t}^2} \cdot Var\left(u_t\right) = \frac{X_{2t}^2}{t} \cdot \frac{t}{X_{2t}^2} = 1$$

Por tanto, teniendo en cuenta la hipótesis de normalidad de las perturbaciones, la perturbación del modelo transformado se comporta como una Normal.

$$u_t^* \sim N(0,1)$$

(b) Partiendo de la matriz de regresores del modelo original

$$X = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 1 & 3 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

se obtiene la matriz del modelo transformado dividiendo cada elemento entre $\frac{\sqrt{t}}{X_2}$. De esta forma se obtiene

$$X^* = \begin{pmatrix} 2 & 4 & 10 \\ 0.707 & 0.707 & 2.120 \\ 0.577 & 0.577 & 0.577 \\ 1 & 2 & 1 \end{pmatrix}$$

EJERCICIO 5.4

Dada la muestra de la Tabla 5.2, donde se dispone del gasto y la renta de las familias en logaritmos:

Tabla 5.2

ln**Y**	lnX_2
2	1
3	2
5	2
8	5
12	5

(a) Estime el siguiente modelo eficientemente:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + u_i$$

teniendo en cuenta que $Var(u_i) = \sigma^2 \ln X_i$.

(b) Contraste que la elasticidad del gasto con respecto a la renta es de 0.8, sabiendo que un estimador insesgado de σ^2 es igual a $\hat{\sigma}^2 = 7.82$.

Solución

(a) Para que la estimación sea eficiente hay que estimar por Mínimos Cuadrados Generalizados (MCG), ya que las perturbaciones, como se muestra a continuación, son heterocedásticas:

$$V(U) = \sigma^2\Sigma = \sigma^2 \begin{pmatrix} 1 & & & & \\ 0 & 2 & & & \\ 0 & 0 & 2 & & \\ 0 & 0 & 0 & 5 & \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

Por tanto, la matriz Σ^{-1} será:

$$\Sigma^{-1} = \begin{pmatrix} 1 & & & & \\ 0 & 0.5 & & & \\ 0 & 0 & 0.5 & & \\ 0 & 0 & 0 & 0.2 & \\ 0 & 0 & 0 & 0 & 0.2 \end{pmatrix}$$

De esta forma, el vector de coeficientes MCG es:

$$\hat{\beta}^{MCG} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y = \begin{pmatrix} 2.4 & 5 \\ 5 & 15 \end{pmatrix}^{-1} \begin{pmatrix} 10 \\ 30 \end{pmatrix} =$$

$$= \begin{pmatrix} 1.3636 & -0.4545 \\ -0.4545 & 0.2181 \end{pmatrix} \begin{pmatrix} 10 \\ 30 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

(b) Para determinar si la elasticidad del gasto con respecto a la renta es de 0.8, el contraste a plantear es:

$$\left.\begin{array}{l} H_0\colon\ \beta_2 = 0.8 \\ H_1\colon\ \beta_2 \neq 0.8 \end{array}\right\}$$

El estadístico de contraste es $t = \dfrac{\hat{\beta}_2 - \beta_2}{S\left(\hat{\beta}_2\right)}$, que, bajo la hipótesis nula, se distribuye como una t-Student de $N-k$ grados de libertad.

La matriz estimada de varianzas y covarianzas de los coeficientes estimados se obtiene a partir de la siguiente expresión:

$$\hat{V}\left(\hat{\beta}^{MCG}\right)=\sigma^2\left(X'\Sigma^{-1}X\right)^{-1}$$

Sustituyendo los valores obtenemos

$$\hat{V}\left(\hat{\beta}^{MCG}\right)=\hat{\sigma}_u^2\left(X'\Sigma^{-1}X\right)^{-1}=7.82\begin{pmatrix}1.3636 & -0.4545\\ -0.4545 & 0.2181\end{pmatrix}$$

con lo que la varianza estimada de $\hat{\beta}_2$ es

$$S^2\left(\hat{\beta}_2\right)=7.82\cdot 0.2181=1.70$$

y el estadístico de contraste es

$$t=\frac{2-0.8}{\sqrt{1.7}}=1.91$$

El valor crítico correspondiente a una distribución t-Student de 3 grados de libertad, para un nivel de significación del 5%, es igual a 3.18. Por tanto, no podemos rechazar la hipótesis nula de que la elasticidad del gasto con respecto a la renta es de 0.8.

EJERCICIO 5.5

Se pretende estimar un modelo lineal que explique la relación entre las ventas (Y) y los gastos en publicidad (X) de una conocida marca de chocolate en los diferentes supermercados del país a través de la siguiente expresión:

$$Y=X\beta+U \quad \text{donde} \quad U\sim N\left(0,\Sigma\right) \tag{5.1}$$

(a) Si suponemos que $\Sigma=3\cdot I$, siendo I la matriz identidad de orden $\left(N\cdot N\right)$, obtenga la expresiones de la esperanza y la varianza de la estimación MCO de β.

(b) ¿Cree que nos encontramos ante un modelo homocedástico?

Solución

(a) La esperanza de los estimadores del modelo (5.1) es:

$$E(\hat{\beta}) = E\left[(X'X)^{-1}X'Y\right] = E\left[(X'X)^{-1}X'(X\beta+U)\right] =$$
$$= E\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'U\right] =$$
$$= (X'X)^{-1}X'XE(\beta) + (X'X)^{-1}E(U) = \beta$$

La matriz de varianzas y covarianzas de los estimadores del modelo (5.1) es:

$$V(\hat{\beta}) = E\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] = E\left[(X'X)^{-1}X'U\left((X'X)^{-1}X'U\right)'\right] =$$
$$= E\left[(X'X)^{-1}X'UU'X(X'X)^{-1}\right] = (X'X)^{-1}X'E[UU']X(X'X)^{-1} =$$
$$= (X'X)^{-1}X'V(U)X(X'X)^{-1} = (X'X)^{-1}X'\Sigma X(X'X)^{-1} =$$
$$= (X'X)^{-1}X'3IX(X'X)^{-1} = 3(X'X)^{-1}X'X(X'X)^{-1} = 3(X'X)^{-1}$$

(b) Sí, estamos ante un modelo homocedástico puesto que la varianza de las perturbaciones es constante.

EJERCICIO 5.6

Suponiendo que la matriz de varianzas y covarianzas de las perturbaciones del modelo (5.1) del Ejercicio 5.5 toma esta nueva expresión

$$V(U) = 3\Sigma = 3\begin{pmatrix} X_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_N^2 \end{pmatrix} \tag{5.2}$$

(a) Calcule nuevamente la esperanza y la varianza de la estimación MCO de β.

(b) ¿Sigue siendo el modelo homocedástico? En caso de no ser así, ¿qué expresión tendría el estimador MCG de β?

Solución

(a) La esperanza de los estimadores MCO del modelo con la matriz de varianzas y covarianzas (5.2) se mantiene igual a la del Ejercicio 5.5, puesto que lo único que ha cambiado es $V(U)$, y este término no afecta al cálculo de $E(\hat{\beta})$. Por tanto, seguimos ante estimadores insesgados.

La matriz de varianzas y covarianzas de los estimadores, en cambio sí varía. En este caso será:

$$V\left(\hat{\beta}\right)=E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right]=E\left[\left(X'X\right)^{-1}X'U\left(\left(X'X\right)^{-1}X'U\right)'\right]=$$
$$=E\left[\left(X'X\right)^{-1}X'UU'X\left(X'X\right)^{-1}\right]=\left(X'X\right)^{-1}X'E\left[UU'\right]X\left(X'X\right)^{-1}=$$
$$=\left(X'X\right)^{-1}X'V\left(U\right)X\left(X'X\right)^{-1}=3\left(X'X\right)^{-1}X'\Sigma X\left(X'X\right)^{-1}$$

(b) El modelo con esta matriz de varianzas y covarianzas de las perturbaciones es heterocedástico, ya que la varianza de la perturbación aleatoria para cada supermercado no es constante, sino que depende del valor que tome el gasto en publicidad de esa marca de chocolate en cada uno de ellos.

Debido a la presencia de heterocedasticidad, lo correcto sería estimar los parámetros del modelo por Mínimos Cuadrados Generalizados (MCG), mediante la siguiente expresión:

$$\hat{\beta}^{MCG}=\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}Y$$

EJERCICIO 5.7

Sea el modelo

$$Y=X\beta+U$$

donde Y y U son vectores de dimensión $\left(N\cdot 1\right)$, X es una matriz de dimensión $\left(N\cdot k\right)$ y donde β es un vector de dimensión $\left(k\cdot 1\right)$.

(a) Suponiendo que $U\sim N\left(0,\sigma_u^2 I\right)$, donde I es una matriz identidad, ¿qué dimensión tiene la matriz I?

(b) ¿Son homocedásticas las perturbaciones?

(c) ¿Qué interpretación tiene el parámetro σ_u^2?

(d) Demuestre que, en este modelo, el estimador $\hat{\sigma}_u^2=\dfrac{e'e}{N-k}$ es insesgado (siendo e el vector de los residuos de la estimación por MCO).

Solución

(a) La matriz I es de dimensión $(N \cdot N)$.

(b) Las perturbaciones son homocedásticas, puesto que la varianza de las mismas es constante $V(U) = \sigma_u^2 I$.

(c) El parámetro σ_u^2 es la varianza de la perturbación aleatoria de cada individuo. Recoge aquellas variaciones de la variable endógena no explicadas por los regresores.

(d) La demostración de la insesgadez del estimador $\hat{\sigma}_u^2$ es la siguiente:

$$E(\hat{\sigma}_u^2) = E\left(\frac{e'e}{N-k}\right) = \frac{1}{N-k} E(e'e) = \frac{1}{N-k} E\left[(MU)'(MU)\right] =$$

$$= \frac{1}{N-k} E(U'M'MU) = \frac{1}{N-k} E(U'MU) = \frac{1}{N-k} E\left[tr(U'MU)\right] =$$

$$= \frac{1}{N-k} E\left[tr(MUU')\right] = \frac{1}{N-k} tr\left[E(MUU')\right] =$$

$$= \frac{1}{N-k} tr\left[M \cdot E(UU')\right] = \frac{1}{N-k} tr(M) E(UU') =$$

$$= \frac{1}{N-k}(N-k)\sigma_u^2 = \sigma_u^2$$

siendo $M = \left(I - X(X'X)^{-1}X'\right)$

En esta demostración hay que tener en cuenta las siguientes cuestiones:

i. la matriz M es una matriz idempotente y por tanto $M'M = M$,

ii. el operador $tr(\cdot)$ se refiere a la traza de una matriz (recuérdese que la traza de una matriz es la suma de los elementos de la diagonal principal de la misma),

iii. además, la traza cumple la siguiente propiedad: $E(tr(A)) = tr(E(A))$ y

iv. se puede demostrar que la traza de M es $N-k$.

EJERCICIO 5.8

Sea el modelo

$$Y = X\beta + U \quad \text{donde} \quad U \sim N(0, \sigma^2 \Sigma) \quad \text{con} \quad \Sigma \neq I$$

(a) ¿Son homocedásticas las perturbaciones de este modelo?

(b) ¿Qué interpretación tiene el parámetro σ^2?

Solución

(a) El que las perturbaciones del modelo sean o no homocedásticas, dependerá de los elementos de la diagonal principal de la matriz Σ. Sólo sabemos que es una matriz diferente a la matriz identidad, pero puede tratarse de una matriz con los elementos de su diagonal principal constantes, en cuyo caso hablaríamos de un modelo homocedástico. En caso contrario, estaríamos ante un modelo heterocedástico.

(b) En este caso, el parámetro σ^2 hace referencia al factor común que tienen todos los elementos de la diagonal principal de la matriz de varianzas y covarianzas de las perturbaciones.

EJERCICIO 5.9

Sea el siguiente modelo:

$$C_i = \beta_1 + \beta_2 RTA_i + \beta_3 S_i + u_i$$

donde C_i es el consumo del individuo i, RTA_i es la renta bruta disponible del individuo i y S_i es una variable dicotómica que hace referencia al sexo del individuo i, tomando el valor 1 si es hombre y cero en caso contrario.

Se estima el modelo a través de dos especificaciones diferentes obteniendo los siguientes resultados, donde entre paréntesis figura la desviación típica de los coeficientes estimados:

$$\hat{C}_i = \underset{(2.6)}{25.18} + \underset{(0.005)}{1.61}\, RTA_i - \underset{(0.06)}{1.43}\, S_i, \quad R^2 = 0.9 \tag{5.3}$$

$$\frac{\hat{C}_i}{RTA_i} = \underset{(2.1)}{21.89}\frac{1}{RTA_i} + \underset{(0.005)}{1.61} - \underset{(0.055)}{1.42}\frac{S_i}{RTA_i}, \quad R^2 = 0.8 \tag{5.4}$$

(a) Si se sospecha que la dispersión de los errores puede estar relacionada con el nivel de renta, ¿qué modelo escogería? En el caso de que la sospecha se confirmara, ¿qué expresión utilizaría para estimar la matriz de varianzas y covarianzas de las estimaciones de los coeficientes del modelo (5.3)?

(b) ¿Qué supuesto sobre los errores habrán hecho los autores de la segunda estimación? En caso de que fuera cierto este supuesto, demuestre que se soluciona la existencia de heterocedasticidad.

Solución

(a) Si existe la sospecha de que la dispersión del consumo aumenta conforme se incrementa la renta, se escogería el modelo (5.4), siempre que al ponderar se solucionase el problema de heterocedasticidad. Si el problema de heterocedasticidad existe, la expresión que se utilizaría para estimar la matriz de varianzas y covarianzas de los coeficientes estimados en (5.3) sería:

$$V\left(\hat{\beta}\right)=\sigma^2\left(X'X\right)^{-1}\left(X'\Sigma X\right)\left(X'X\right)^{-1}$$

(b) Para obtener la estimación (5.4) se ha partido del supuesto de que la varianza de las perturbaciones del modelo están directamente relacionadas con el cuadrado del nivel de renta, de la siguiente forma:

$$Var\left(u_i\right)=\sigma^2 RTA_i^2$$

por ello se ha optado por ponderar el modelo dividiendo todos sus regresores entre la renta bruta disponible. De esta forma, la varianza de las perturbaciones del modelo transformado queda constante, por lo que el modelo transformado es homocedástico.

$$V\left(u_i^*\right)=V\left(\frac{u_i}{RTA_i}\right)=\frac{1}{RTA_i^2}V\left(u_i\right)=\frac{1}{RTA_i^2}\sigma^2 RTA_i^2=\sigma^2$$

EJERCICIO 5.10

Sea el modelo general

$$Y=X\beta+U \quad \text{con} \quad E\left(u_i\right)=0 \quad \text{y} \quad Var\left(u_i\right)=\sigma^2 X_{3i}$$

¿Cuánto valdrá la varianza de los estimadores si se estima el modelo por MCO? Demuéstrelo.

Solución

Al tratarse de un modelo que presenta heterocedasticidad, la varianza de los estimadores MCO será la siguiente:

$$V\left(\hat{\beta}^{MCO}\right)=\sigma^2\left(X'X\right)^{-1}X'\Sigma X\left(X'X\right)^{-1}$$

$$\text{siendo} \quad \Sigma=\begin{pmatrix} X_{3_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_{3_N} \end{pmatrix}$$

Demostración:

$$V\left(\hat{\beta}_{MCO}\right)=E\left[\left(\hat{\beta}^{MCO}-E\left(\hat{\beta}^{MCO}\right)\right)\left(\hat{\beta}^{MCO}-E\left(\hat{\beta}^{MCO}\right)\right)'\right]=$$

$$=E\left[\left(\hat{\beta}^{MCO}-\beta\right)\left(\hat{\beta}^{MCO}-\beta\right)'\right]=E\left[\left((X'X)^{-1}X'U\right)\left((X'X)^{-1}X'U\right)'\right]=$$

$$=E\left[(X'X)^{-1}X'UU'X(X'X)^{-1}\right]=(X'X)^{-1}X'E[UU']X(X'X)^{-1}=$$

$$=(X'X)^{-1}X'V(U)X(X'X)^{-1}=(X'X)^{-1}X'\sigma^2\Sigma X(X'X)^{-1}=$$

$$=\sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

EJERCICIO 5.11

Se desea estimar la relación existente entre el consumo de las familias y su renta. Dentro de las familias, la mitad son numerosas y la otra mitad no lo son. Además, se sabe que la varianza de las perturbaciones asociadas a las familias numerosas es el doble de las asociadas a las no numerosas.

(a) Determine si los estimadores de los parámetros serían insesgados, eficientes y consistentes si se estimara la relación consumo-renta por MCO.

(b) ¿Qué método de estimación debería utilizarse para eliminar o resolver estas dificultades? Plantee la expresión de la matriz de varianzas y covarianzas de la perturbación a utilizar en este caso.

Solución

(a) Si se estima por MCO las perturbaciones del modelo serían heterocedásticas y los estimadores MCO serían insesgados y consistentes pero no eficientes.

(b) La solución consistiría en estimar el modelo por MCG y, en ese caso, los estimadores se obtienen a partir de la expresión

$$\hat{\beta}^{MCG}=\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}Y \text{ con } V(U)=\sigma^2\Sigma,$$

siendo

$$\Sigma = \begin{pmatrix} a & & & & & \\ & a & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & 2a & \\ & & & & & 2a \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ \ddots \end{matrix}}\right\} \text{familias no numerosas} \\ \left.\vphantom{\begin{matrix} \ddots \\ 2a \\ 2a \end{matrix}}\right\} \text{familias numerosas} \end{matrix}$$

donde a representa un valor constante.

EJERCICIO 5.12

En la Tabla 5.3 se denota por Y a la variable endógena y por X_2 a la variable exógena. Además, se dispone de los datos de la variable Z, que se sabe que es la causante de la heterocedasticidad de este modelo de tal manera que $Var(u_i) = \sigma^2 Z_i$.

Tabla 5.3

Y	X_2	Z
110 240.4	1 002.2290	1.120722
109 995.2	999.6933	1.527862
110 159.2	1 000.7870	5.222250
110 051.4	999.9260	3.473066
110 147.9	1 001.0770	2.680817
110 202.4	1 001.3070	4.930057
110 106.3	1 000.1920	5.495203
110 246.7	1 001.0420	10.165500
110 273.1	1 000.7450	13.650850
110 055.2	998.4051	14.668230

(a) Estime los parámetros de posición del modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ por MCO y la varianza de las perturbaciones.

(b) Utilizando la expresión $\hat{V}(\hat{\beta}) = \hat{\sigma}_u^2 (X'X)^{-1}$, estime la varianza de los coeficientes estimados en el modelo del apartado anterior.

(c) Enuncie las características que presentan las estimaciones realizadas en los apartados (a) y (b).

Solución

(a) El estimador MCO de los parámetros de posición del modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ se calcula como $\hat{\beta}^{MCO} = (X'X)^{-1} X'Y$, en donde las matrices están definidas como se muestra a continuación:

$$X = \begin{pmatrix} 1 & 1\,002.2290 \\ 1 & 999.6933 \\ 1 & 1\,000.7870 \\ 1 & 999.9260 \\ 1 & 1\,001.0770 \\ 1 & 1\,001.3070 \\ 1 & 1\,000.1920 \\ 1 & 1\,001.0420 \\ 1 & 1\,000.7450 \\ 1 & 998.4051 \end{pmatrix} \quad Y = \begin{pmatrix} 110\,240.4 \\ 109\,995.2 \\ 110\,159.2 \\ 110\,051.4 \\ 110\,147.9 \\ 110\,202.4 \\ 110\,106.3 \\ 110\,246.7 \\ 110\,273.1 \\ 110\,055.2 \end{pmatrix}$$

A partir de estas matrices es inmediato comprobar los siguientes resultados:

$$(X'X) = \begin{pmatrix} 10.0000 & 10\,005.4034 \\ 10\,005.4034 & 10\,010\,819.6000 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 101\,558.299000 & -101.50335200 \\ -101.503352 & 0.10144854 \end{pmatrix}$$

Finalmente, los estimadores MCO de los coeficientes del modelo son los siguientes:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 41\,428.6679 \\ 68.6820 \end{pmatrix}$$

Para estimar la varianza de las perturbaciones a partir de la estimación MCO utilizamos la expresión $\hat{\sigma}_u^2 = \dfrac{e'e}{N-k}$, lo que implica suponer que las perturbaciones son homocedásticas y con autocorrelación nula. En nuestro ejemplo, $N = 10$, $k = 2$ y los errores los calculamos como

$$e_i = Y_i - (41428.6679 - 68.682 \cdot X_{2i})$$

Los valores de los errores, así como sus cuadrados, se muestran en la Tabla 5.4, junto con los valores reales de la variable y los predichos.

Tabla 5.4

Y	$\hat{Y}$	e	$(e)^2$
110 240.4	110 263.761	–23.36054400	545.71501600
109 995.2	110 089.604	–94.40359530	8 912.03881000
110 159.2	110 164.721	–5.52109928	30.48253720
110 051.4	110 105.586	–54.18589680	2 936.11142000
110 147.9	110 184.639	–36.73887940	1 349.74526000
110 202.4	110 200.436	1.96426046	3.85831916
110 106.3	110 123.855	–17.55530900	308.18887300
110 246.7	110 182.235	64.46499059	4 155.73501000
110 273.1	110 161.836	111.26354470	12 379.57640000
110 055.2	110 001.127	54.07255772	2 923.84150000
		$e'e$ =	33 545.29310000

Por tanto, el estimador MCO de la varianza de la perturbación aleatoria se obtiene de la siguiente manera:

$$\hat{\sigma}_u^2 = \frac{e'e}{N-k} = \frac{33\,545.2931}{10-2} = 4\,193.1616$$

(b) Dado que disponemos de toda la información necesaria, es inmediato obtener los resultados que se muestran a continuación:

$$\hat{V}\left(\hat{\beta}\right) = \hat{\sigma}_u^2 \left(X'X\right)^{-1} = \begin{pmatrix} 425\,850\,362 & -425\,619.963000 \\ -425\,619.963 & 425.390108 \end{pmatrix}$$

Por tanto, las dispersiones medidas a través de las varianzas son 425 850 362 y 425.390108 para el estimador de la constante y la pendiente, respectivamente.

(c) Si la perturbación aleatoria es heterocedástica, sabemos que las estimaciones realizadas en los apartados (a) y (b) presentan las siguientes características:

- Con respecto a los estimadores de los coeficientes, estos son:
  - Insesgados
  - No son eficientes
  - Generalmente consistentes
  - Además, se ha obtenido erróneamente la estimación de la matriz de varianzas y covarianzas de la estimación MCO de los coeficientes del modelo, pues, bajo la existencia de heterocedasticidad, la fórmula empleada es incorrecta.

- Con respecto al estimador de la varianza de las perturbaciones, éste es:
  - Sesgado
- Los contrastes de hipótesis dejan de ser válidos

EJERCICIO 5.13

Teniendo en cuenta la información que nos da el enunciado del Ejercicio 5.12,

(a) escriba la matriz de varianzas y covarianzas de la perturbación aleatoria,

(b) estime el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ por MCG, incluido el factor común de la varianza de las perturbaciones.

Solución

(a) La matriz de varianzas y covarianzas de la perturbación aleatoria es proporcional a la matriz Σ definida como

$$\Sigma = \begin{pmatrix} 1.1207 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.5279 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5.2222 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.4731 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.6808 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4.9300 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5.4952 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10.1655 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13.6509 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 14.6682 \end{pmatrix}$$

Por tanto, la matriz de varianzas y covarianzas será:

$$V(U) = \sigma^2 \Sigma$$

(b) El estimador MCG del modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ viene dado por la expresión $\hat{\beta}^{MCG} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y$. Con un poco de paciencia o una hoja electrónica es fácil obtener los siguientes resultados:

$$\left(X'\Sigma^{-1}X\right)=\begin{pmatrix} 3.0238 & 3\,026.5143 \\ 3\,026.5143 & 3\,029\,187.7200 \end{pmatrix}$$

$$\left(X'\Sigma^{-1}X\right)^{-1}=\begin{pmatrix} 297\,412.1800 & -297.14970 \\ -297.1497 & 0.29698 \end{pmatrix}$$

$$\left(X'\Sigma^{-1}Y\right)=\begin{pmatrix} 333\,043.381 \\ 333\,337\,480.000 \end{pmatrix}$$

$$\Rightarrow \hat{\beta}^{MCG}=\begin{pmatrix} 26\,278.6536000 \\ 83.7864082 \end{pmatrix}$$

El estimador de la varianza de las perturbaciones viene dado por la siguiente expresión:

$$Var(U)=\sigma^2\,\Sigma$$

donde los valores de Σ ya se definieron en el apartado (a), con lo que ahora será necesario, tan solo, estimar el factor común de la varianza de las perturbaciones $\left(\sigma^2\right)$ a través de la expresión:

$$\hat{\sigma}^2=\frac{\left(Y-X\hat{\beta}^{MCG}\right)'\Sigma^{-1}\left(Y-X\hat{\beta}^{MCG}\right)}{N-k}$$

Nuevamente los cálculos son tediosos, pero no suponen ninguna dificultad conceptual. Como ya hemos comentado, $N=10$, $k=2$, la matriz Σ ya ha sido definida en el apartado (a) y la expresión $\left(Y-X\hat{\beta}^{MCG}\right)$ coincide con la matriz columna formada por los N elementos calculados como

$$\left[Y_i-\left(26278.6536+83.7864082\cdot X_{2i}\right)\right]$$

Es decir, la matriz $\left(Y-X\hat{\beta}^{MCG}\right)$ tiene como resultado la matriz formada por la última columna de la Tabla 5.5:

Tabla 5.5

Y	$\hat{Y}^{MCG} = X\hat{\beta}^{MCG}$	$\left(Y - X\hat{\beta}^{MCG}\right)$
110 240.4	110 251.822	–11.42175980
109 995.2	110 039.365	–44.16456440
110 159.2	110 131.002	28.19824092
110 051.4	110 058.862	–7.46166158
110 147.9	110 155.300	–7.39981747
110 202.4	110 174.571	27.82930863
110 106.3	110 081.149	25.15115382
110 246.7	110 152.367	94.33270682
110 273.1	110 127.483	145.61727010
110 055.2	109 931.431	123.76908670

Realizando la operación

$$\hat{\sigma}^2 = \frac{\left(Y - X\hat{\beta}^{MCG}\right)' \Sigma^{-1} \left(Y - X\hat{\beta}^{MCG}\right)}{N - k},$$

se obtiene un valor igual a 665.8774.

EJERCICIO 5.14

Siguiendo con los datos del Ejercicio 5.12,

(a) estime la dispersión de los estimadores MCG y MCO,

(b) compare los resultados del apartado anterior con los obtenidos en el apartado (b) del Ejercicio 5.12.

Solución

(a) La expresión a utilizar para estimar la matriz de varianzas y covarianzas de los estimadores mínimo cuadrático generalizados es $\hat{V}\left(\hat{\beta}^{MCG}\right) = \hat{\sigma}^2 \left(X'\Sigma^{-1}X\right)^{-1}$ y, teniendo en cuenta los resultados de los apartados anteriores, se obtiene una matriz de varianzas y covarianzas estimadas para los estimadores MCG igual a:

$$\hat{V}\left(\hat{\beta}^{MCG}\right) = \begin{pmatrix} 198\,040\,150.000 & -197\,865.369000 \\ -197\,865.369 & 197.690963 \end{pmatrix}$$

La estimación de la matriz de varianzas y covarianzas de los estimadores MCO se obtiene mediante la expresión:

$$\hat{V}\left(\hat{\beta}^{MCO}\right) = \hat{\sigma}^2 \left(X'X\right)^{-1} \left(X'\Sigma X\right) \left(X'X\right)^{-1}$$

Teniendo en cuenta que

$$\left(X'\Sigma X\right) = \begin{pmatrix} 62.934557 & 62\,948.19328 \\ 62\,948.193280 & 62\,961\,906.18000 \end{pmatrix}$$

y, dado que el resto de los elementos que intervienen en la expresión de $\hat{V}\left(\hat{\beta}^{MCO}\right)$ ya han sido obtenidos en preguntas anteriores, especialmente teniendo en cuenta que, para la estimación MCO $\sigma^2 = 4193.1616$, se tendrá que

$$\hat{V}\left(\hat{\beta}^{MCO}\right) = \begin{pmatrix} 3\,469\,260\,762.000 & -3\,466\,518.046000 \\ -3\,466\,518.046 & 3\,463.779916 \end{pmatrix}$$

(b) En la Tabla 5.6 se muestran los resultados obtenidos por MCO y por MCG:

Tabla 5.6

MCO	$\hat{\beta}^{MCO} = \begin{pmatrix} 41\,428.6679 \\ 68.6820 \end{pmatrix}$	
	$\hat{V}\left(\hat{\beta}^{MCO}\right) = \begin{pmatrix} 3\,469\,260\,762.000 & -3\,466\,518.046000 \\ -3\,466\,518.046 & 3\,463.779916 \end{pmatrix}$	
MCG	$\hat{\beta}^{MCG} = \begin{pmatrix} 26\,278.6536000 \\ 83.7864082 \end{pmatrix}$	
	$\hat{V}\left(\hat{\beta}^{MCG}\right) = \begin{pmatrix} 198\,040\,150.000 & -197\,865.369000 \\ -197\,865.369 & 197.690963 \end{pmatrix}$	

Si nos centramos en los estimadores, la característica fundamental, en términos numéricos, es su gran diferencia al utilizar un método de estimación u otro. Para las propiedades teóricas entre ambos estimadores se remite al apartado (c) del Ejercicio 5.12 para los estimadores MCO y se recuerda que los estimadores MCG de los coeficientes son ELIO, mientras que la matriz de covarianzas de estos es insesgada.

Si nos centramos en los valores de la dispersión de los estimadores, la característica más relevante es la fuerte reducción que presenta la dispersión de los estimadores MCG frente a los estimadores MCO.

EJERCICIO 5.15

Se pretende estimar el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{5.5}$$

a partir de los datos muestrales de la Tabla 5.7.

Tabla 5.7

Y	X_2
3	1
10	3
4	4
5	2
11	5

Además, se sabe que los valores de la diagonal principal de la matriz de varianzas y covarianzas de las perturbaciones del modelo (5.5) es igual a $diag\{1,9,16,4,25\}$, mientras que los valores fuera de la diagonal principal son iguales a cero.

(a) Estime de forma eficiente el modelo (5.5).

(b) Estime la matriz de varianzas y covarianzas de las estimaciones obtenidas en la pregunta anterior.

(c) Contraste la hipótesis $H_0: \beta_2 = 0$.

Solución

(a) Debido a la existencia de heterocedasticidad en las perturbaciones del modelo, la estimación eficiente del modelo se realiza por el método de mínimos cuadrados generalizados de esta manera:

$$\hat{\beta}^{MCG} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y$$

donde:

$$\Sigma^{-1} = \begin{pmatrix} 1 & & & & \\ & 0.111 & & & \\ & & 0.0625 & & \\ & & & 0.25 & \\ & & & & 0.04 \end{pmatrix}$$

$$X'\Sigma^{-1} = \begin{pmatrix} 1 & 0.111 & 0.0625 & 0.25 & 0.04 \\ 1 & 0.333 & 0.2500 & 0.50 & 0.20 \end{pmatrix}$$

$$X'\Sigma^{-1}X = \begin{pmatrix} 1.4636111 & 2.283333 \\ & 5 \end{pmatrix}$$

$$\left(X'\Sigma^{-1}X\right)^{-1} = \begin{pmatrix} 2.375924 & -1.0850050 \\ & 0.6954857 \end{pmatrix}$$

con lo que:

$$\hat{\beta}^{MCG} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y = \begin{pmatrix} 1.3207497 \\ 1.8035243 \end{pmatrix}$$

(b) La expresión que se debe utilizar para la estimación de la matriz de varianzas y covarianzas de las estimaciones del modelo es:

$$\hat{V}\left(\hat{\beta}\right) = \hat{\sigma}^2\left(X'\Sigma^{-1}X\right)^{-1}$$

Para calcular el valor de esta expresión se debe obtener la suma de cuadrados de los errores del modelo, $e'e = Y'\Sigma^{-1}Y - \hat{\beta}'X'\Sigma^{-1}Y$, para lo que es necesario saber que:

$$Y'\Sigma^{-1} = \begin{pmatrix} 3 & 1.1111 & 0.25 & 1.25 & 0.44 \end{pmatrix} \quad Y'\Sigma^{-1}Y = 32.201111$$

$$\Rightarrow e'e = 32.201111 - 29.694412 = 2.506699$$

Una vez se conoce la suma de cuadrados de los errores, se puede obtener la estimación del factor común de la varianza de las perturbaciones:

$$\hat{\sigma}^2 = \frac{e'e}{N-k} = \frac{2.506699}{5-2} = 0.835566$$

y la matriz estimada de varianzas y covarianzas de las estimaciones tal y como se pedía en el apartado es igual a:

$$\hat{V}\left(\hat{\beta}\right)=\hat{\sigma}^2\left(X'\Sigma^{-1}X\right)=\begin{pmatrix}1.985242 & -0.906659\\ & 0.581124\end{pmatrix}$$

(c) La hipótesis nula de este contraste es $H_0:\beta_2=0$, cuyo estadístico de contraste es una ratio t tal que:

$$t=\frac{\hat{\beta}_2^{MCG}}{S\left(\hat{\beta}_2^{MCG}\right)}=\frac{1.8035243}{\sqrt{0.5811244}}=2.3658512$$

Como este valor es inferior a una t_3 para un nivel de significación de 5%, que toma el valor 3.1824, no se rechaza la hipótesis nula.

EJERCICIO 5.16

Sea el modelo

$$Y_i=\beta_1+\beta_2X_{2i}+u_i \tag{5.6}$$

del cual se dispone de la siguiente información muestral:

$$\sum_{i=1}^{75}Y_i=45.177041 \quad \sum_{i=1}^{75}X_{2i}=10.501578 \quad \sum_{i=1}^{75}X_{2i}^2=72.216278$$

$$\sum_{i=1}^{75}X_{2i}Y_i=41.175407 \quad \sum_{i=1}^{75}X_{2i}^2Y_i=50.534224 \quad \sum_{i=1}^{75}X_{2i}^3Y_i=92.644026$$

$$\sum_{i=1}^{75}X_{2i}^3=29.466035 \quad \sum_{i=1}^{75}X_{2i}^4=156.865226$$

(a) Estime el modelo (5.6) por MCO.

(b) Suponga que conoce que la varianza de la perturbación del modelo (5.6) se comporta según la siguiente expresión:

$$Var\left(u_i\right)=\frac{\sigma^2}{X_{2i}^2} \tag{5.7}$$

En este caso, ¿sería correcta la estimación de la matriz de varianzas y covarianzas de los coeficientes estimados en el apartado (a) a partir de la expresión habitual $\sigma_u^2\left(X'X\right)^{-1}$? ¿Por qué? ¿Qué expresión utilizaría para obtener la matriz de varianzas y covarianzas de la estimación de los coeficiente MCO de (5.6) teniendo en cuenta (5.7)?

(c) Estime eficientemente el modelo (5.6) teniendo en cuenta el comportamiento de la varianza de las perturbaciones aleatorias recogido en (5.7).

Solución

(a) La estimación del modelo por MCO se realiza a partir de los siguientes cálculos previos:

$$X'X = \begin{pmatrix} N & \sum_{i=1}^{N} X_{2i} \\ & \sum_{i=1}^{N} X_{2i}^2 \end{pmatrix} = \begin{pmatrix} 75 & 10.501578 \\ & 72.216278 \end{pmatrix}$$

$$\Rightarrow (X'X)^{-1} = \begin{pmatrix} 0.013610 & -0.001979 \\ & 0.014135 \end{pmatrix}$$

$$(X'Y) = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} Y_i X_{2i} \end{pmatrix} = \begin{pmatrix} 45.177041 \\ 41.175407 \end{pmatrix}$$

Por tanto,

$$\hat{\beta}^{MCO} = (X'X)^{-1} X'Y = \begin{pmatrix} 0.533386 \\ 0.492604 \end{pmatrix}$$

(b) La estimación de la matriz de varianzas y covarianzas de los coeficientes sería errónea si se utiliza la expresión $\sigma_u^2 (X'X)^{-1}$, dado que ésta sólo es correcta bajo la hipótesis de que las perturbaciones del modelo son esféricas. En este caso la expresión adecuada es:

$$\sigma^2 (X'X)^{-1} (X'\Sigma X)(X'X)^{-1}$$

(c) Teniendo en cuenta el comportamiento de la varianza de las perturbaciones, tal y como queda recogido en la ecuación (5.7), la estimación eficiente del modelo debe realizarse mediante el método de MCG o su equivalente, el método de Mínimos Cuadrados Ponderados (MCP). El modelo ponderado en este caso es

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + u_i^*$$

donde

$$Y_i^* = Y_i X_{2i} \quad X_{1i}^* = X_{2i} \quad X_{2i}^* = X_{2i}^2 \quad u_i^* = X_{2i} u_i$$

A partir de éste se pueden obtener las siguientes expresiones:

$$(X'X)=\begin{pmatrix}\sum_{i=1}^{N}X_{2i}^2 & \sum_{i=1}^{N}X_{2i}^3\\ & \sum_{i=1}^{N}X_{2i}^4\end{pmatrix}=\begin{pmatrix}72.216278 & 29.466035\\ & 156.865226\end{pmatrix}$$

$$\Rightarrow(X'X)^{-1}=\begin{pmatrix}0.014997 & -0.002812\\ & 0.006904\end{pmatrix}$$

$$X'Y=\begin{pmatrix}\sum_{i=1}^{N}Y_iX_{2i}^2\\ \sum_{i=1}^{N}Y_iX_{2i}^3\end{pmatrix}=\begin{pmatrix}50.534224\\ 92.644026\end{pmatrix}$$

que nos permiten estimar los coeficientes del modelo tal que:

$$\hat{\beta}^{MCP}=\left(X'^{*}X^{*}\right)^{-1}\left(X^{*\prime}Y^{*}\right)=\begin{pmatrix}0.496866\\ 0.497263\end{pmatrix}$$

EJERCICIO 5.17

Sea el siguiente modelo:

$$\hat{Y}_i=-21.77+0.00207X_{2i}+0.123X_{3i}+13.85X_{4i}\quad N=88$$

Si en la regresión del cuadrado de los residuos estandarizados en función de todas las explicativas de Y, el coeficiente de determinación es $R^2=0.1601$ y la suma cuadrática de la regresión es $SCR=28.18$,

(a) ¿qué problema presenta este modelo?

(b) Si a continuación se estima el modelo

$$\ln\hat{Y}_i=5.61+0.168\ln\left(X_{2i}\right)+0.7\ln\left(X_{3i}\right)+0.037X_{4i}$$

y en la regresión

$$e_i^2=\alpha_1+\alpha_2\ln\left(X_{2i}\right)+\alpha_3\ln\left(X_{3i}\right)+\alpha_4X_{4i}+u_i$$

el coeficiente de determinación es $R^2=0.0480$ y la suma cuadrática de la regresión es $SCR=8.44$. ¿Cuál cree que es la razón por la que se ha reestimado el modelo en logaritmos? Compruebe que se ha solucionado el problema que se pretendía con esta nueva estimación.

(c) ¿Por qué cree que en la regresión de $\ln(Y)$ se han introducido como explicativas otras variables en logaritmo?

Solución

(a) Las perturbaciones del modelo así estimado presentan un problema de heterocedasticidad a la vista del resultado del contraste de Breusch-Pagan:

$$\left.\begin{array}{l} H_0\text{: Perturbaciones homocedásticas} \\ H_1\text{: Perturbaciones heterocedásticas} \end{array}\right\}$$

El contraste de Breusch-Pagan parte de la estimación de una regresión auxiliar en la que la endógena la constituyen los residuos estandarizados al cuadrado y las explicativas son las mismas del modelo original $e_i^{*2} = f(X_{2i}, X_{3i}, X_{4i})$. De esta estimación se obtiene el estadístico de contraste de la siguiente manera: $SCR/2$, que, bajo la hipótesis nula, sigue una distribución χ^2_{p-1}, siendo p el número de regresores de la regresión auxiliar (incluida la constante). Concretamente, y para el caso que nos ocupa,

$$\frac{SCR}{2} = \frac{28.18}{2} = 14.09$$

Como el valor crítico para un nivel de significación del 5% de una función de distribución χ^2_3 es 7.81 rechazamos la hipótesis nula de homocedasticidad en las perturbaciones.

(b) Una ventaja de utilizar la forma funcional logarítmica en la variable dependiente es que a menudo se resuelven los problemas de heterocedasticidad.

Si estimamos $e_i^{*2} = f\left(\ln(X_{2i}), \ln(X_{3i}), X_{4i}\right)$ y la SCR toma el valor 8.44, el valor del estadístico de contraste de Breusch y Pagan será en esta ocasión $\frac{SCR}{2} = \frac{8.44}{2} = 4.22$, lo que nos lleva a no rechazar la hipótesis nula de perturbaciones homocedásticas.

(c) Se han introducido otras variables en logaritmo bien para interpretar los coeficientes en términos de elasticidades, o bien como consecuencia de una mala especificación funcional.

EJERCICIO 5.18

Dado el siguiente modelo, donde Y es el gasto de las familias y X la renta de las mismas:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

se sabe que $Var(u_i) = \sigma^2 \ln X_i$.

(a) Si se realiza una transformación dividiendo la ecuación entre $\ln(X_i)$ ¿cómo serán las perturbaciones del modelo transformado? Demuéstrelo.

(b) Una vez obtenido el modelo transformado propuesto según el apartado (a), se ha estimado éste a partir de una muestra. ¿Cuál/es de los siguientes gráficos puede/n corresponder al modelo estimado? Razone su respuesta.

Gráfico 5.1

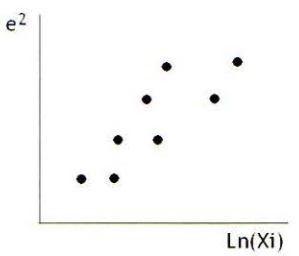


Gráfico 5.2

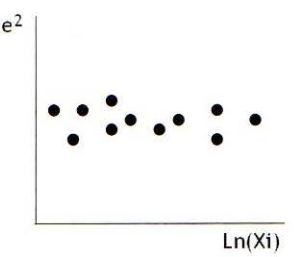
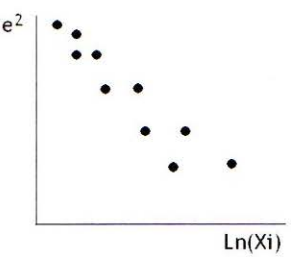


Gráfico 5.3



(c) ¿Qué transformación habría que realizar al modelo transformado en el apartado (a), para obtener perturbaciones esféricas?

Solución

(a) Dividiendo entre $\ln(X_i)$ el modelo transformado queda como sigue:

$$\frac{\ln Y_i}{\ln X_i} = \frac{\beta_1}{\ln X_i} + \frac{\beta_2 \ln X_i}{\ln X_i} + v_i \quad \text{donde} \quad v_i = \frac{u_i}{\ln X_i}$$

De esta forma:

$$Var(v_i) = \frac{1}{(\ln X_i)^2} Var(u_i) = \frac{\sigma^2 \ln X_i}{(\ln X_i)^2} = \frac{\sigma^2}{\ln X_i}$$

Por lo tanto, las perturbaciones del modelo transformado presentan heterocedasticidad.

(b) A medida que aumenta la renta, $\ln(X)$, se reduce la varianza de la perturbación del modelo incorrectamente transformado Por tanto, el gráfico que se corresponde con esta situación es el Gráfico 5.3.

(c) La transformación apropiada para obtener perturbaciones esféricas sería:

$$\frac{\ln Y_i}{\ln X_i}\sqrt{\ln X_i} = \frac{\beta_1}{\ln X_i}\sqrt{\ln X_i} + \frac{\beta_2 \ln X_i}{\ln X_i}\sqrt{\ln X_i} + v_i \quad \text{donde} \quad v_i = \frac{u_i}{\ln X_i}\sqrt{\ln X_i}$$

ya que entonces la varianza de las perturbaciones del nuevo modelo sería:

$$Var(v_i) = \sigma^2$$

EJERCICIO 5.19

Se ha estimado la siguiente regresión:

$$\hat{Y}_i = 0.55 + 0.77X_{2i} - 0.266X_{3i} \quad N = 500 \tag{5.8}$$

cuyo coeficiente de determinación es: $R^2 = 0.784$.

A partir de ella se ha obtenido la siguiente regresión auxiliar:

$$\hat{e}_{1i}^2 = \underset{(0.12)}{0.086} + \underset{(0.019)}{0.0179}\,X_{2i} - \underset{(-0.12)}{0.041}\,X_{2i}^2 - \underset{(-0.19)}{0.015}\,X_{3i} + \underset{(4.04)}{0.043}\,X_{3i}^2 \tag{5.9}$$

donde e_1 son los residuos de la regresión (5.8), figurando entre paréntesis el valor de los estadísticos *t*-Student para el contraste de significación individual.

Se sabe, además que:

$$\sum_{i=1}^{N} e_{1i}^4 = 692.81 \qquad \sum_{i=1}^{N} e_{1i}^2 = 268.43 \qquad \sum_{i=1}^{N} e_{1i}^2 X_{2i} = 356.28$$

$$\sum_{i=1}^{N} e_{1i}^2 X_{3i} = 1287.6 \qquad \sum_{i=1}^{N} e_{1i}^2 X_{3i}^2 = 6892.89 \qquad \sum_{i=1}^{N} e_{1i}^2 X_{2i}^2 = 493.76$$

(a) Realice el contraste de White argumentando las conclusiones obtenidas.

(b) A partir de la información suministrada, proponga un modelo alternativo cuyas perturbaciones sean esféricas.

Solución

(a) El contraste de White tiene como hipótesis nula la homocedasticidad. Para llevar a cabo este contraste se utiliza una regresión auxiliar como la planteada

en la ecuación (5.9), mientras que su estadístico de contraste es igual a NR^2, siendo R^2 el coeficiente de determinación de la regresión auxiliar. Bajo la hipótesis nula este estadístico se distribuye como una χ^2_{p-1}, siendo p el número de regresores de la regresión auxiliar (incluida la constante). El coeficiente de determinación de la regresión auxiliar se obtiene tal que

$$R^2 = 1 - \frac{SCE}{SCT}$$

Para obtener SCE, tenemos en cuenta que, en general, y para cualquier regresión MCO:

$$SCE = Y'Y - \hat{\beta}' X'Y$$

En nuestro caso concreto, para la regresión auxiliar del contraste de White

$$SCE = e^{2\prime} e^2 - \hat{\alpha}' X' e^2$$

siendo $\hat{\alpha}'$ un vector de dimensión $(1 \cdot p)$ que recoge las estimaciones de los coeficientes de la regresión auxiliar y X' la matriz de dimensión $(p \cdot N)$ que recoge en cada una de sus filas los valores de los regresores de la regresión auxiliar. De esta manera, tendremos que:

$$X'e^2 = \begin{pmatrix} \sum_{i=1}^{N} e_i^2 \\ \sum_{i=1}^{N} e_i^2 X_{2i} \\ \sum_{i=1}^{N} e_i^2 X_{2i}^2 \\ \sum_{i=1}^{N} e_i^2 X_{3i} \\ \sum_{i=1}^{N} e_i^2 X_{3i}^2 \end{pmatrix}$$

$$e^{2'}e^2 = \sum_{i=1}^{N} e_i^4 - \begin{pmatrix} 0.086 & 0.0179 & -0.041 & -0.015 & 0.043 \end{pmatrix} \begin{pmatrix} 268.43 \\ 356.28 \\ 493.76 \\ 1287.60 \\ 6892.89 \end{pmatrix} =$$

$$= 692.81 - 286.29 = 406.51$$

Igualmente, y para el caso de la regresión auxiliar, tendremos que:

$$SCT = e^{2'}e^2 - N\left(\overline{e^2}\right)^2 = e^{2'}e^2 - 500\left(\frac{\sum_{i=1}^{N} e_i^2}{N}\right)^2 =$$

$$= 692.81 - 500\left(\frac{268.43}{500}\right)^2 = 548.7$$

de donde

$$R^2 = 1 - \frac{406.51}{548.7} = 0.259$$

Siendo el estadístico de contraste

$$NR^2 = 500 \cdot 0.259 = 129.56$$

El valor crítico de una χ_4^2 al 95% de nivel de confianza es 9.48, que nos lleva a rechazar la hipótesis nula de homocedasticidad en las perturbaciones.

(b) Teniendo en cuenta la ecuación (5.9) y el grado de significatividad de X_3^2, la siguiente forma funcional de la varianza de las perturbaciones parece adecuada:

$$Var(u_i) = \sigma^2 X_{3i}^2$$

con lo que el modelo ponderado quedaría como:

$$\frac{Y_i}{X_{3i}} = \frac{\beta_1}{X_{3i}} + \beta_2 \frac{X_{2i}}{X_{3i}} + \beta_3 + \frac{u_i}{X_{3i}}$$

EJERCICIO 5.20

Se estima un modelo uniecuacional para explicar la emisión de kilogramos de dióxido de carbono a la atmósfera ($CO2$) a nivel mundial. Se espera que dicha emisión dependa del consumo per cápita de KW/h de electricidad ($ELECTRIC$) y Kg de carbón ($CARBÓN$), del consumo anual de barriles de petróleo ($PETROL$), del porcentaje de energía generada por quema de combustibles fósiles ($FÓSIL$), del nivel porcentual en el grado de concentración urbana ($URBAN$), de si el país está en el Norte o en el Sur (SUR) y de si ha ratificado o no el tratado de Kyoto (SI_KYOTO). Los términos que figuran en la ecuación entre paréntesis son la desviación típica estimada del estimador correspondiente.

$$\begin{aligned} \hat{CO2}_i = &-\underset{(104.28)}{169.42} + \underset{(0.012)}{0.03}\,ELECTRIC_i + \underset{(0.028)}{0.29}\,CARBON_i + \underset{(5.77)}{126.56}\,PETROL_i + \\ &+ \underset{(0.95)}{2.36}\,FOSIL_i + \underset{(1.85)}{3.30}\,URBAN_i + \underset{(94.73)}{190.9}\,SUR_i - \underset{(62.75)}{107.24}\,SI\_KYOTO_i \end{aligned} \tag{5.10}$$

Si se quisiera contrastar la presencia de heterocedasticidad en el modelo (5.10), ¿qué expresión tendría la regresión auxiliar del contraste de White sin términos cruzados?

Solución

El contraste de White sin términos cruzados utiliza como regresión auxiliar la estimación de los errores al cuadrado de la estimación (5.10) en función de cada una de las variables explicativas con forma lineal y cuadrática, por lo que su expresión sería la siguiente:

$$\begin{aligned} error_i^2 = &\beta_0 + \beta_1 ELECTRIC_i + \beta_2 ELECTRIC_i^2 + \beta_3 PETROL_i + \beta_4 PETROL_i^2 + \\ &+ \beta_5 CARBON_i + \beta_6 CARBON_i^2 + \beta_7 FOSIL_i + \beta_8 FOSIL_i^2 + \beta_9 URBAN_i + \\ &+ \beta_{10} URBAN_i^2 + \beta_{11} SUR_i + + \beta_{12} SI\_KYOTO_i + v_i \end{aligned}$$

EJERCICIO 5.21

A continuación, en el Cuadro 5.1, se muestra parte de la salida de regresión del modelo auxiliar del contraste de White sin términos cruzados del modelo (5.10). ¿Podemos suponer que el modelo presenta heterocedasticidad? Responda en términos econométricos precisos.

Cuadro 5.1

Test Equation:
(...)
Included observations: 161

Variable	Coefficient	Std. Error	t-Statistic	Prob.
		(...)		
R-squared	0.318526	Mean dependent var		0.132411
Adjusted R-squared	0.248029	S.D. dependent var		0.307104
S.E. of regression	0.266309	Akaike info criterion		0.285768
Sum squared resid	10.283470	Schwarz criterion		0.591994
Log likelihood	–7.004299	F-statistic		4.518280
Durbin-Watson stat	2.049889	Prob(F-statistic)		0.000001

Solución

El estadístico de contraste es NR^2. Sustituyendo los valores del Cuadro 5.1, nos da 51.28. Este estadístico, bajo la hipótesis nula, se distribuye como una $\chi^2_{p-1} = \chi^2_{15}$. El valor crítico para una chi-cuadrado con 15 grados de libertad, para un nivel de significación del 5%, es 25, lo que nos lleva a rechazar la hipótesis nula de homocedasticidad y concluir que el modelo es heterocedástico.

EJERCICIO 5.22

Se quiere estimar una regresión para una muestra representativa de 100 ciudades españolas, en la que los gastos en educación (G) vengan explicados por los ingresos (I) de la ciudad, el número de niños en edad escolar (N) y las subvenciones recibidas con propósitos educativos (S).

$$G_i = \beta_0 + \beta_1 I_i + \beta_2 N_i + \beta_3 S_i + u_i \qquad i = 1,..,100 \qquad (5.11)$$

(a) ¿Esperaría que la heterocedasticidad fuera un problema en este caso? Argumente su respuesta.

(b) ¿Podríamos tratar de detectar el problema a través de la prueba de Goldfeld-Quandt? De ser así, detalle cómo se haría el contraste para este modelo concreto.

(c) Suponga que finalmente se estima el siguiente modelo:

$$G_i^* = \beta_0 \frac{1}{I_i} + \beta_1 + \beta_2 N_i^* + \beta_3 S_i^* + u_i^* \qquad i = 1,..,100 \qquad (5.12)$$

¿Qué supuesto se ha hecho acerca del comportamiento de la perturbación aleatoria del modelo (5.11)? Bajo la premisa de que este supuesto es verdadero demuestre que, con esta transformación, el posible problema de heterocedasticidad queda solventado.

(d) Bajo el supuesto de que existe heterocedasticidad en el modelo (5.11), ¿cómo serían los estimadores MCO del mismo? ¿Y los del modelo (5.12)?

Solución

(a) Una de las causas estructurales de heterocedasticidad surge en los modelos de corte transversal con unidades muestrales de diferente tamaño. En este caso estamos planteando la estimación de un modelo de gasto educativo a partir de la información de 100 ciudades españolas que, en términos de gasto educativo, son de distinto tamaño. Es de esperar que las perturbaciones aleatorias asociadas a las ciudades con ingresos elevados tengan varianzas mayores que las asociadas a ciudades con ingresos bajos, por lo que sí es factible la presencia de heterocedasticidad en el modelo planteado.

(b) El contraste de Goldfeld y Quandt se utiliza para detectar heterocedasticidad en modelos en los que se sospecha que dicha heterocedasticidad ha sido causada por una única variable explicativa. Si éste fuera el caso, es muy probable que la heterocedasticidad sea producida por la variable I (ingresos).

Los pasos a seguir en la realización de dicho contraste son los siguientes:

i. Ordenar todas las observaciones por valores crecientes de la variable ingresos.

ii. Eliminar el tercio central de las observaciones.

iii. Estimar el modelo original (5.11) para las dos submuestras restantes.

iv. Contrastar que las varianzas de las perturbaciones de ambas estimaciones son iguales a través del siguiente contraste:

$$\left.\begin{array}{l} H_0: \sigma_{u_1}^2 = \sigma_{u_2}^2 \\ H_1: \sigma_{u_1}^2 \neq \sigma_{u_2}^2 \end{array}\right\}$$

cuyo estadístico es:

$$F = \frac{SCE_2}{SCE_1}$$

que, bajo la hipótesis nula como cierta, se distribuye como una F_{N_2-k,N_1-k}, donde $N_1 = N_2$.

(c) La transformación aplicada al modelo (5.11) para llegar al modelo (5.12) ha consistido en dividir todas las variables entre la variable ingreso

$$\frac{G_i}{I_i} = \frac{\beta_0}{I_i} + \frac{\beta_1 I_i}{I_i} + \frac{\beta_2 N_i}{I_i} + \frac{\beta_3 S_i}{I_i} + \frac{u_i}{I_i} = \beta_0 \frac{1}{I_i} + \beta_1 + \beta_2 \frac{N_i}{I_i} + \beta_3 \frac{S_i}{I_i} + \frac{u_i}{I_i}$$

$$G_i^* = \beta_0 \frac{1}{I_i} + \beta_1 + \beta_2 N_i^* + \beta_3 S_i^* + u_i^*$$

Por tanto, el supuesto del que se parte es

$$\sigma_{u_i}^2 = \sigma_u^2 \cdot I_i^2$$

A continuación demostraremos cómo mediante esta transformación desaparece el problema de heterocedasticidad.

$$E\left(u_i^*\right) = E\left(\frac{u_i}{I_i}\right) = \frac{1}{I_i} E\left(u_i\right) = 0$$

$$\sigma_{u_i^*}^2 = Var\left(u_i^*\right) = Var\left(\frac{u_i}{I_i}\right) = \frac{1}{I_i^2} Var\left(u_i\right) = \frac{1}{I_i^2} \sigma_u^2 I_i^2 = \sigma_u^2$$

(d) Los estimadores del modelo (5.11), al ser heterocedástico y estar estimado por MCO, serán lineales, insesgados y consistentes, pero no eficientes.

Por su parte, los estimadores MCO del modelo (5.12), al ser un modelo homocedástico, serán lineales, insesgados, consistentes y eficientes y, por tanto, óptimos (ELIO).

EJERCICIO 5.23

Una vez estimado el siguiente modelo:

$$C_i = \beta_1 + \beta_2 R_i + u_i \quad i = 1, \ldots, 500 \tag{5.13}$$

donde C_i y R_i son el consumo y la renta anual de la familia i-ésima, respectivamente, se ha estimado la regresión que figura en la salida del Cuadro 5.2

Cuadro 5.2

Dependent Variable: *ZRESID2*
Method: Least Squares
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RENTA	5.31E–05	4.30E–06	12.345760	0.000000
C	–0.651296	0.154339	–4.219902	0.000000
R-squared	0.234338	Mean dependent var		0.998000
Adjusted R-squared	0.232801	S.D. dependent var		1.973121
S.E. of regression	1.728255	Akaike info criterion		3.936093
Sum squared resid	1 487.459000	Schwarz criterion		3.952952
Log likelihood	–982.023400	F-statistic		152.417800
Durbin-Watson stat	1.963568	Prob(F-statistic)		0.000000

siendo *ZRESID2* el cuadrado de los residuos estandarizados del modelo (5.13).

A partir de la información aportada, contraste la existencia de heterocedasticidad en el modelo (5.13).

Solución

Teniendo en cuenta la información que proporciona la pregunta, vamos a utilizar el contraste de Breusch y Pagan para contrastar la presencia de heterocedasticidad en las perturbaciones del modelo (5.13). El estadístico de contraste es:

$$\frac{SCR}{2}$$

que, bajo la hipótesis nula de homocedasticidad, sigue una distribución chi-cuadrado de $p-1$ grados de libertad, siendo p el número de regresores de la regresión auxiliar del contraste (incluida la constante). La *SCR* del estadístico del contraste se puede calcular de la siguiente manera:

$$SCR = SCE\frac{R^2}{1-R^2} = 1487.459\frac{0.234338}{1-0.234338} = 455.25$$

con lo que el estadístico del contraste es igual a

$$\frac{SCR}{2} = \frac{455.25}{2} = 227.625$$

que es superior al valor crítico de una distribución chi-cuadrado de 1 grado de libertad que, para un nivel de significación del 5%, toma el valor 3.8415. Por tanto, se rechaza la hipótesis nula de homocedasticidad.

EJERCICIO 5.24

Dado el siguiente modelo en el que se estima el gasto anual en comida fuera de casa (Y) en función del salario anual para una serie de familias (X_2):

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{5.14}$$

del que, además, conocemos los siguientes datos:

$$X'X = \begin{pmatrix} 10.0 & 31.80 \\ 31.8 & 211.08 \end{pmatrix}$$

$$\sum_{i=1}^{N} e_i^2 = 167.21 \quad \sum_{i=1}^{N} e_i^4 = 4033.153 \quad \sum_{i=1}^{N} X_i^3 = 2010.64$$

$$\sum_{i=1}^{N} X_i^4 = 22050.39 \quad \sum_{i=1}^{N} e_i^2 X_i = 258.18 \quad \sum_{i=1}^{N} X_i^2 e_i^2 = 685.68$$

siendo e los residuos correspondientes a la estimación MCO de (5.14).

(a) Plantee la regresión auxiliar correspondiente al contraste de White y construya las matrices X'^*X^* y X'^*Y^* de dicha regresión auxiliar.

(b) Si se sabe que los coeficientes estimados de la anterior regresión auxiliar son

$$\hat{\alpha} = \begin{pmatrix} 31.36 \\ -7.13 \\ 0.381 \end{pmatrix}$$

correspondiendo a la constante y a los coeficientes de X_2 y X_2^2, respectivamente, calcule el estadístico del contraste de White e interprete el resultado.

(c) Conociendo que, en la estimación de la siguiente regresión auxiliar:

$$\left(\frac{e_i}{S_e}\right)^2 = \gamma_1 + \gamma_2 X_{2i} + w_i$$

donde S_e es la desviación típica de los errores, se obtiene un coeficiente de determinación igual a $R^2 = 0.55$ y una $\hat{\sigma}_w = 0.448$, contraste la existencia de heterocedasticidad en las perturbaciones del modelo (5.14). Compare los resultados con los obtenidos en el apartado (b).

Solución

(a) Las hipótesis del contraste de White son:

$$\left.\begin{array}{l} H_0\text{: Perturbaciones homocedásticas} \\ H_1\text{: Perturbaciones heterocedásticas} \end{array}\right\}$$

mientras que la regresión auxiliar correspondiente a este contraste es:

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{2i}^2 + v_i$$

Genéricamente podemos denominarla:

$$Y^* = X^* \alpha + v$$

El estadístico de contraste es NR^2, donde R^2 y p son el coeficiente de determinación de la regresión auxiliar y el número de coeficientes (incluida la constante) de esta misma regresión. Bajo la hipótesis nula, este estadístico se distribuye como una chi-cuadrado de $p-1$ grados de libertad.

La matriz X^* de la regresión auxiliar es:

$$X^* = \begin{pmatrix} 1 & X_{21} & X_{21}^2 \\ 1 & X_{22} & X_{22}^2 \\ \ldots & \ldots & \ldots \\ 1 & X_{2,N-1} & X_{2,N-1}^2 \\ 1 & X_{2N} & X_{2N}^2 \end{pmatrix}$$

Por tanto,

$$X'^*X^* = \begin{pmatrix} N & \sum_{i=1}^N X_{2i} & \sum_{i=1}^N X_{2i}^2 \\ \sum_{i=1}^N X_{2i} & \sum_{i=1}^N X_{2i}^2 & \sum_{i=1}^N X_{2i}^3 \\ \sum_{i=1}^N X_{2i}^2 & \sum_{i=1}^N X_{2i}^3 & \sum_{i=1}^N X_{2i}^4 \end{pmatrix} = \begin{pmatrix} 10.00 & 31.80 & 211.08 \\ 31.80 & 211.08 & 2010.64 \\ 211.08 & 2010.64 & 22050.39 \end{pmatrix}$$

Por otro lado, dado que $Y_i^* = e_i^2$, la matriz $X'^*Y^* = X'^*e^2$ será

$$X'^{*}e^{2} = \begin{pmatrix} \sum_{i=1}^{N} e_i^2 \\ \sum_{i=1}^{N} e_i^2 X_{2i} \\ \sum_{i=1}^{N} e_i^2 X_{2i}^2 \end{pmatrix} = \begin{pmatrix} 167.21 \\ 258.18 \\ 685.68 \end{pmatrix}$$

(b) Conociendo los coeficientes estimados de la regresión auxiliar tenemos información para calcular el coeficiente de determinación:

$$R^2 = \frac{SCR}{SCT} = \frac{\hat{\alpha}' X'^{*} Y^{*} - N\bar{Y}^{*2}}{Y'^{*} Y^{*} - N\bar{Y}^{*2}} = \frac{\hat{\alpha}' X'^{*} e^{2} - N\left(\overline{e^2}\right)^2}{\sum_{i=1}^{N} e_i^4 - N\left(\overline{e^2}\right)^2}$$

Para ello, primero calculamos $\overline{e^2}$:

$$\overline{e^2} = \frac{\sum_{i=1}^{N} e_i^2}{N} = \frac{167.21}{10} = 16.721$$

Con lo cual, podemos obtener el valor de la *SCT*:

$$SCT = \sum_{i=1}^{N} e_i^4 - N\left(\overline{e^2}\right)^2 = 4033.153 - 10 \cdot 16.721^2 = 1\,237.23$$

A continuación, calculamos la *SCR* como

$$SCR = \hat{\alpha}' X'^{*} e^{2} - N\left(\overline{e^2}\right)^2 = \begin{pmatrix} 31.36 & -7.13 & 0.381 \end{pmatrix} \begin{pmatrix} 167.21 \\ 258.18 \\ 685.68 \end{pmatrix} - 10 \cdot 16.721^2 =$$

$$= 3\,662.89 - 10 \cdot 16.721^2 = 868.207$$

Conociendo la *SCR* y la *SCT* obtenemos

$$R^2 = \frac{SCR}{SCT} = 0.7$$

Por último, el estadístico de contraste es

$$NR^2 = 10 \cdot 0.7 = 7$$

Como el valor crítico de una $\chi_2^{2(0.05)}$ es 5.99, se rechaza la hipótesis nula de homocedasticidad.

(c) La regresión auxiliar planteada en este apartado se construye para realizar el contraste de Breusch y Pagan.

Las hipótesis del contraste son:

$$\left.\begin{array}{l} H_0\text{: Perturbaciones homocedásticas} \\ H_1\text{: Perturbaciones heterocedásticas} \end{array}\right\}$$

El estadístico de contraste es:

$$\frac{SCR}{2}$$

donde *SCR* es la suma de cuadrados explicada correspondiente a la regresión auxiliar que utiliza este contraste. Bajo la hipótesis nula, el estadístico de contraste se distribuye como una chi-cuadrado de $p-1$ grados de libertad, siendo p el número de regresores de la regresión auxiliar (incluida la constante). La *SCR* se puede obtener calculando primero la *SCE* de la regresión auxiliar y sustituyendo posteriormente este valor en la expresión del coeficiente de determinación:

$$R^2 = 1 - \frac{SCE}{SCT} \tag{5.15}$$

Para calcular la *SCE* sabemos que $\sigma_w^2 = \dfrac{SCE}{N-p}$.

Por tanto, $SCE = \sigma_w^2 (N-p) = 0.4488^2 (10-2) = 1.611$

A partir de (5.15) despejamos la *SCT*:

$$SCT = \frac{SCE}{1-R^2} = \frac{1.611}{1-0.55} = 3.58$$

Por otro lado, sabemos que $SCT = SCR + SCE$. Despejando *SCR*, tenemos que:

$$SCR = 3.58 - 1.611 = 1.969$$

El estadístico de contraste es $1.969/2 = 0.9845$. El valor crítico al 95% de nivel de confianza de una χ_1^2 vale 3.84. Por tanto, no podemos rechazar la hipótesis nula de homocedasticidad.

Como vemos, el resultado difiere con el contraste de White, sin embargo, hemos de tener en cuenta que los grados de libertad en estos casos son realmente pequeños, especialmente en el caso del contraste de White, lo

que lleva a una mayor probabilidad de rechazar la hipótesis nula de estos contrastes. No se debe perder de vista el hecho de que estos contrastes tienen una validez asintótica y que en estos ejercicios se utilizan muestras pequeñas únicamente a efectos meramente ilustrativos.

EJERCICIO 5.25

Dada la salida de regresión del Cuadro 5.3, referida a la regresión auxiliar de un contraste de heterocedasticidad, donde e representa a los errores de la estimación de un determinado modelo:

Cuadro 5.3

Test Equation:
Dependent Variable: e^2
Method: Least Squares
Sample: 1 1980
Included observations: 1980

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1 224.704000	1 702.039000	0.719551	0.4719
X2	0.365144	1.108049	0.329538	0.7418
X2^2	–2.78E–05	0.000245	–0.113420	0.9097
X2*X3	0.040795	0.050002	0.815862	0.4147
X2*X4	–0.196071	0.114199	–1.716917	0.0862
X3	8.347079	239.991100	0.034781	0.9723
X3^2	3.179366	12.710690	0.250133	0.8025
X3*X4	–42.162440	25.993600	–1.622032	0.1050
X4	1 633.378000	488.238300	3.345453	0.0008
X4^2	–198.022500	74.198690	–2668815	0.0077
R-squared	0.008564	Mean dependent var		3 551.284000
Adjusted R-squared	0.004035	S.D. dependent var		3 306.045000
S.E. of regression	3 299.369000	Akaike info criterion		19.045890
Sum squared resid	2.14E+10	Schwarz criterion		19.074120
Log likelihood	–18 845.430000	F-statistic		1.890745
Durbin-Watson stat	2.059028	Prob(F-statistic)		0.000000

(a) ¿Cuál es el modelo para el que se está contrastando la heterocedasticidad?

(b) ¿Qué contraste de heterocedasticidad se está realizando? Especifique sus hipótesis.

(c) Concluya sobre el contraste anterior.

Solución

(a) El modelo para el que se está contrastando la heterocedasticidad es

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \qquad (5.16)$$

(b) Se está realizando el contraste de White, ya que la salida de regresión muestra la estimación de los errores al cuadrado en función de la constante, las variables X_2, X_3 y X_4, de sus cuadrados y de sus interacciones. Esta regresión coincide con la regresión auxiliar que se utiliza para contrastar la hipótesis nula de que el modelo (5.16) es homocedástico frente a la alternativa de que es heterocedástico.

(c) El estadístico de prueba para este contraste es NR^2. Si se cumple la hipótesis nula este estadístico se distribuye como una χ^2_{p-1}, en donde p es el número de parámetros de la regresión auxiliar. De dicha salida se obtiene que $N = 1980$, $R^2 = 0.008564$ y $p = 10$, con lo que el estadístico de contraste es igual a $NR^2 = 1980 \cdot 0.008564 = 16.956$. Además, teniendo en cuenta la tabulación de la función de distribución de una variable chi-cuadrado, se sabe que el punto que deja a su derecha una probabilidad igual al 5%, cuando la distribución tiene 9 grados de libertad, es 16.92. Tal y como se puede observar, este valor es ligeramente inferior al estadístico de contraste, con lo que se rechazaría la hipótesis nula de homocedasticidad para este nivel de significación. Si el nivel de significación elegido fuera del 1%, la hipótesis nula de homocedasticidad no sería rechazada.

EJERCICIO 5.26

Con la información del Cuadro 5.4 y del Cuadro 5.5, calcule el primer y el último dato de la variable endógena a utilizar en la regresión auxiliar que usa el contraste de Breusch-Pagan, que analiza la heterocedasticidad/homocedasticidad de la perturbación aleatoria del siguiente modelo:

$$FORMA_i = \beta_1 + \beta_2 INGRE_i + u_i$$

Cuadro 5.4

Dependent Variable: *FORMA*
Method: Least Squares
Sample: 1 5
Included observations: 5

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	100.00000	117.842100	–0.845762	0.4598
INGRE	0.10000	0.013743	5.364932	0.0127
R-squared	0.905608	Mean dependent var		480.473100
Adjusted R-squared	0.874145	S.D. dependent var		295.199500
S.E. of regression	104.725300	Akaike info criterion		12.429730
Sum squared resid	32 902.170000	Schwarz criterion		12.273510
Log likelihood	–29.074330	F-statistic		28.782490
Durbin-Watson stat	1.255836	Prob(F-statistic)		0.012675

Cuadro 5.5

Errores
79.956
–154.059

Solución

El contraste de Breusch-Pagan utiliza una regresión auxiliar cuya variable endógena es

$$\left(\frac{\text{Error}}{\text{Desviación típica muestral del error}}\right)^2$$

Dado que el modelo tiene constante, la media de los errores es cero y la varianza muestral del error se calcula como el cociente entre la SCE y N. La salida de regresión nos muestra que $SCE = 32902.17$ y $N = 5$. Por tanto, la varianza muestral del error es 6580.4 y su raíz cuadrada vale 81.1. En consecuencia, el primer dato de la variable pedida es $(79.956/81.1)^2$ y el último dato es $(-154.059/81.1)^2$.

EJERCICIO 5.27

Se tiene el modelo estimado:

$$Y_i = 1.93 + 1.15X_{2i} + e_i \quad N = 15 \quad e'e = 0.90 \tag{5.17}$$

del que además se dispone de la siguiente información:

$$\sum X_{2i} = 15.09 \quad \sum X_{2i}^2 = 16.3883 \quad \sum X_{2i}^3 = 19.11636 \quad \sum X_{2i}^4 = 23.78774$$

$$\sum e_i^2 X_{2i} = 0.922429 \quad \sum e_i^2 X_2^2 = 0.961396 \quad \sum e_i^4 = 0.190748$$

Con los datos aportados, realice el contraste de heterocedasticidad de White del modelo (5.17). A la vista del resultado, ¿qué conclusiones saca?

Solución

La regresión auxiliar del contraste de White es:

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{2i}^2 + v_i$$

y el estadístico de contraste es NR^2, siendo R^2 el coeficiente de determinación de la regresión auxiliar. La hipótesis nula del contraste de White es la existencia de homocedasticidad.

La estimación de la regresión auxiliar implica la necesidad de obtener las siguientes expresiones:

$$X'X = \begin{pmatrix} N & \sum_{i=1}^{N} X_{2i} & \sum_{i=1}^{N} X_{2i}^2 \\ & \sum_{i=1}^{N} X_{2i}^2 & \sum_{i=1}^{N} X_{2i}^3 \\ & & \sum_{i=1}^{N} X_{2i}^4 \end{pmatrix} = \begin{pmatrix} 15 & 15.0900 & 16.38830 \\ & 16.3883 & 19.11636 \\ & & 23.78774 \end{pmatrix}$$

$$\Rightarrow (X'X)^{-1} = \begin{pmatrix} 8.59354 & -16.08205 & 7.00347 \\ & 31.07081 & -13.88965 \\ & & 6.37911 \end{pmatrix}$$

$$X'e^2 = \begin{pmatrix} \sum_{i=1}^{N} e_i^2 \\ \sum_{i=1}^{N} e_i^2 X_{2i} \\ \sum_{i=1}^{N} e_i^2 X_{2i}^2 \end{pmatrix} = \begin{pmatrix} 0.900000 \\ 0.922429 \\ 0.961396 \end{pmatrix}$$

donde la matriz X recoge en cada una de sus columnas los regresores de la regresión auxiliar y el vector e^2 contiene, en cada uno de sus elementos, el cuadrado de los errores de la estimación que se presenta en la ecuación (5.17).

La estimación de la regresión auxiliar es:

$$\hat{\alpha} = (X'X)^{-1} X'e^2 = \begin{pmatrix} -0.367252 \\ 0.833315 \\ -0.376242 \end{pmatrix}$$

El sumatorio del cuadrado de los residuos de la regresión auxiliar y su coeficiente de determinación, respectivamente, son

$$\hat{v}'\hat{v} = \sum_{i=1}^{N} e_i^4 - \hat{\alpha}' X'e^2 = 0.190748 - 0.0764309 = 0.1143171$$

$$R^2 = 1 - \frac{\hat{v}'\hat{v}}{\sum_{i=1}^{N} e_i^4 - N\left(\overline{e^2}\right)^2} = 1 - \frac{0.1143171}{0.190748 - 15\left(\frac{0.90}{15}\right)^2} = 0.1640$$

El estadístico de contraste da como resultado:

$$NR^2 = 15 \cdot 0.1640 = 2.46$$

Por tanto, como dicho valor es inferior al valor tabulado ($\chi_2^{2(\alpha=0.05)} = 5.99$), no se rechaza la hipótesis nula de homocedasticidad.

Esta obra contiene una relación de ejercicios resueltos de Econometría, que aborda los principales tópicos que permiten introducir al lector en el modelo de regresión lineal múltiple y profundizar en el mismo.

La principal aportación de la misma es que incentiva el autoaprendizaje, en la línea de los objetivos perseguidos con el Espacio Europeo de Educación Superior. En este afán, la resolución de los ejercicios se realiza de forma detallada, facilitando al lector la comprensión y aplicabilidad de los conceptos teóricos.

En el texto se ha incrementado el número de tópicos que se tratan habitualmente en otras obras, incluyendo temas como los de forma funcional, observaciones atípicas y caracteres cualitativos. Además, aunque se utilicen algunas salidas de programas econométricos, los objetivos perseguidos van más allá de la mera interpretación de éstas.

Los autores son todos ellos profesores de la Facultad de Ciencias Económicas y Empresariales de la Universidad de Las Palmas de Gran Canaria y entre sus líneas de trabajo e investigación descatan las de Econometría Financiera, Economía Laboral y de la Educación, Economía Regional y Números Índices.

DELTA PUBLICACIONES

C/ Luarca, 11
28230 Las Rozas de Madrid
MADRID
Tel. 91 637 16 88

www.deltapublicaciones.com

ISBN 84-96477-55-X