
Introducción a la econometría
Ezequiel Uriel
Universidad de Valencia 2019

Diseño portada: Jordi Uriel

Introducción a la Econometría

Ezequiel Uriel

2019

Universidad de Valencia

Deseo agradecer a los profesores Luisa Moltó, Amado Peiró, Paz Rico, Pilar Beneito y Javier Ferri sus sugerencias por las erratas que han detectado en versiones previas, y por haberme facilitado datos para formular ejercicios. También en la detección de erratas han colaborado algunos alumnos. En cualquier caso, soy el único responsable de las erratas que no han sido detectadas

Tabla de contenido

1 ECONOMETRÍA Y DATOS ECONÓMICOS

1.1 ¿Qué es la econometría?

En primer lugar, veamos algo sobre el origen de la econometría como disciplina. El término econometría se cree que fue acuñado por Ragnar Frisch co-ganador del primer Premio Nobel en Ciencias Económicas en 1969, junto con el también económetra Jan Tinbergen. Ambos fueron fundadores de la Econometric Society en 1933. En la sección I de la constitución de esta sociedad, se afirma que

> "La Econometric Society es una sociedad internacional para el avance de la teoría económica en su relación con la estadística y las matemáticas. Su principal objetivo será promover que tengan como objetivo la unificación de los enfoques cuantitativo-teórico y cuantitativo-empírico de los problemas económicos y que son abordados de forma constructiva y rigurosa similar al que es el enfoque predominante en las ciencias naturales"

En el primer número de *Econometrica* (1933), revista de la Econometric Society, Ragnar Frisch nos da una explicación sobre el significado de la econometría:

> "Pero hay varios aspectos del enfoque cuantitativo de la economía, aunque ninguno de ellos, tomado aisladamente, debería confundirse con la econometría. Así, la econometría no es lo mismo que la estadística económica. Tampoco es idéntica a lo que llamamos teoría económica general, aunque una parte considerable de esta teoría tenga un carácter definitivamente cuantitativo. Tampoco debe tomarse la econometría como sinónimo de la aplicación de las matemáticas a la economía. La experiencia ha demostrado que cada uno de estos tres puntos de vista, el de la estadística, la teoría económica y las matemáticas, es una condición necesaria, pero no por sí misma una condición suficiente, para una verdadera comprensión de las relaciones cuantitativas en la vida económica moderna. Se trata de la unificación de los tres aspectos lo que le da gran alcance. Y es esa unificación, lo que constituye la econometría"

Hoy en día, también se dice que la econometría es el estudio combinado de los modelos económicos, estadística matemática y datos económicos. Dentro del campo de la econometría se puede distinguir la teoría econométrica de la econometría aplicada.

La *teoría econométrica* se refiere al desarrollo de las herramientas y métodos, y al estudio de las propiedades de los métodos econométricos. La teoría econométrica pertenece al ámbito de la estadística.

La *econometría aplicada* es un término que describe la elaboración de modelos económicos cuantitativos y la aplicación de métodos econométricos a estos modelos utilizando datos económicos. La econometría aplicada se encuentra, básicamente, dentro del campo de la economía aplicada.

¿Cuáles son los objetivos de la econometría? Vamos a considerar tres objetivos de la econometría:

1) *El conocimiento de la economía real*. Los métodos econométricos nos permiten estimar las magnitudes económicas, como la propensión marginal al consumo o la elasticidad de la mano de obra con respecto al output. Estas estimaciones se sitúan en un determinado tiempo y espacio: por ejemplo, en España en el último cuarto del siglo XX. Además de la estimación, en la que se obtienen los valores numéricos, los métodos econométricos nos permiten realizar contrastes de hipótesis, por ejemplo, ¿en una función de producción es admisible la hipótesis de rendimientos a escala constantes?
2) *Simulación de políticas económicas*. Los métodos de econometría pueden ser utilizados para simular los efectos de políticas alternativas. Por ejemplo, con un modelo econométrico apropiado, se podría determinar, en términos cuantitativos, el efecto de diferentes tipos del impuesto del tabaco sobre el consumo de tabaco.
3) Predicción. Muy a menudo los métodos econométricos se utilizan para predecir valores de variables económicas. Cuando hacemos predicciones tratamos de reducir nuestra incertidumbre sobre el futuro de la economía. Esto no es una tarea fácil, ya que, en general, las predicciones sólo son satisfactorias cuando no hay cambios drásticos en la economía. Sería muy conveniente también predecir estos cambios drásticos, pero las predicciones con métodos econométricos por lo general no son muy buenas en esos casos, aunque tampoco funcionan otros métodos alternativos.

1.2 Etapas en la elaboración de un modelo econométrico

En la elaboración de un modelo econométrico se pueden distinguir las siguientes etapas: especificación, estimación y validación.

Si bien en una primera aproximación estas etapas siguen un orden secuencial, en la elaboración de un modelo econométrico es necesario, por regla general, retroceder en más de una ocasión dentro de este orden secuencial. Es decir, en el análisis econométrico no se sigue un orden establecido de antemano, sino que es necesario confrontar continuamente el modelo con los datos y con cualquier otra fuente de información, con la finalidad de obtener un modelo econométrico compatible con los datos, que permita analizar la realidad, ofrezca mejores predicciones o constituya una buena base para tomar decisiones. Se procede a continuación a describir las etapas enumeradas anteriormente.

(a) Especificación

La primera etapa de la elaboración de un modelo econométrico la constituye la especificación.

En la etapa de especificación, vamos a considerar cuatro elementos: el modelo económico, el modelo econométrico, los supuestos estadísticos del modelo y los datos. En este apartado vamos a referirnos a los tres primeros elementos, mientras que en el epígrafe 1.3 examinaremos los diferentes tipos de datos utilizados en el análisis econométrico.

El primer elemento que necesitamos es disponer de un modelo económico. En algunos casos, un modelo formal económico se especifica completamente mediante la

utilización de la teoría económica. En otros casos, la teoría económica se utiliza menos formalmente en la construcción de un modelo económico.

Después de obtener un modelo económico, tenemos que convertirlo a un modelo econométrico. Vamos a ver con dos ejemplos como se realiza este proceso.

EJEMPLO 1.1 Función de consumo keynesiana

Keynes formuló su conocida función de consumo a través de las siguientes proposiciones:

Proposición 1: El consumo es una función de la renta, y ambas variables están medidas en términos reales. Si las variables se miden en términos reales, esto significa que cuando los consumidores deciden la proporción de renta que van a dedicar al consumo, no se ven afectados por ilusión monetaria.

Analíticamente, la proposición 1 se puede expresar de la siguiente manera:

$$cons = f(renta) \tag{1-1}$$

Proposición 2: El consumo es una función creciente de la renta, pero un aumento de la renta produce siempre un aumento de menor magnitud en el consumo.

Esta proposición implica que la propensión marginal al consumo es mayor que 0 (que es una función creciente), pero es menor que 1 (un aumento de la renta siempre causa un aumento de menor magnitud en el consumo).

Analíticamente, la proposición 2 se puede expresar de la siguiente manera:

$$0 < \frac{\mathrm{d}\,cons}{\mathrm{d}\,renta} < 1 \tag{1-2}$$

Proposición 3: La proporción de la renta dedicada al consumo es menor cuando aumenta la renta. Es decir, la proporción del último euro ganado destinado al consumo es más pequeña que la proporción de la renta total destinada al consumo.

Analíticamente, la proposición 3 se puede expresar de la siguiente manera:

$$\frac{\mathrm{d}con}{\mathrm{d}\,renta} < \frac{cons}{renta} \tag{1-3}$$

En otras palabras, la propensión marginal al consumo es menor que la propensión media al consumo.

Estas tres proposiciones constituyen un modelo económico: la función de consumo keynesiana.

Para estimar y contrastar este modelo tenemos que convertirlo en un modelo econométrico. Para esta conversión deben cumplirse dos requisitos.

De acuerdo con el primer requisito, es necesario especificar la forma matemática de la función. En este caso se ha utilizado la función lineal, debido a que, además de ser simple, es compatible con la descripción hecha por Keynes.

Con el fin de cumplir con la segunda exigencia, debe tenerse en cuenta que el modelo formulado en la proposición 1 es determinista. Es decir, la renta es el único factor que se tiene en cuenta para la determinación del consumo. Pero en la vida real hay muchos otros factores, distintos de la renta, que tienen una influencia en el consumo. En un modelo econométrico todos los factores diferentes de las variables independientes incluidas se reúnen en una variable denominada perturbación aleatoria o error (u). Por lo tanto, el segundo requisito es la introducción del término de error en la ecuación.

En general, todos los factores relevantes deben ser introducidos de forma explícita en el modelo econométrico, y el resto de los factores se agrupan en una única variable: el error o perturbación aleatoria. En la función de consumo keynesiana el único factor considerado relevante es la renta.

Teniendo en cuenta estos dos requisitos la función de consumo keynesiana se puede expresar de la siguiente manera:

$$cons = \beta_1 + \beta_2 renta + u \tag{1-4}$$

Éste es un modelo econométrico que puede estimarse si se dispone de datos sobre consumo y renta. Veamos ahora las otras dos proposiciones. En este modelo lineal la propensión marginal al consumo es la siguiente:

$$\frac{\mathrm{d}\,cons}{\mathrm{d}\,renta} = \beta_2 \tag{1-5}$$

En consecuencia, la proposición 2 en este modelo es la siguiente:

$$0 < \beta_2 < 1 \tag{1.6}$$

Una vez que el modelo se ha estimado, es posible comprobar si la estimación de β_2 se encuentra entre 0 y l.

En el modelo lineal la propensión media al consumo, considerando que el error es igual a 0, es la siguiente:

$$\frac{cons}{renta} = \frac{\beta_1 + \beta_2 renta}{renta} = \frac{\beta_1}{renta} + \beta_2 \tag{1-7}$$

Por lo tanto, la proposición 3 implica que

$$\frac{\beta_1}{renta} + \beta_2 > \beta_2 \quad \text{or} \quad \frac{\beta_1}{renta} > 0 \tag{1-8}$$

Es decir,

$$\beta_1 > 0 \tag{1-9}$$

Una vez que el modelo se ha estimado, contrastar la proposición 3 es equivalente a contrastar si el término independiente es significativamente mayor que 0.

EJEMPLO 1.2 Determinación de los salarios

Modelo económico:

La teoría económica formal - la teoría del capital humano- dice que la educación (*educ*), la experiencia (*exper*) y el aprendizaje (*aprend*) son factores que afectan la productividad y por lo tanto al *salario*. Entonces, un modelo económico para explicar el salario podría ser el siguiente:

$$salario = f(educ, exper, aprend) \tag{1-10}$$

Por cierto, en su opinión, ¿cree Ud. que falta alguna variable en este modelo?

Modelo econométrico:

El modelo econométrico, que corresponde utilizando una forma lineal matemática, es el siguiente:

$$salario = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 aprend + u \tag{1-11}$$

En resumen, para convertir un modelo económico en un modelo econométrico:

a) Se ha especificado la forma de la función *f* (.).

b) Se ha incluido en el modelo una perturbación aleatoria que recoge el efecto conjunto de otras variables que también afectan a los salarios, pero que no figuran en el modelo.

Un elemento importante en la especificación del modelo es la formulación de un conjunto de supuestos estadísticos, que se utilizan en las etapas siguientes. Estos supuestos estadísticos juegan un papel clave en el contraste de hipótesis y, en general, en todo el proceso de inferencia llevado a cabo con el modelo.

(b) Estimación

En la estimación se obtienen los valores numéricos de los coeficientes de un modelo econométrico. Para completar esta etapa se debe disponer de un conjunto de observaciones de todas las variables observables que aparecen en el modelo econométrico especificado, y, por otro lado, es necesario seleccionar el método de estimación apropiado, teniendo en cuenta las implicaciones de esta elección en las propiedades estadísticas de los estimadores de los coeficientes. La distinción entre un

estimador y una estimación debe quedar clara. Un estimador es el resultado de aplicar un método de estimación a una especificación econométrica. Por otra parte, una estimación consiste en la obtención de un valor numérico de un estimador para una muestra dada. Por ejemplo, la aplicación del método de *mínimos cuadrados* a la especificación de la función de consumo (1-4) proporciona expresiones que determinan los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$. Sustituyendo los datos muestrales en esas expresiones, se obtienen dos valores: un valor para $\hat{\beta}_1$ y otro para $\hat{\beta}_2$, que son las estimaciones de los parámetros β_1 y β_2.

En general, es posible obtener expresiones analíticas de los estimadores, particularmente en el caso de la estimación de relaciones lineales. Sin embargo, en los procedimientos de estimación no lineal es a menudo difícil establecer su expresión analítica.

(c) Validación

En la etapa de validación se evalúan los resultados. En esta etapa se evalúa si las estimaciones obtenidas en la etapa anterior son aceptables, tanto por la teoría económica como desde el punto de vista estadístico. Se analiza, por un lado, si las estimaciones de los parámetros del modelo tienen los signos y magnitudes esperados, es decir, si satisfacen las limitaciones establecidas por la teoría económica.

Desde el punto de vista estadístico, por otro lado, se llevan a cabo contrastes estadísticos sobre la significatividad de los parámetros del modelo en los que se utilizan los supuestos estadísticos formulados en la etapa de especificación. A su vez, es importante contrastar si los supuestos estadísticos del modelo econométrico se cumplen, aunque hay que tener en cuenta que no todos los supuestos son contrastables. La violación de alguno de estos supuestos implica, en general, la aplicación de otros métodos de estimación, que permitan obtener estimadores que gocen de las mejores propiedades estadísticas posibles.

Una manera de establecer si el modelo es adecuado para hacer predicciones es utilizar el modelo fuera del período muestral, y después comparar los valores predichos de la variable endógena con los valores realmente observados.

1.3 Datos económicos

Como hemos visto, el análisis empírico utiliza datos para contrastar una teoría o para estimar una relación. Es importante destacar que en Econometría utilizamos datos no experimentales. Los datos no experimentales se recogen mediante la observación del mundo real de una manera pasiva. En este caso los datos no son el resultado de experimentos controlados. Los datos experimentales se recogen a menudo en entornos de laboratorio, como ocurre en las ciencias naturales.

Ahora, vamos a ver tres tipos de datos que se pueden utilizar en la estimación de un modelo econométrico: series temporales, datos de corte transversal y datos panel.

Series temporales

En las series temporales, los datos son observaciones de una variable a lo largo del tiempo. Por ejemplo: magnitudes de las cuentas nacionales, como el consumo, las importaciones, ingresos, etc. El orden cronológico de las observaciones proporciona

información potencialmente importante. En consecuencia, en una serie temporal la ordenación de las observaciones es relevante.

No se puede asumir que los datos de series temporales sean independientes a través del tiempo. La mayoría de las series económicas se relacionan con sus historias recientes. Ejemplos típicos son los agregados macroeconómicos como los precios y los tipos de interés. Este tipo de datos se caracteriza por la dependencia serial, de forma que el supuesto de muestreo aleatorio resulta inapropiado en este caso.

La mayoría de los datos económicos agregados sólo están disponibles para frecuencias bajas (anual, trimestral o mensual en algunas ocasiones) por lo que el tamaño de la muestra suele ser mucho menor que en los típicos estudios de corte transversal. La excepción son los datos financieros, donde se dispone de datos para frecuencias más elevadas (semanal, diaria, por hora, etc.) de forma que los tamaños muestrales pueden ser muy grandes.

Datos de corte transversal

En los datos de corte transversal se dispone de una observación por individuo y se refieren a un punto determinado en el tiempo. En la mayoría de los estudios, los individuos encuestados son personas (por ejemplo, en la Encuesta de Población Activa (EPA), más de 100000 personas son entrevistadas cada trimestre), hogares (por ejemplo, la Encuesta de Presupuestos Familiares), empresas (por ejemplo, la Encuesta de Empresas Industriales) u otros agentes económicos. Las encuestas son una fuente típica para datos de corte transversal. En muchos estudios econométricos contemporáneos de corte transversal el tamaño muestral es bastante elevado.

En los datos de corte transversal, las observaciones deben ser obtenidas mediante un muestreo aleatorio, lo que implica que las observaciones sean independientes entre sí. El orden de las observaciones en los datos de corte transversal no importa para el análisis econométrico. Si los datos no se obtienen con una muestra aleatoria, tendremos un problema de selección muestral.

Hasta ahora nos hemos referido a datos de tipo de micro, pero también se pueden tener datos de corte transversal relativos a unidades agregadas, como países, regiones, etc. Por supuesto, los datos de este tipo no se obtienen mediante un muestreo aleatorio.

Datos de panel

Los datos de panel (o datos longitudinales) consisten en observaciones de corte transversal repetidas a lo largo del tiempo. Así pues, los datos panel combinan elementos de datos de corte transversal y de series temporales. Estos conjuntos de datos consisten en un conjunto de individuos (por lo general personas, hogares o empresas) encuestados repetidamente a lo largo del tiempo. En la modelización se adopta generalmente el supuesto de que los individuos son independientes entre sí, pero que, para un individuo dado, las observaciones a lo largo del tiempo son mutuamente dependientes. Por lo tanto, el orden dentro de un corte transversal de un conjunto de datos panel no importa, pero el orden en la dimensión temporal es relevante. Si no tenemos en cuenta el tiempo en datos de panel, se dice que estamos utilizando datos de corte transversal agrupados (*pooled*).

2 EL MODELO DE REGRESIÓN LINEAL SIMPLE: ESTIMACIÓN Y PROPIEDADES

2.1 Algunas definiciones en el modelo de regresión simple

2.1.1 El modelo de regresión poblacional y la función de regresión poblacional

En el modelo de regresión simple, el *modelo de regresión poblacional* o, simplemente, el *modelo poblacional* es el siguiente:

$$y = \beta_1 + \beta_2 x + u \tag{2-1}$$

Vamos a ver los diferentes elementos del modelo (2-1) y la terminología utilizada para designarlos. En primer lugar, en el modelo hay tres tipos de variables: *y*, *x* y *u*. En este modelo el único un factor explícito para explicar *y* es *x*. El resto de los factores que afectan a *y* están recogidos en *u*.

Denominamos a *y* variable endógena (del griego: generada dentro) o variable dependiente. Se utilizan también otras denominaciones para designar a *y*: variable explicada o regresando. En este modelo todas estas denominaciones son equivalentes, pero en otros modelos, como veremos más adelante, puede haber algunas diferencias.

En la regresión lineal simple de *y* sobre *x*, a la variable *x* se le denomina variable exógena (del griego: generado fuera de) o variable independiente. Otras denominaciones utilizadas también para designar a *x* son: variable explicativa, regresor, covariable o variable de control. Todas estas denominaciones son equivalentes, pero en otros modelos, como veremos más adelante, puede haber algunas diferencias.

La variable *u* recoge todos aquellos factores distintos de *x* que afectan a *y*. Es denominada error o perturbación aleatoria. El término de perturbación puede captar también el error de medición de la variable dependiente. La perturbación es una variable no observable.

Los parámetros β_1 y β_2 son fijos y desconocidos.

En el segundo miembro de (2-1) se pueden distinguir dos componentes: un componente sistemático $\beta_1 + \beta_2 x$ y la perturbación aleatoria *u*. Llamando μ_y al componente sistemático, podemos escribir:

$$\mu_y = \beta_1 + \beta_2 x \tag{2-2}$$

Esta ecuación es conocida como la función de regresión poblacional (*FRP*) o recta poblacional. Por lo tanto, como puede verse en la figura 2.1, μ_y es una función lineal de *x* con término independiente igual a β_1 y pendiente igual a β_2 .

La linealidad significa que un aumento de una unidad en *x* implica que el *valor esperado* de *y* - $\mu_y = E(y)$ - varíe en β_1 unidades.

Ahora, supongamos que disponemos de una muestra aleatoria de tamaño *n* {(y_i, x_i): i = 1, ...,n} extraída de la población estudiada. En el diagrama de dispersión de la figura 2.2, se muestran los hipotéticos valores de la muestra.

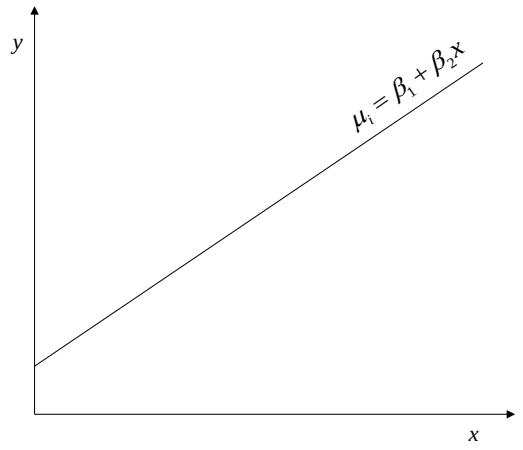


FIGURA 2.1. La función de regresión poblacional. (*FRP*)

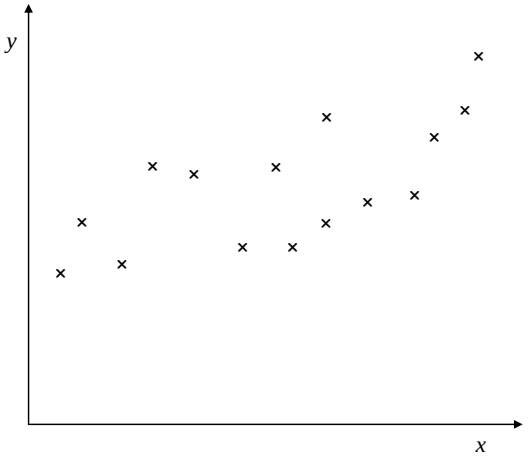


FIGURA 2.2. Diagrama de dispersión.

El modelo poblacional para cada observación de la muestra se puede expresar de la siguiente forma:

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad i = 1, 2, \cdots, n \tag{2-3}$$

En la figura 2.3 se ha representado conjuntamente la función de regresión poblacional y el diagrama de dispersión, pero es importante no olvidar que β_1 y β_2 son fijos, pero desconocidos. De acuerdo con este modelo es posible, desde un punto de vista teórico, hacer la siguiente descomposición:

$$y_i = \mu_{yi} + u_i \quad i = 1, 2, \cdots, n \tag{2-4}$$

que ha sido representada en la figura 2.3 para la observación *i*-ésima. Sin embargo, desde un punto de vista empírico, no es posible hacerlo debido a que β_1 y β_2 son desconocidos y, consecuentemente, u_i es no observable.

2.1.2 La función de regresión muestral

El objetivo principal del *modelo de regresión* es la determinación o estimación de β_1 y β_2 a partir de una muestra dada.

La *función de regresión muestral* (*FRM*) es la contrapartida de la función de regresión poblacional (*FRP*). Dado que la *FRM* se obtiene para una muestra dada, una nueva muestra generará otra estimación distinta.

La *FRM*, que es una estimación de la *FRP*, viene dada por

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-5}$$

y permite calcular el *valor ajustado* ($\hat{y}_i$) para *y* cuando $x = x_i$. En la *FRM* $\hat{\beta}_1$ y $\hat{\beta}_2$ son los estimadores de los parámetros β_1 y β_2 . Para cada x_i tenemos un valor observado (y_i) y un valor ajustado ($\hat{y}_i$).

A la diferencia entre y_i e $\hat{y}_i$ se le denomina residuo $\hat{u}_i$:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \tag{2-6}$$

En otras palabras, el residuo $\hat{u}_i$ es la diferencia entre el valor muestral y_i y el valor ajustado de $\hat{y}_i$, según puede verse en la figura 2.4. En este caso sí es posible calcular empíricamente la descomposición para una muestra dada:

$$y_i = \hat{y}_i + \hat{u}_i$$

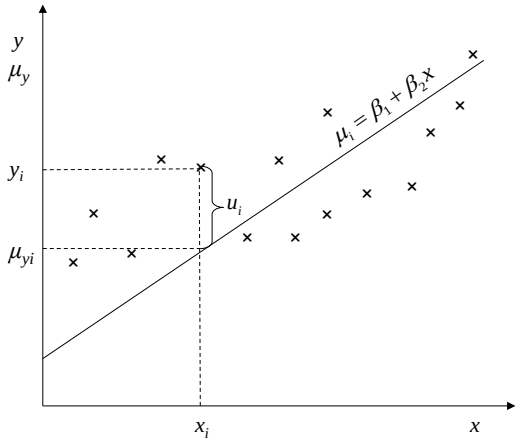


FIGURA 2.3. La función de regresión poblacional y el diagrama de dispersión.

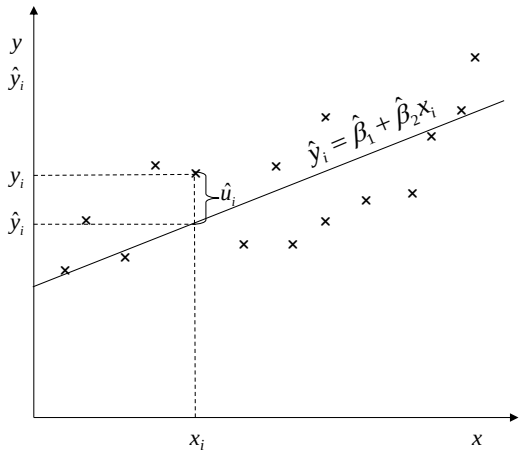


FIGURA 2.4. La función de regresión muestral y el diagrama de dispersión.

Resumiendo, $\hat{\beta}_1, \hat{\beta}_2, \hat{y}_i$ y $\hat{u}_i$ son la contrapartida muestral de β_1, β_2, μ_{yi} y u_i respectivamente. Es posible calcular $\hat{\beta}_1$ y $\hat{\beta}_2$, para una muestra dada, pero para cada muestra las estimaciones serán distintas. Por el contrario, β_1 y β_2 son fijos pero desconocidos.

2.2 Obtención de las estimaciones por Mínimos Cuadrados Ordinarios (*MCO*)

2.2.1 Diferentes criterios de estimación

Antes de obtener las estimaciones por mínimos cuadrados, vamos a examinar tres métodos alternativos para ilustrar el problema que tenemos planteado. Estos tres métodos tienen en común que tratan de minimizar, de alguna forma, el valor de los residuos en su conjunto.

Criterio 1

Un primer criterio consistiría en tomar como estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ a aquellos valores que hagan la suma de todos los residuos tan próxima a cero como sea posible. Con este criterio la expresión a minimizar sería la siguiente:

$$\text{Min}\left|\sum_{i=1}^{n}\hat{u}_i\right| \tag{2-7}$$

El problema principal de este método de estimación radica en que los residuos de distinto signo pueden compensarse. Tal situación puede observarse gráficamente en la figura 2.5, en la que se representan tres observaciones alineadas, (x_i,y_i), (x_2, y_2) y (x_3, y_3). En este caso, ocurre lo siguiente:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}$$

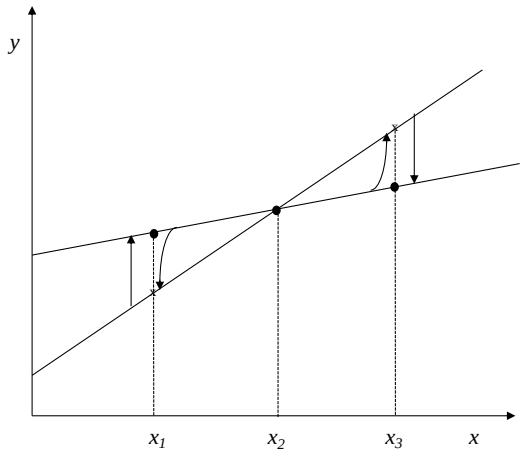


FIGURA 2.5. Los problemas del criterio 1.

Si una línea recta se ajusta de forma que pase a través de los tres puntos, cada uno de los residuos tomará el valor cero, de modo que

$$\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$$

Este ajuste podría ser considerado óptimo. Pero también es posible obtener $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$, mediante la rotación de la línea recta - desde el punto x_2, y_2 - en cualquier dirección, como muestra la figura 2.5, porque $\hat{u}_3 = -\hat{u}_1$. En otras palabras, haciendo girar de esta manera la recta, se obtiene siempre el resultado de que $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$. Este simple ejemplo muestra que este criterio no es adecuado para la estimación de los parámetros, ya que, para cualquier conjunto de observaciones, existe un número infinito de líneas rectas que satisfacen este criterio.

Criterio 2

Con el fin de evitar la compensación de los residuos positivos con los negativos, de acuerdo con este criterio se toman los valores absolutos de los residuos. En este caso se minimizaría la siguiente expresión:

$$\text{Min}\sum_{i=1}^{n}\left|\hat{u}_i\right| \tag{2-8}$$

Desgraciadamente, aunque los estimadores así obtenidos tienen algunas propiedades interesantes, su cálculo es complicado, requiriendo la resolución de un

problema de programación lineal o la aplicación de un procedimiento de cálculo iterativo.

Criterio 3

Un tercer método consiste en minimizar la suma de los cuadrados de los residuos, es decir,

(2-9)

Los estimadores obtenidos se denominan estimadores de mínimos cuadrados (*MC*), y gozan de ciertas propiedades estadísticas deseables, que se estudiarán más adelante. Por otra parte, frente al primero de los criterios examinados, al tomar los cuadrados de los residuos se evita que se compensen, mientras que, a diferencia del segundo de los criterios, los estimadores de mínimos cuadrados son sencillos de obtener. Es importante señalar que, desde el momento en que tomamos los cuadrados de los residuos, estamos penalizando más que proporcionalmente a los residuos grandes frente a los pequeños (si un residuo es el doble que otro, su cuadrado será cuatro veces mayor). Esto caracteriza la estimación de mínimos cuadrados con respecto a otros procedimientos posibles.

2.2.2 Aplicación del criterio de mínimo cuadrados

A continuación, se expone el proceso de obtención de los estimadores de *MC*. El objetivo es minimizar la suma de los cuadrados de los residuos (*S*). Para ello, en primer lugar expresamos *S* como una función de los estimadores, utilizando (2-6).

Por lo tanto

$$\underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}}\, S = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{t=1}^{T} \hat{u}_t^2 = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{2-10}$$

Para minimizar *S*, derivamos parcialmente con respecto a $\hat{\beta}_1$ y $\hat{\beta}_2$:

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i$$

Los estimadores de *MC* se obtienen igualando las anteriores derivadas a cero:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2-11}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = 0 \tag{2-12}$$

Las ecuaciones (2-11) y (2-12) se denominan *ecuaciones normales* o *condiciones de primer orden de MC*.

En las operaciones con sumatorios se deben tener en cuenta las siguientes reglas: $\sum_{i=1}^{n} a = na$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

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

Operando con las ecuaciones normales, se tiene que

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i \tag{2-13}$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 \tag{2-14}$$

Dividiendo ambos miembros de (2-13) por *n*, se tiene que

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x} \tag{2-15}$$

Por tanto,

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \tag{2-16}$$

Sustituyendo este valor de $\hat{\beta}_1$ en la segunda ecuación normal (2-14), se obtiene que

$$\sum_{i=1}^{n} y_i x_i = (\bar{y} - \hat{\beta}_2 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = \bar{y} \sum_{i=1}^{n} x_i - \hat{\beta}_2 \bar{x} \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

Resolviendo para $\hat{\beta}_2$ se tiene que:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \tag{2-17}$$

O, como se puede ver en el apéndice 2.1,

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-18}$$

Si dividimos numerador y denominador de (2-18) por *n*, se puede ver que $\hat{\beta}_2$ es el cociente entre la covarianza de las dos variables y la varianza de *x*. Por lo tanto, el signo de $\hat{\beta}_2$ es el mismo que el signo de la covarianza.

Una vez calculado $\hat{\beta}_2$, se puede obtener $\hat{\beta}_1$ utilizando la ecuación (2-16).

Estos son los estimadores de *MC*. Dado que existen métodos más complejos, que también se denominan de *MC*, al método que acabamos de desarrollar le denominaremos método de mínimos cuadrados ordinarios (*MCO*), debido a su simplicidad.

En los epígrafes precedentes, $\hat{\beta}_1$ y $\hat{\beta}_2$ se han utilizado para designar estimadores genéricos. A partir de ahora con esta notación sólo designaremos a los estimadores *MCO*.

EJEMPLO 2.1 La estimación de la función de consumo

Dada la función de consumo keynesiana,

$$cons = \beta_1 + \beta_2 renta + u_i$$

vamos a estimarla utilizando los datos de 6 hogares que aparecen en el cuadro 2.1.

CUADRO 2.1. Datos y cálculos para estimar la función de consumo.

Observ.	$cons_i$	$renta_i$	$cons_i \times renta_i$	$renta_i^2$	$cons_i - \overline{cons}$	$renta_i - \overline{renta}$	$(cons_i - \overline{cons}) \times (renta_i - \overline{renta})$	$(renta_i - \overline{renta})^2$
1	5	6	30	36	-4	-5	20	25
2	7	9	63	81	-2	-2	4	4
3	8	10	80	100	-1	-1	1	1
4	10	12	120	144	1	1	1	1
5	11	13	143	169	2	2	4	4
6	13	16	208	256	4	5	20	25
Suma	**54**	**66**	**644**	**786**	**0**	**0**	**50**	**60**

Calculando $\overline{cons}$ y $\overline{renta}$, y aplicando la fórmula (2-17), o alternativamente (2-18), a los datos de la cuadro 2.1, obtenemos:

$$\overline{cons} = \frac{54}{6} = 9\ ;\ \overline{renta} = \frac{66}{6} = 11\ ;\text{(2-17)}:\ \hat{\beta}_2 = \frac{644 - 9 \times 66}{786 - 11 \times 66} = 0.8\hat{3}\ ;\text{(2-18)}:\ \hat{\beta}_2 = \frac{50}{60} = 0.8\hat{3}$$

Aplicando después (2-16), obtenemos que $\hat{\beta}_1 = 9 - 0.8\hat{3} \times 11 = -0.1\hat{6}$

2.3 Algunas características de los estimadores de *MCO*

2.3.1 Implicaciones algebraicas de la estimación

Las implicaciones algebraicas de la estimación son derivadas exclusivamente de la aplicación del procedimiento de *MCO* al modelo de regresión lineal simple:

1. *La suma de los residuos de MCO es igual a 0*:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{2-19}$$

De la definición de los residuos:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \qquad i = 1, 2, \cdots, n \tag{2-20}$$

Si sumamos para las *n* observaciones, se obtiene:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2-21}$$

que es precisamente la primera ecuación (2-11) del sistema de ecuaciones normales.

Obsérvese que, si (2-21) se cumple, esto implica que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \tag{2-22}$$

y, dividiendo (2-19) y (2-22) por *n*, obtenemos

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \tag{2-23}$$

2. La recta de regresión de MCO pasa necesariamente por el punto ($\bar{x}, \bar{y}$).

Efectivamente, dividiendo la ecuación (2-13) por *n*, se obtiene:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x} \tag{2-24}$$

3. El producto cruzado muestral entre cada uno de los regresores y los residuos de MCO es cero.

Es decir,

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \tag{2-25}$$

Puede verse que (2-25) es igual a la segunda ecuación normal dada en (2-14):

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

4. El producto cruzado muestral entre los valores ajustados ($\hat{y}$) y los residuos de MCO es igual a cero.

Es decir,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{2-26}$$

Demostración

En efecto, teniendo en cuenta las implicaciones algebraicas 1 -(2-19)- y 3 -(2-25)-, se obtiene que

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_i) \hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i \hat{u}_i = \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 = 0$$

2.3.2 Descomposición de la varianza de *y*

Por definición

$$y_i = \hat{y}_i + \hat{u}_i \tag{2-27}$$

Restando $\bar{y}$ en ambos miembros de la expresión anterior (recordar que $\bar{\hat{y}}$ es igual a $\bar{y}$), se obtiene

$$y_i - \bar{y} = \hat{y}_i - \bar{\hat{y}} + \hat{u}_i$$

Elevando al cuadrado ambos miembros:

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

Sumando para todo *i*:

$$\sum\left[y_i-\bar{y}\right]^2=\sum(\hat{y}_i-\bar{\hat{y}})^2+\sum\hat{u}_i^2+2\sum\hat{u}_i(\hat{y}_i-\bar{\hat{y}})$$

Teniendo en cuenta las propiedades algebraicas 1 y 4, el tercer término del segundo miembro es igual a 0. Analíticamente,

$$\sum\hat{u}_i(\hat{y}_i-\bar{\hat{y}})=\sum\hat{u}_i\hat{y}_i-\bar{\hat{y}}\sum\hat{u}_i=0 \tag{2-28}$$

Por lo tanto, obtenemos

$$\sum\left[y_i-\bar{y}\right]^2=\sum(\hat{y}_i-\bar{\hat{y}})^2+\sum\hat{u}_i^2 \tag{2-29}$$

En palabras,

Suma de cuadrados totales (*SCT*) =

Suma de cuadrados explicados (*SCE*)+Suma de los cuadrados de los residuos (*SCR*)

Debe recalcarse que se debe cumplir la relación (2-15) para asegurar que (2-28) es igual a 0. Hay que recordar que (2-15) está asociada a la primera ecuación normal, es decir, a la ecuación correspondiente al término independiente. Si en el modelo ajustado no hay término independiente, entonces, en general, no se cumplirá la descomposición obtenida en (2-29).

Esta descomposición puede aplicarse a las varianzas, dividiendo ambos miembros de (2-29) por *n*:

$$\frac{\sum\left(y_i-\bar{y}\right)^2}{n}=\frac{\sum(\hat{y}_i-\bar{\hat{y}})^2}{n}+\frac{\sum\hat{u}_i^2}{n} \tag{2-30}$$

En palabras,

Varianza total=varianza explicada+varianza residual

2.3.3 Bondad del ajuste: Coeficiente de determinación (R^2)

A priori, se han obtenido unos estimadores que minimizan la suma de los cuadrados de los residuos.

Ahora, una vez hecha la estimación, podremos ver en qué medida la recta de regresión muestral se ajusta a los datos.

Una medida que indique el grado de ajuste de la recta de regresión muestral con los datos se denomina medida de *bondad del ajuste*. Vamos a estudiar ahora la medida más conocida: el *coeficiente de determinación* o *R* cuadrado (R^2). Esta medida se define de la siguiente manera:

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{2-31}$$

Por lo tanto, R^2 es la proporción de la suma de cuadrados totales (*SCT*), que se explica por la regresión (*SCE*), es decir, que se explica por el modelo. También podemos decir que 100 R^2 es el porcentaje de variación muestral de *y* explicada por *x*.

Alternativamente, teniendo en cuenta (2-29), tenemos:

$$\sum(\hat{y}_i - \bar{\hat{y}})^2 = \sum(y_i - \bar{y})^2 - \sum \hat{u}_i^2$$

Substituyendo en (2-31), tenemos

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum(y_i - \bar{y})^2 - \sum \hat{u}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SCR}{SCT} \qquad (2\text{-}32)$$

Por lo tanto, R^2 es igual a 1 menos la proporción de la suma de cuadrados totales (*SCT*), que no es explicada por la regresión (*SCR*).

De acuerdo con la definición de R^2 , debe cumplirse que

$$0 \le R^2 \le 1$$

Casos extremos:

a) Si el ajuste es perfecto, entonces se verificará $\hat{u}_i = 0 \quad \forall i$. Esto implica que

$$\hat{y}_i = y_i \quad \forall i \quad \Rightarrow \sum(\hat{y}_i - \bar{\hat{y}})^2 = \sum(y_i - \bar{y})^2 \quad \Rightarrow R^2 = 1$$

b) Si $\hat{y}_i = c \quad \forall i$, esto implica que

$$\bar{\hat{y}} = c \quad \Rightarrow \hat{y}_i - \bar{\hat{y}} = c - c = 0 \quad \forall i \quad \Rightarrow \sum(\hat{y}_i - \bar{\hat{y}})^2 = 0 \quad \Rightarrow R^2 = 0$$

Si R^2 está próximo a cero, esto implica que el ajuste no es bueno. En otras palabras, hay muy poca variación de *y* que sea explicada por *x*.

En muchos casos, se obtiene un R^2 elevado cuando se ajusta un modelo utilizando datos de series temporales, debido al efecto de una tendencia común. Por el contrario, cuando utilizamos datos de corte transversal es frecuente obtener valores bajos, pero esto no significa que el modelo ajustado sea malo.

¿Cuál es la relación entre el coeficiente de determinación y el coeficiente de correlación estudiados en estadística descriptiva? El coeficiente de determinación es igual al coeficiente de correlación al cuadrado, como puede verse en el apéndice 2.2:

$$r_{xy}^2 = R^2 \qquad (2\text{-}33)$$

(Esta igualdad es válida en el modelo de regresión lineal simple, pero no en el modelo de regresión lineal múltiple)

EJEMPLO 2.2 Cumplimiento de las propiedades algebraicas y R^2 en la función de consumo

En la columna 2 del cuadro 2.2, se calcula $cons_i$; en las columnas 3, 4 y 5, puede verse el cumplimiento de las implicaciones algebraicas 1, 3 y 4, respectivamente. En el resto de las columnas se realizan cálculos con el fin de obtener

$$SCT = 42 \quad SCE = 41.67 \quad SCR = 42 - 41.67 = 0.33 \quad R^2 = \frac{41.67}{42} = 0.992$$

o, alternativamente, $R^2 = 1 - \frac{0.33}{42} = 0.992$

CUADRO 2.2. Datos y cálculos para estimar la función de consumo.

Observ.	$cons_i$	$\hat{u}_i$	$\hat{u}_i \times renta_i$	$cons_i \times \hat{u}_i$	$cons_i^2$	$(cons_i - \overline{cons})^2$	$cons_i^{\,2}$	$(cons_i - \overline{cons})^2$
1	4.83	0.17	1.00	0.81	25	16	23.36	17.36
2	7.33	-0.33	-3.00	-2.44	49	4	53.78	2.78
3	8.17	-0.17	-1.67	-1.36	64	1	66.69	0.69
4	9.83	0.17	2.00	1.64	100	1	96.69	0.69
5	10.67	0.33	4.33	3.56	121	4	113.78	2.78
6	13.17	-0.17	-2.67	-2.19	169	16	173.36	17.36
	54.00	**0.00**	**0.00**	**0.00**	**528**	**42**	**527.67**	**41.67**

2.3.4 Regresión a través del origen

Si forzamos a que la línea de regresión pase por el punto (0,0) estamos imponiendo la restricción de que el término independiente sea cero, como puede verse en la figura 2.6. A esta regresión se le denomina regresión a través del origen.

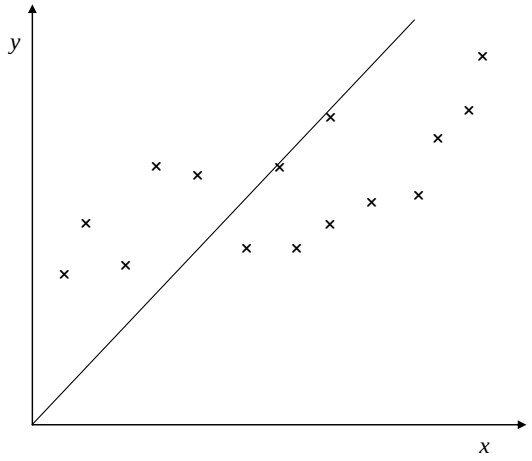


FIGURA 2.6. Una regresión a través del origen.

Ahora, vamos a estimar una recta de regresión a través del origen. El modelo ajustado es el siguiente:

$$\tilde{y}_i = \tilde{\beta}_2 x_i \tag{2-34}$$

Por lo tanto, debemos minimizar

$$\underset{\tilde{\beta}_2}{\text{Min}}\, S = \underset{\tilde{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i)^2 \tag{2-35}$$

Para minimizar *S*, derivamos con respecto a $\tilde{\beta}_2$ e igualaremos a 0:

$$\frac{dS}{d\tilde{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i) x_i = 0 \tag{2-36}$$

Resolviendo para $\tilde{\beta}_2$

$$\tilde{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \tag{2-37}$$

Otro problema que se plantea al ajustar una recta de regresión para que pase por el origen es que sucede en general que:

$$\sum (y_i - \bar{y})^2 \neq \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum \hat{u}_i^2$$

Si no es posible la descomposición de la varianza de *y* en dos componentes (explicada y residual), entonces R^2 no tiene sentido. Este coeficiente puede tomar valores negativos o superiores a 1 en el modelo sin término independiente.

Resumiendo, se debe incluir siempre un término independiente en las regresiones, a menos que haya fuertes razones en contra sustentadas por la teoría económica.

2.4 Las unidades de medida y la forma funcional

2.4.1 Unidades de medida

Cambio de unidades de medida (cambio de escala) en x

Si *x* es multiplicada/dividida por una constante $c \neq 0$, entonces la pendiente de *MCO* queda dividida/multiplicada por la misma constante, *c*. Así

$$\hat{y}_i = \hat{\beta}_1 + \left[\frac{\hat{\beta}_2}{c} \right] (x_i \times c) \tag{2-38}$$

EJEMPLO 2.3

Supongamos la siguiente función del consumo estimado, en la que ambas variables se miden en miles de euros:

$$cons_i = 0.2 + 0.85 \times renta_i \tag{2-39}$$

Si ahora se expresan la renta en euros (multiplicando por 1000) y se designa por *rentae*, el modelo ajustado a las nuevas unidades de medida de la renta será el siguiente:

$$cons_i = 0.2 + 0.00085 \times rentae_i$$

Como puede verse, el cambio de las unidades de medida de la variable explicativa no afecta al término independiente.

Cambio de unidades de medida (cambio de escala) en y

Si *y* es multiplicada/dividida por una constante $c \neq 0$, entonces la pendiente y el término independiente calculados por *MCO* se multiplican/dividen por la misma constante, *c*. Así,

$$(\hat{y}_i \times c) = (\hat{\beta}_1 \times c) + (\hat{\beta}_2 \times c) x_i \tag{2-40}$$

EJEMPLO 2.4

Si expresamos, en el modelo (2-39), el consumo en euros (multiplicando por 1000) y lo denominamos *conse*, el modelo ajustado a las nuevas unidades de medida del consumo será el siguiente:

$$conse_i = 200 + 850 \times inc_i$$

Cambio del origen

Si se suma/resta una constante *d* a *x* y/o *y*, entonces la pendiente *MCO* no se ve afectada. Sin embargo, si se cambia el origen de *x* y/o *y* el término independiente de la regresión sí se ve afectado.

Si se resta una constante *d* a *x*, el término independiente cambia de la siguiente manera:

$$\hat{y}_i = (\hat{\beta}_1 + \hat{\beta}_2 \times d) + \hat{\beta}_2 (x_i - d) \tag{2-41}$$

Si se resta una constante *d* a *y*, el término independiente cambia de la siguiente manera:

$$\hat{y}_i - d = (\hat{\beta}_1 - d) + \hat{\beta}_2 x_i \tag{2-42}$$

EJEMPLO 2.5

Supongamos que la renta media es de 20 mil euros. Si definimos la variable $rentad_i = renta_i - \overline{renta}$ y ambas variables se miden en miles de euros, el modelo ajustado con este cambio en el origen será el siguiente:

$$cons_i = (0.2 + 0.85 \times 20) + 0.85 \times (renta_i - 20) = 17.2 + 0.85 \times rentad_i$$

EJEMPLO 2.6

Supongamos que el consumo medio es de 15 mil euros. Si definimos la variable $consd_i = cons_i - \overline{cons}$ y medimos ambas variables en euros, el modelo ajustado con el cambio en el origen será el siguiente:

$$cons_i - 15 = 0.2 - 15 + 0.85 \times renta_i$$

Es decir,

$$consd_i = -14.8 + 0.85 \times renta_i$$

Hay que observar que R^2 no varía al realizar cambios de unidades de *x* y/o *y*, y tampoco varia cuando se cambia el origen de las variables.

2.4.2 Forma funcional

En muchos casos las relaciones lineales no son adecuadas en las aplicaciones económicas. Sin embargo, en el modelo de regresión simple podemos incorporar no linealidades (en las variables) redefiniendo de forma apropiada la variable dependiente y la variable independiente.

Algunas definiciones

Vamos a estudiar ahora algunas definiciones de las medidas de variación que serán útiles en la interpretación de los coeficientes de distintas formas funcionales. En concreto, vamos a estudiar las siguientes medidas: cambio proporcional y cambio en logaritmos.

El *cambio proporcional* (o tasa de variación relativa) entre x_1 y x_0 viene dado por:

$$\frac{\Delta x_1}{x_0} = \frac{x_1 - x_0}{x_0} \tag{2-43}$$

Multiplicando un *cambio proporcional* por 100 se obtiene un *cambio proporcional en %*. Es decir:

$$100\frac{\Delta x_1}{x_0}\% \tag{2-44}$$

El *cambio en logaritmos* y el *cambio en logaritmos en %* entre x_1 y x_0, vienen dados por

$$\begin{gathered}\Delta \ln(x) = \ln(x_1) - \ln(x_0) \\ 100\Delta \ln(x)\%\end{gathered} \tag{2-45}$$

El *cambio en logaritmos* es una aproximación del *cambio proporcional*, como puede verse en el apéndice 2.3. Esta aproximación es buena cuando la variación es pequeña, pero las diferencias pueden ser importantes cuando el *cambio proporcional* es grande, como puede observarse en el cuadro 2.3.

CUADRO 2.3. Ejemplos de cambios proporcionales y cambios en logaritmos.

x_1	202	210	220	240	300
x_0	200	200	200	200	200
Cambio proporcional en %	1%	5.0%	10.0%	20.0%	50.0%
Cambio en logaritmos en %	1%	4.9%	9.5%	18.2%	40.5%

La *elasticidad* es la razón entre los cambios relativos de dos variables. Si se utilizan cambios proporcionales, la elasticidad de la variable *y* con respecto a la variable *x* viene dada por

$$\varepsilon_{y/x} = \frac{\Delta y / y_0}{\Delta x / x_0} \tag{2-46}$$

Si se toman logaritmos se obtienen cambios infinitesimales, entonces, la elasticidad de la variable *y* con respecto a una variable *x* viene dada por

$$\varepsilon_{y/x} = \frac{dy / y}{dx / x} = \frac{d\ln(y)}{d\ln(x)} \tag{2-47}$$

En general, en los modelos econométricos, la elasticidad se define según (2-47).

Formas funcionales alternativas

El método *MCO* también se puede aplicar a modelos en los que se hayan transformado la variable endógena y/o la exógena. El modelo (2-1) nos muestra que la variable exógena y el regresor son términos equivalentes. Pero a partir de ahora, denominaremos regresor a la forma específica en la que una variable exógena aparece en la ecuación. Por ejemplo, en el modelo

$$y = \beta_1 + \beta_2 \ln(x) + u$$

la variable exógena es *x*, pero el regresor es ln(*x*).

El modelo de (2-1) también nos indica que la variable endógena y el regresando son equivalentes. Pero de ahora en adelante, denominaremos regresando a la forma

específica en la que una variable endógena aparece en la ecuación. Por ejemplo, en el modelo

$$\ln(y) = \beta_1 + \beta_2 x + u$$

la variable endógena es *y*, pero el regresando es ln(*y*).

Ambos modelos son lineales en los parámetros, aunque no son lineales en la variable *x* (el primero) o en la variable *y* (el segundo). En cualquier caso, si un modelo es lineal en los parámetros, se puede estimar aplicando el método de *MCO*. Por el contrario, si un modelo no es lineal en los parámetros, la estimación debe hacerse por métodos iterativos.

Sin embargo, existen ciertos modelos no lineales que, por medio de transformaciones adecuadas, pueden convertirse en lineales. Estos modelos son denominados linealizables.

Así, en algunas ocasiones se postulan modelos potenciales en la teoría económica, como es el caso de la conocida función de producción de Cobb-Douglas. Un modelo potencial con una única variable explicativa viene dado por

$$y = e^{\beta_1} x^{\beta_2}$$

Si se introduce el término de perturbación de forma multiplicativa se obtiene

$$y = e^{\beta_1} x^{\beta_2} e^{u} \tag{2-48}$$

Tomando logaritmos en ambos miembros de (2-48), se obtiene un modelo lineal en los parámetros:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + u \tag{2-49}$$

Por el contrario, si se introduce el término de perturbación de forma aditiva, se obtiene

$$y = e^{\beta_1} x^{\beta_2} + u$$

En este caso no existe una transformación que permita convertirlo en un modelo lineal. Será, por tanto, un modelo no linealizable.

Ahora, vamos a considerar algunos modelos con formas funcionales alternativas, pero todos ellos son lineales en los parámetros. Estudiaremos en cada caso la interpretación del coeficiente $\hat{\beta}_2$.

a) Modelo lineal

El coeficiente $\hat{\beta}_2$ mide el efecto del regresor *x* sobre *y*. Veamos esto con detalle. La observación *i* de la función de regresión muestral se expresa de acuerdo con (2-24) por

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-50}$$

Consideremos ahora la observación *h* del modelo ajustado en la cual el valor del regresor y, en consecuencia, del regresando han cambiado con respecto a (2-50):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_h \tag{2-51}$$

Si restamos (2-51) de (2-50), vemos que *x* tiene un efecto lineal sobre $\hat{y}$:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x \tag{2-52}$$

donde $\Delta\hat{y} = \hat{y}_i - \hat{y}_h$ y $\Delta x = x_i - x_h$

Por lo tanto, $\hat{\beta}_2$ es el cambio producido en *y* (en las unidades en qué esté medida *y*) al cambiar *x* en una unidad (en las unidades en qué esté medida *x*).

Por ejemplo, en la función ajustada (2-39), si la renta aumenta en una unidad, el consumo se incrementará en 0.85 unidades.

La linealidad de este modelo implica que un cambio de una unidad en *x* tiene siempre el mismo efecto en *y*, con independencia del valor de *x* considerado.

EJEMPLO 2.7 Cantidad de café vendido como una función de su precio. Modelo lineal

En un experimento de marketing[1] se formuló el siguiente modelo para explicar la cantidad de café vendido por semana (*coffqty*) en función del precio del café (*coffpric*).

$$coffqty = \beta_1 + \beta_2 coffpric + u$$

La variable *coffpric* toma el valor 1, el precio habitual, y también los valores 0.95 y 0.85 en dos acciones cuyos efectos están bajo investigación. El experimento duró 12 semanas, *coffqty* está expresado en miles de unidades y *coffpric* en francos franceses. Los datos aparecen en el cuadro 2.4 y en el fichero *coffee1*.

El modelo ajustado es el siguiente:

$$\widehat{coffqty} = 774.9 - 693.33 coffpric \quad R^2 = 0.95 \quad n = 12$$

Interpretación del coeficiente $\hat{\beta}_2$: si el precio del café se incrementa en 1 franco francés, la cantidad vendida de café se reducirá en 693.33 miles de unidades. En la medida que el precio del café es una magnitud pequeña, es preferible dar la siguiente interpretación: si aumenta el precio del café en 1 céntimo de franco francés, la cantidad vendida de café se reducirá en 6.93 miles de unidades.

CUADRO 2.4. Datos sobre cantidades y precios del café.

semana	*coffpric*	*coffqty*
1	1.00	89
2	1.00	86
3	1.00	74
4	1.00	79
5	1.00	68
6	1.00	84
7	0.95	139
8	0.95	122
9	0.95	102
10	0.85	186
11	0.85	179
12	0.85	187

[1]*Los datos de este ejercicio se han obtenido de un experimento controlado de marketing, sobre el gasto en café en tiendas de París. La referencia es* A. C. Bemmaor and D. Mouchoux, "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment". *Journal of Marketing Research*, 28 (1991), 202–14.

EJEMPLO 2.8 Explicando el valor de mercado de los bancos españoles. Modelo lineal

Utilizando datos de la Bolsa de Madrid (*Bolsa de Madrid*) del 18 de agosto de 1995 (fichero *bolmad95*, 20 primeras observaciones), se ha estimado el siguiente modelo para explicar el valor de mercado de bancos e instituciones financieras:

$$marktval = 29.42 + 1.219 bookval$$

$$R^2=0.836 \qquad n=20$$

donde

- *marktval* es el valor en mercado de una empresa. Se calcula multiplicando el precio de la acción por el número de acciones emitidas.
- *bookval* es el valor contable o el valor neto de la compañía. El valor contable se calcula como la diferencia entre los activos de una empresa y sus pasivos.
- Los datos de *marktval* y *bookval* están expresados en millones de pesetas.

Interpretación del coeficiente β_2: si el valor contable de un banco se incrementa en 1 millón de pesetas, la capitalización de mercado de ese banco se incrementará en 1.219 millones de pesetas.

b) Modelo lineal logarítmico

Un modelo lineal logarítmico se expresa como

$$y = \beta_1 + \beta_2 \ln(x) + u \tag{2-53}$$

La función ajustada correspondiente es la siguiente:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln(x) \tag{2-54}$$

Tomando primeras diferencias en (2-54), y multiplicando y dividiendo el segundo miembro por 100, se tiene

$$\Delta\hat{y} = \frac{\hat{\beta}_2}{100} 100 \times \Delta \ln(x)\%$$

Por lo tanto, si x aumenta un 1%, $\hat{y}$ se incrementará en $(\hat{\beta}_2 / 100)$ unidades.

c) Modelo logarítmico lineal

Un modelo logarítmico lineal se expresa como

$$\ln(y) = \beta_1 + \beta_2 x + u \tag{2-55}$$

El modelo anterior se obtiene tomando logaritmos naturales en ambos miembros del siguiente modelo:

$$y = \exp(\beta_1 + \beta_2 x + u)$$

Por esta razón, el modelo (2-55) también se llama también exponencial.

La función de regresión muestral correspondiente a (2-55) es la siguiente

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 x \tag{2-56}$$

Tomando las primeras diferencias en (2-56), y multiplicando ambos miembros por 100, se tiene

$$100 \times \Delta \ln(y)\% \; = \; 100 \times \hat{\beta}_2 \Delta x$$

Por lo tanto, si *x* aumenta en una unidad, entonces $\hat{y}$ se incrementará un $100\times\hat{\beta}_2$%.

d) Modelo doblemente logarítmico

El modelo que figura en (2-49) es un modelo doblemente logarítmico o, antes de la transformación, un modelo potencial (2-48). A este modelo también se le denomina modelo de elasticidad constante.

El modelo ajustado correspondiente a (2-49) es el siguiente:

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 \ln(x) \tag{2-57}$$

Tomando primeras diferencias en (2-57), se tiene

$$\Delta\ln(y) = \hat{\beta}_2 \Delta\ln(x)$$

Por lo tanto, si *x* aumenta en 1%, entonces $\hat{y}$ se incrementará un $\hat{\beta}_2$ %. Hay que resaltar que, en este modelo, $\hat{\beta}_2$ es la elasticidad estimada de *y* con respecto a *x*, para cualquier valor de *x* e *y*. En consecuencia, en este modelo la elasticidad es constante.

En el anexo 1 en un caso de estudio de la curva de Engel para la demanda de productos lácteos se analizan seis formas funcionales alternativas.

EJEMPLO 2.9 Cantidad de café vendido en función de su precio. Modelo doblemente logarítmico (Continuación del ejemplo 2.7)

Como una alternativa al modelo lineal se ha estimado el modelo doblemente logarítmico:

$$\ln(coffqty) = 4.415 - 5.132\ln(coffpric) \quad R^2 = 0.90 \quad n = 12$$

Interpretación del coeficiente $\hat{\beta}_2$: si el precio del café aumenta en un 1%, la cantidad vendida de café se reducirá en un 5,13%. En este caso, $\hat{\beta}_2$ es el estimador de la elasticidad de la demanda/precio.

EJEMPLO 2.10 Explicando el valor de mercado de los bancos españoles. Modelo doblemente logarítmico (Continuación del ejemplo 2.8)

Utilizando datos del ejemplo 2.8, se ha estimado el siguiente modelo doblemente logarítmico:

$$\ln(marktval) = 0.6756 + 0.938\ln(bookval)$$

$$R^2=0.928 \qquad n=20$$

Interpretación del coeficiente $\hat{\beta}_2$: si el valor contable de un banco se incrementa en 1%, el valor de mercado de ese banco se incrementará en un 0.938%. En este caso $\hat{\beta}_2$ es el estimador de la elasticidad del valor de mercado/valor contable.

En el cuadro 2.5 se muestra, para el modelo ajustado, la interpretación de los cuatro modelos estudiados. Si hubiéramos considerado el modelo poblacional en lugar del muestral, la interpretación de β_2 es la misma pero teniendo en cuenta que Δu debería ser igual a 0.

CUADRO 2.5. Interpretación de $\hat{\beta}_2$ en los diferentes modelos.

Modelo	Si *x* aumenta en	entonces *y* se incrementará en
lineal	1 unidad	$\hat{\beta}_2$ unidades
lineal logarítmico	1%	$(\hat{\beta}_2 / 100)$ unidades
logarítmico lineal	1 unidad	$(100\hat{\beta}_2)\%$
doblemente logarítmico	1%	$\hat{\beta}_2\%$

2.5 Supuestos y propiedades estadísticas de los *MCO*

Vamos ahora a estudiar las propiedades estadísticas de los estimadores de *MCO*, $\hat{\beta}_1$ *y* $\hat{\beta}_2$, del modelo de regresión lineal simple. Previamente, es necesario formular un conjunto de supuestos estadísticos. Específicamente, al conjunto de supuestos que vamos a formular se les denomina *supuestos del modelo lineal cásico* (*MLC*). Es de resaltar que los *supuestos* del *MLC* son sencillos, y que los estimadores *MCO* tienen, bajo estos supuestos, muy buenas propiedades.

2.5.1 Supuestos estadísticos del *MLC* en regresión lineal simple

a) Supuesto sobre la forma funcional

1) La relación entre el regresando, regresor y perturbación aleatoria es lineal en los parámetros:

$$y = \beta_1 + \beta_2 x + u \tag{2-58}$$

El regresando y los regresores pueden ser cualquier función de la variable endógena y de las variables explicativas, respectivamente, a condición de que entre los regresores y el regresando exista una relación lineal. Es decir, el modelo es lineal en los parámetros. La aditividad de la perturbación garantiza la relación lineal con el resto de los elementos.

b) Supuestos sobre el regresor x

2) Los valores que toma x son fijos en repetidas muestras:

De acuerdo con este supuesto, cada observación del regresor toma el mismo valor para diferentes muestras del regresando. Este es un supuesto fuerte en el caso de las ciencias sociales, donde, en general, no es posible la experimentación. Los datos se obtienen mediante observación, no mediante experimentación. Es importante destacar que los resultados obtenidos basados en este supuesto permanecen virtualmente idénticos a los que se obtienen cuando asumimos que los regresores son estocásticos, siempre, que postulemos el supuesto adicional de independencia entre los regresores y la perturbación aleatoria. Este supuesto alternativo se puede formular así:

2) El regresor x se distribuye de forma independiente de la perturbación aleatoria.*

En cualquier caso, a lo largo de este capítulo y los siguientes vamos a adoptar el supuesto 2.

3) El regresor x no contiene errores de medición

Se trata de un supuesto que a menudo no se cumple en la práctica, ya que los instrumentos de medición no son siempre fiables en la economía. Pensemos, por ejemplo, en la multitud de errores que se pueden cometer en la recopilación de información cuando se realizan encuestas a las familias.

4) La varianza muestral de x es distinta de 0 y tiene un límite finito cuando n tiende a infinito

Por lo tanto, este supuesto implica que

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n} \neq 0 \qquad (2\text{-}59)$$

c) Supuesto sobre los parámetros

5) Los parámetros β_1 y β_2 son fijos

Si no se adopta este supuesto, el modelo de regresión sería muy difícil de aplicar. En cualquier caso, puede ser aceptable postular que los parámetros del modelo son estables en el tiempo (si no es un período muy largo) o en el espacio (si es relativamente limitado).

d) Supuestos sobre las perturbaciones aleatorias

6) La esperanza de las perturbaciones es cero,

$$E(u_i) = 0, \quad i = 1, 2, 3, \ldots, n \qquad (2\text{-}60)$$

Éste no es un supuesto restrictivo, ya que siempre se puede utilizar β_1 para normalizar $E(u)$ a 0. Supongamos, por ejemplo, que $E(u) = 4$, entonces podríamos redefinir el modelo del siguiente modo:

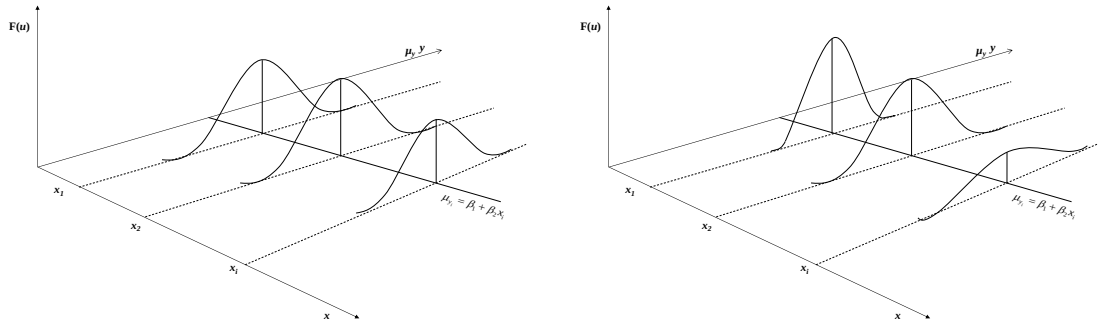
$$y = (\beta_1 + 4) + \beta_2 x + v$$

dónde $v = u - 4$. Por lo tanto, la esperanza de la nueva perturbación, *v*, es 0 y la esperanza de *u* ha sido absorbida por el término independiente.

7) Las perturbaciones tienen una varianza constante

$$var(u_i) = \sigma^2 \quad i = 1, 2, \cdots n \qquad (2\text{-}61)$$

A este supuesto se le denomina supuesto de *homoscedasticidad*. Esta palabra viene del griego: *homo* (igual) y *scedasticidad* (variabilidad). Esto significa que la variabilidad en torno a la línea de regresión es la misma en toda la muestra de *x*; es decir, que no aumenta o disminuye cuando *x* varía, como puede verse en la figura 2.7, parte a), donde las perturbaciones son homoscedásticas.



a) **b)**

FIGURA 2.7. Perturbaciones aleatorias: a) homoscedasticidad; b) heteroscedasticidad.

Si este supuesto no se cumple, como ocurre en la parte b) de la figura 2.7, los estimadores de *MCO* no son eficientes. Las perturbaciones en ese caso se dice que son heteroscedásticas (*hetero* significa distinta).

8) Las perturbaciones con diferentes subíndices no están correlacionadas entre sí (supuesto de no autocorrelación):

$$E(u_i u_j) = 0 \qquad i \neq j \tag{2-62}$$

Es decir, las perturbaciones correspondientes a diferentes individuos o a diferentes momentos de tiempo, no están correlacionadas entre sí. Este supuesto de no autocorrelación o no correlación serial, al igual que en el caso de homoscedasticidad, es contrastable *a posteriori*. La transgresión de este supuesto se produce con bastante frecuencia en los modelos que utilizan datos de series temporales.

9) Las perturbaciones se distribuyen normalmente

Teniendo en cuenta los supuestos 6, 7 y 8 se tiene que

$$u_i \sim NID(0, \sigma^2) \qquad i = 1, 2, \cdots, n \tag{2-63}$$

donde *NID* indica que las perturbaciones están *n*ormal e *i*ndependientemente *d*istribuidas.

La razón de este supuesto es que si *u* se distribuye normalmente, también lo harán *y* y los coeficientes estimados de regresión, lo cual es útil en la realización de contrastes de hipótesis y en la construcción de intervalos de confianza para β_1 y β_2. La justificación de este supuesto se basa en el Teorema Central del Límite. En esencia, este teorema indica que, si una variable aleatoria es el resultado agregado de los efectos de un número indefinido de variables, tendrá una distribución aproximadamente normal, incluso si sus componentes no la tienen, a condición de que ninguno de ellos sea dominante.

2.5.2 Propiedades deseables de los estimadores

Antes de examinar las propiedades de los estimadores mínimo-cuadráticos bajo los supuestos estadísticos del *MLC*, se puede plantear la siguiente cuestión previa: ¿cuáles son las propiedades deseables para un estimador?

Dos propiedades deseables para un estimador es que sea insesgado y que su varianza sea lo más pequeña posible. Si esto sucede el proceso de inferencia se podrá llevar a cabo de una forma satisfactoria.

Vamos a ilustrar estas propiedades de forma gráfica. Consideremos en primer lugar la propiedad de insesgadez. En las figuras 2.8 y 2.9 se han representado las funciones de densidad de dos hipotéticos estimadores obtenidos por dos procedimientos diferentes:

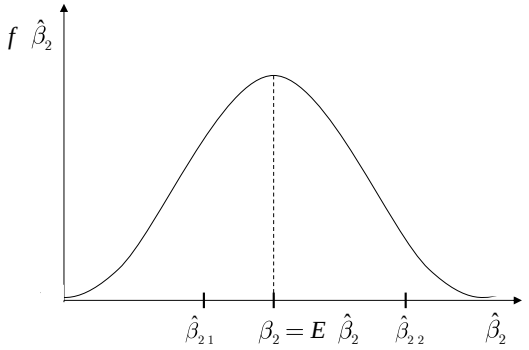


FIGURA 2.8. Estimador insesgado.

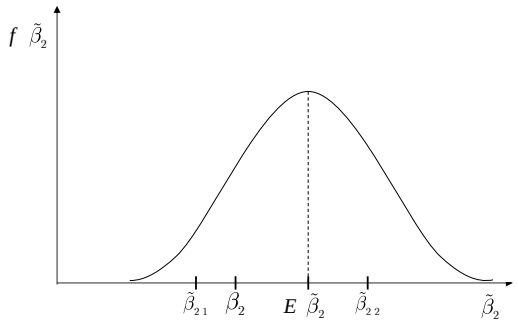


FIGURA 2.9. Estimador sesgado.

El estimador $\hat{\beta}_2$ es insesgado, es decir, su esperanza matemática es igual al parámetro que trata de estimar, β_2. El estimador $\hat{\beta}_2$ es una variable aleatoria que en cada muestra de *y*– las *x* son fijas en repetidas muestra según el supuesto 2- toma un valor diferente, pero en *promedio*, es decir, teniendo en cuenta los infinitos valores que puede tomar $\hat{\beta}_2$, es igual al parámetro β_2. Con cada muestra de *y* se obtiene un valor específico de $\hat{\beta}_2$, es decir, una *estimación*. En la figura 2.8 se han representado dos estimaciones de β_2: $\hat{\beta}_{2(1)}$ y $\hat{\beta}_{2(2)}$. La primera estimación está relativamente cerca de β_2, mientras que la segunda está mucho más alejada. En todo caso, la insesgadez es una propiedad deseable, ya que nos asegura que el estimador en promedio está centrado sobre el parámetro.

El estimador $\tilde{\beta}_2$, en la figura 2.9, es sesgado, ya que su esperanza no es igual a β_2. El sesgo es precisamente $E\ \tilde{\beta}_2\ -\beta_2$. En este caso también se han representado dos hipotéticas estimaciones: $\tilde{\beta}_{2(1)}$ y $\tilde{\beta}_{2(2)}$. Como puede verse $\tilde{\beta}_{2(1)}$ está más cerca de β_2 que el estimador insesgado $\hat{\beta}_{2(1)}$: es una cuestión de azar. En todo caso, por ser sesgado no está centrado en promedio sobre el parámetro. No cabe duda que siempre es preferible un estimador insesgado puesto que, con independencia de lo que ocurra en una muestra concreta, no tiene una desviación sistemática respecto al valor del parámetro.

La otra propiedad deseable es la eficiencia. Esta propiedad hace referencia a la varianza de los estimadores. En las figuras 2.10 y 2.11 se han representado dos hipotéticos estimadores insesgados a los que seguiremos llamando $\hat{\beta}_2$ y $\tilde{\beta}_2$. El primero de ellos tiene una varianza más pequeña que el segundo.

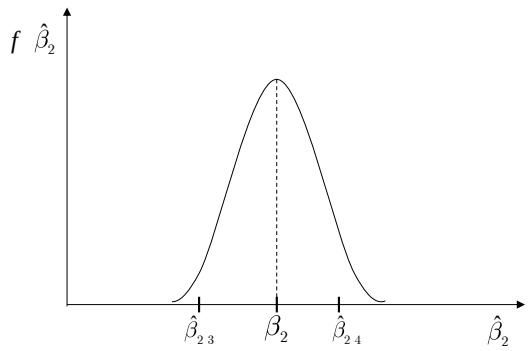


FIGURA 2.10. Estimador con varianza pequeña.

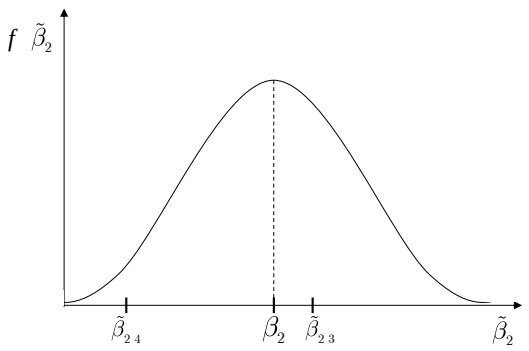


FIGURA 2.11. Estimador con una varianza grande.

En ambas figuras hemos representado dos estimaciones: $\hat{\beta}_{2(3)}$ y $\hat{\beta}_{2(4)}$ en el estimador con varianza más pequeña; $\tilde{\beta}_{2(3)}$ y $\tilde{\beta}_{2(4)}$ en el estimador con varianza más grande. También aquí, para resaltar el papel jugado por el azar, la estimación que está más cerca de β_2 es precisamente $\tilde{\beta}_{2(3)}$. En cualquier caso, siempre es preferible que la varianza del estimador sea lo más pequeña posible. Así por ejemplo, utilizando el estimador $\hat{\beta}_2$ es prácticamente imposible que una estimación esté tan alejada de β_2 como lo está $\tilde{\beta}_{2(4)}$, debido a que el recorrido de $\hat{\beta}_2$ es mucho más reducido que el que tiene $\tilde{\beta}_2$.

2.5.3 Propiedades estadísticas de los estimadores *MCO*

Bajo los supuestos anteriores, los estimadores *MCO* poseen algunas propiedades ideales. Así, podemos decir que los *MCO* son estimadores lineales insesgados y óptimos.

Linealidad e insesgadez de los MCO

El estimador $\hat{\beta}_2$ de *MCO* es insesgado. En el apéndice 2.4 se demuestra que es un estimador insesgado utilizando implícitamente los supuestos 3, 4 y 5, y explícitamente los supuestos 1, 2 y 6. En dicho anexo también se puede ver que es un estimador lineal, utilizando los supuestos 1 y 2. Del mismo modo, se puede demostrar que el estimador *MCO* $\hat{\beta}_1$ es insesgado.

Recordemos que la insesgadez es una propiedad general del estimador, pero que para una muestra determinada la estimación puede estar más "cerca" o más "lejos" del verdadero parámetro. En cualquier caso, la distribución del estimador está centrada en el parámetro poblacional.

Varianzas de los estimadores de MCO

Ahora sabemos que la distribución muestral de nuestro estimador está centrada en el parámetro poblacional, pero ¿cuál es la dispersión de su distribución? La varianza, que es una medida de dispersión, de un estimador es un indicador de la precisión de ese estimador.

Para obtener las varianzas de $\hat{\beta}_1$ y $\hat{\beta}_2$ se requieren los supuestos 7 y 8, además de los seis primeros. Estas varianzas son las siguientes:

$$Var(\hat{\beta}_1) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-64}$$

En el apéndice 2.5 se muestra cómo se obtiene la varianza de $\hat{\beta}_2$.

Los estimadores de MCO son ELIO

Los estimadores de *MCO* tienen la menor varianza de entre todos los estimadores lineales e insesgados. Por esta razón se dice que los estimadores de *MCO* son *estimadores lineales insesgados y óptimos (ELIO)*, como se ilustra en la figura 2.12. Esta propiedad se conoce como el teorema de Gauss-Markov. Para la demostración de este teorema se utilizan los supuestos 1 a 8, como puede verse en el apéndice 2.6. Este conjunto de supuestos se conoce como los supuestos de Gauss-Markov.

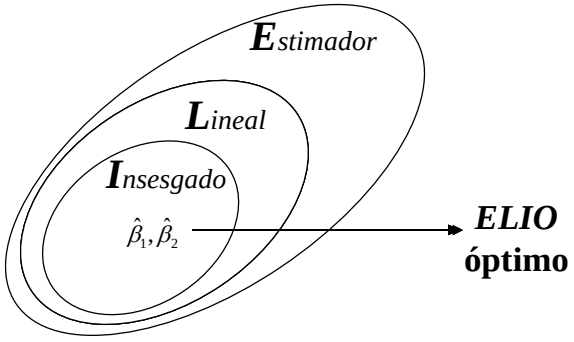


FIGURA 2.12. Los estimadores *MCO* son *ELIO*.

La estimación de la varianza de las perturbaciones y de la varianza de los estimadores

Dado que no conocemos el valor de la varianza de la perturbación, σ^2, tenemos que estimarlo. Sin embargo, no podemos estimarlo utilizando los valores muestrales de las perturbaciones u_i porque no son observables En su lugar, tenemos que utilizar los residuos de *MCO* ($\hat{u}_i$).

La relación entre las perturbaciones y los residuos viene dada por

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + u_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\ &= u_i - \left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i \end{aligned} \tag{2-65}$$

Por tanto, $\hat{u}_i$ no es lo mismo que u_i, aunque la diferencia entre ellos - $\left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i$ - tiene un valor esperado que es igual a cero. Por ello, un primer estimador de σ^2 podría ser la varianza residual:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n} \tag{2-66}$$

Sin embargo, este estimador es sesgado, esencialmente porque no tiene en cuenta las dos siguientes restricciones que deben ser satisfechas por los residuos de *MCO* en el modelo de regresión lineal simple:

$$\begin{cases} \sum_{i=1}^{n} \hat{u}_i = 0 \\ \sum_{i=1}^{n} x_i \hat{u}_i = 0 \end{cases} \tag{2-67}$$

Una forma de ver estas restricciones es la siguiente: si conocemos *n*-2 de los residuos, podemos obtener los otros dos residuos mediante el uso de las restricciones implícitas en las ecuaciones normales (2-67).

Por lo tanto, sólo hay *n*-2 grados de libertad en los residuos de *MCO*, a diferencia de los *n* grados de libertad que tendrían las correspondientes *n* perturbaciones. En el estimador insesgado de σ^2 mostrado a continuación se realiza un ajuste en el que se tiene en cuenta los grados de libertad:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \tag{2-68}$$

Bajo los supuestos 1-8 (supuestos Gauss-Markov), se obtiene, como puede verse en el apéndice 2.7, que

$$E(\hat{\sigma}^2) = \sigma^2 \tag{2-69}$$

Si $\hat{\sigma}^2$ se introduce en las fórmulas de la varianza obtenemos entonces los estimadores insesgados de var($\hat{\beta}_1$) y var($\hat{\beta}_2$)

El estimador natural de σ es $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ y se llama *error estándar de la regresión*. La raíz cuadrada de la varianza se denomina *desviación estándar* de $\hat{\beta}_2$, es decir,

$$de(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2-70}$$

Por lo tanto, su estimador natural, al que se denomina *error estándar* de $\hat{\beta}_2$, viene dado por

$$ee(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2-71}$$

Nótese que $ee(\hat{\beta}_2)$, debido a la presencia del estimador $\hat{\sigma}$ en (2-71), es una variable aleatoria igual que $\hat{\beta}_2$. El error estándar de una estimación nos ofrece una idea de lo preciso que es el estimador.

La consistencia de los MCO y otras propiedades asintóticas

A veces no es posible obtener un estimador insesgado. Entonces, se considera que la *consistencia* es el requisito mínimo que debe cumplir el estimador. Según un enfoque intuitivo, *consistencia* significa que a medida que $n \to \infty$, la función de

densidad del estimador converge al valor del parámetro. Esta propiedad puede expresarse para el estimador $\hat{\beta}_2$ como:

$$\operatorname{plim}_{n\to\infty} \hat{\beta}_2 = \beta_2 \tag{2-72}$$

donde plim es el límite en probabilidad. En otras palabras, $\hat{\beta}_2$ converge en probabilidad a β_2.

Es importante tener en mente que las propiedades de insesgadez y consistencia son conceptualmente diferentes. La propiedad de insesgadez se mantiene para cualquier tamaño muestral, mientras que la consistencia es una propiedad estrictamente de grandes muestras o, de forma más precisa, es una *propiedad asintótica*.

Bajo los supuestos 1 a 6, los estimadores *MCO*, $\hat{\beta}_1$ y $\hat{\beta}_2$ son consistentes. La demostración de la consistencia de $\hat{\beta}_2$ puede verse en el apéndice 2.8.

Otras propiedades asintóticas de $\hat{\beta}_1$ y $\hat{\beta}_2$: Bajo los supuestos de Gauss-Markov 1 a 8, $\hat{\beta}_1$ y $\hat{\beta}_2$ tienen *una distribución asintóticamente normal* y es asintóticamente eficiente dentro de la clase de estimadores consistentes y asintóticamente normales.

Los estimadores MCO son estimadores de máxima verosimilitud (MV) y estimadores insesgados de mínima varianza (EIMV)

Ahora vamos a introducir el supuesto 9 en la normalidad de las perturbaciones *u*. El conjunto de supuestos 1 a 9 se conocen como los supuestos del *modelo lineal clásico* (*MLC*)

Bajo los supuestos del *MLC*, los estimadores de *MCO* son también estimadores de *máxima verosimilitud* (*MV*), como puede verse en el apéndice 2.8.

Por otro lado, bajo los supuestos del *MLC*, los estimadores de *MCO* además de ser *ELIO*, son *estimadores insesgados* de *mínima varianza* (*EIMV*). Esto significa que los estimadores de *MCO* tienen la varianza más pequeña entre todos los estimadores insesgados, lineales o no lineales, según se ilustra en la figura 2.13. Por lo tanto, ya no tenemos que restringirnos a los estimadores que son lineales en y_i.

También se cumple que cualquier combinación lineal de $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_k$ se distribuye normalmente, y cualquier subconjunto de las $\hat{\beta}_j$ tiene una distribución normal conjunta.

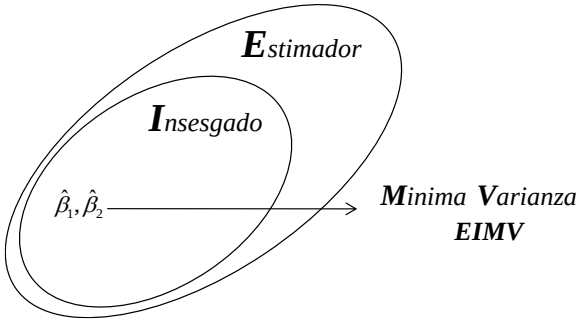


FIGURA 2.13. Los estimadores MCO son EIMV.

En resumen, hemos visto que los estimadores de *MCO* tienen propiedades muy deseables cuando se cumplen los supuestos estadísticos del *MLC*.

Ejercicios

Ejercicio 2.1 El siguiente modelo ha sido formulado para explicar las ventas anuales (*ventas*) de empresas fabricantes de productos de limpieza doméstica en función de un índice de precios relativo (*ipr*):

$$ventas = \beta_1 + \beta_2 ipr + u$$

donde la variable *ventas* está expresada en millones de euros e *ipr* es un índice de precios relativos (precios de la empresa/precios de la empresa 1 de la muestra). Así, el valor 110 de la empresa 2 indica que su precio es un 10% mal elevado que en la empresa 1.

Para ello se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

empresa	*ventas*	*ipr*
1	10	100
2	8	110
3	7	130
4	6	100
5	13	80
6	6	80
7	12	90
8	7	120
9	9	120
10	15	90

a) Estime β_1 y β_2 por *MCO*.
b) Obtenga la suma de los cuadrados de los residuos.
c) Calcule el coeficiente de determinación.
d) Compruebe si se cumplen las implicaciones algebraicas 1, 3 y 4 en la estimación por *MCO*.

Ejercicio 2.2 Para estudiar la relación entre consumo de combustible (*y*) y el tiempo de vuelo (*x*) en una compañía aérea se ha formulado el siguiente modelo:

$$y = \beta_1 + \beta_2 x + u$$

donde *y* está expresado en miles de libras y *x* en horas, utilizándose como unidades de orden inferior fracciones decimales de la hora.

De las estadísticas de «Tiempos de vuelo y consumos de combustible» de una compañía aérea se han obtenido datos relativos a tiempos de vuelo y consumos de combustible de 24 trayectos distintos realizados por aviones DC-9. A partir de estos datos se han elaborado los siguientes estadísticos:

$$\sum y_i = 219.719;\ \sum x_i = 31.470;\ \sum x_i^2 = 51.075;$$
$$\sum x_i y_i = 349.486;\ \sum y_i^2 = 2396.504$$

Se pide

a) La estimación de β_1 y β_2.

b) La descomposición de la varianza de *y* en varianza explicada por la regresión y varianza residual.

c) El coeficiente de determinación.

d) ¿Qué consumo total estimaría, en miles de libras, para un programa de vuelos compuesto por 100 vuelos de media hora, 200 de una hora y 100 de dos horas?

Ejercicio 2.3 Un analista formula el siguiente modelo:

$$y = \beta_1 + \beta_2 x + u$$

Utilizando una muestra dada, se estima el modelo obteniendo los siguientes resultados:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = 20 \qquad \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = 10 \qquad \overline{y} = 8 \quad \overline{x} = 4 \quad \hat{\beta}_2 = 3$$

¿Le parecen coherentes los resultados obtenidos por el analista?

Ejercicio 2.4 Un económetra ha estimado el siguiente modelo con una muestra de cinco observaciones:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Una vez realizada la estimación el económetra pierde toda la información excepto la que aparece en el siguiente cuadro:

Obs.	x_i	$\hat{u}_t$
1	1	2
2	3	-3
3	4	0
4	5	¿?
5	6	¿?

Con la información anterior el económetra debe calcular la varianza residual. Hágalo en su lugar.

Ejercicio 2.5 Sea el siguiente modelo

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad 1 = 1, 2, \ldots, n$$

Al estimar este modelo con una muestra de tamaño 11 se han obtenido los siguientes resultados:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

a) Obtenga la estimación de β_2 y β_1.
b) Obtenga la suma de cuadrados de los residuos.
c) Calcule el coeficiente de determinación.
d) Calcule el coeficiente de determinación bajo el supuesto de que $2F^2 = BE$.

Ejercicio 2.6 La empresa A se dedica a montar paneles prefabricados para naves industriales. Hasta el momento ha realizado 8 obras, en las cuales el número de metros cuadrados de paneles y el de horas de trabajo directamente empleadas en el montaje han sido los siguientes:

Nº de metros cuadrados (miles)	**Nº de horas**
4	7400
6	9800
2	4600
8	12200
10	14000
5	8200
3	5800
12	17000

La empresa A desea participar en un concurso para montar 14000 m^2 de panel para una nave industrial, para lo cual tiene que presentar un presupuesto.

Como datos a tener en cuenta en la elaboración del presupuesto, se conocen los siguientes:

a) El presupuesto debe referirse exclusivamente a los costes de montaje, ya que el material lo proporciona la empresa que ha convocado el concurso.

b) El coste por hora trabajada para la empresa A es de 1100 pesetas.

c) Para cubrir los restantes costes, la empresa A debe cargar un 20% sobre el importe total del coste de mano de obra empleada en el montaje.

Por la situación en que se encuentra, a la empresa A le interesa participar en el concurso con un presupuesto en el que únicamente se cubran los costes. En estas condiciones, y bajo el supuesto de que el número de horas trabajadas es función lineal del número de metros cuadrados de paneles montados, ¿cuál debería ser el importe del presupuesto de la empresa A?

Ejercicio 2.7 Considere las siguientes igualdades:

1. $E[u] = 0$.
2. $E[\hat{u}] = 0$.
3. $\bar{u} = 0$.
4. $\bar{\hat{u}} = 0$.

En el contexto del modelo lineal, indique si cada una de las anteriores igualdades se cumple o no, razonando la respuesta.

Ejercicio 2.8 Se han estimado por mínimos cuadrados ordinarios los parámetros β_1 y β_2 del modelo

$$y = \beta_1 + \beta_2 x + u$$

con una muestra de tamaño 3.

Los valores de x_i son {1,2,3}. Se sabe también que el residuo correspondiente a la primera observación es de 0.5.

A partir de la anterior información, ¿es posible calcular la suma de los cuadrados de los residuos y obtener una estimación de σ^2? En caso afirmativo, realice los correspondientes cálculos.

Ejercicio 2.9 Se tienen los siguientes datos, para estimar una relación entre *y* y *x*:

y	*x*
-2	-2
-1	0
0	1
1	0
2	1

a) Estime por *MCO* los parámetros α y β del siguiente modelo:

$$y = \alpha + \beta x + \varepsilon$$

b) Estime var(ε_i).

c) Por otra parte, estime por *MCO* los parámetros γ y δ del siguiente modelo:

$$x = \gamma + \delta y + \upsilon$$

d) ¿Son las dos líneas de regresión ajustadas iguales? Explique el resultado en términos de la metodología mínimo-cuadrática.

Ejercicio 2.10 Responda a las siguientes preguntas:

a) Un investigador, después de realizar la estimación de un modelo por *MCO*, calcula $\sum \hat{u}_i$ y comprueba que no es 0. ¿Es esto posible? Razone la respuesta indicando en su caso las condiciones en las cuales puede haberse producido este hecho.

b) Obtenga un estimador insesgado de σ^2, indicando los supuestos utilizados. Razone la respuesta.

Ejercicio 2.11 En el contexto del modelo de regresión lineal

$$y = \beta_1 + \beta_2 x + u$$

a) Indique en que se basa el cumplimiento, en su caso, de las siguientes igualdades

$$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = 0; \qquad \bar{\hat{u}} = \frac{\sum_{i=1}^{n} \hat{u}_i}{n} = 0; \qquad E\left[x_i u_i\right] = 0; \qquad E\left[u_i\right] = 0;$$

b) Establezca la relación entre las dos expresiones siguientes:

$$E\left[u_i^2\right] = \sigma^2; \qquad \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$$

Ejercicio 2.12 Responda a las siguientes preguntas:

a) Defina las propiedades probabilísticas de los estimadores de *MCO* bajo los supuestos del *MLC*. Razone la respuesta.

b) ¿Qué sucede con la estimación del modelo de regresión lineal si la varianza muestral de la variable explicativa es nula? Razone su respuesta.

Ejercicio 2.13 Un investigador considera que la relación entre consumo (*cons*) y renta disponible (*renta*) debe ser estrictamente proporcional. Por ello, plantea el siguiente modelo:

$$cons=\beta_2 renta+u$$

a) Deduzca la fórmula para estimar β_2.
b) Deduzca la fórmula para estimar σ^2.
c) En este modelo, ¿a qué es igual $\sum_{i=1}^{n}\hat{u}_i$?

Ejercicio 2.14 En el contexto del modelo de regresión lineal simple

$$y=\beta_1+\beta_2 x+u$$

a) ¿Qué supuestos deben cumplirse para que los estimadores de mínimos cuadrados ordinarios sean insesgados?
b) ¿Qué supuestos se requieren para que su varianza sea mínima dentro del conjunto de estimadores lineales e insesgados?

Ejercicio 2.15 En lenguaje estadístico se suelen hacer en muchas ocasiones afirmaciones como la siguiente:

"Sea una muestra aleatoria de tamaño *n* extraída de una variable *x* con distribución normal $N(\alpha,\sigma)$".

a) Exprese la afirmación anterior con lenguaje econométrico, introduciendo un término de perturbación.
b) Deduzca la fórmula para estimar α.
c) Deduzca la fórmula para estimar σ^2.
d) En este modelo, ¿a qué sería igual $\sum_{i=1}^{n}\hat{u}_i$?

Ejercicio 2.16 Sea el siguiente modelo que relaciona el gasto en educación (*geduc*) con la renta disponible (*renta*):

$$geduc=\beta_1+\beta_2 renta+u$$

Utilizando la información obtenida de una muestra de 10 familias se han obtenido los siguientes resultados:

$$\overline{geduc}=7 \quad \overline{renta}=50 \quad \sum_{i=1}^{10} renta_i^2=30.650 \quad \sum_{i=1}^{10} geduc_i^2=622 \quad \sum_{i=1}^{10} renta_i\times geduc_i=4.345$$

a) Estime β_1 y β_2 por *MCO*.
b) Estime la elasticidad gasto en educación/renta para el promedio de las familias de la muestra.
c) Descomponga la varianza total del gasto en educación de la muestra en varianza explicada y varianza residual.
d) Calcule el coeficiente de determinación.
e) Estime la varianza de las perturbaciones

Ejercicio 2.17 Dado el modelo poblacional

$$y_i=3+2x_i+u_i \qquad i=1,\ 2,\ 3$$

y siendo los valores de $x_i=\{1,2,3\}$:

a) Genere 15 muestras de u_1, u_2 y u_3, y obtenga los correspondientes valores de *y*, utilizando los números aleatorios $N(0,1)$.
b) Realice las correspondientes estimaciones de β_1 y β_2 en el modelo:

$$y = \beta_1 + \beta_2 x + u$$

c) Compare las medias y varianzas muestrales de $\hat{\beta}_1$ y $\hat{\beta}_2$ con sus esperanzas y varianzas poblacionales.

Ejercicio 2.18 Basándose en la información suministrada en el ejercicio 2.17, y con las distintas estimaciones de β_1 y β_2 obtenidas:

a) Calcule los residuos correspondientes a cada una de las estimaciones.
b) Explique el motivo por el cual los residuos adoptan siempre la forma

$$\hat{u}_1 = -\hat{u}_2$$

$$\hat{u}_3 = 0$$

Ejercicio 2.19 El siguiente modelo se formuló para explicar el tiempo dedicado a dormir (*sleep*) en función del tiempo dedicado al trabajo remunerado (*paidwork*):

$$sleep = \beta_1 + \beta_2 paidwork + u$$

donde el *sleep* y la *paidwork* se miden en minutos por día.

Usando una sub-muestra aleatoria, extraída del archivo *timuse03*, fueron obtenidos los siguientes resultados:

$$\widehat{sleep}_i = 550.17 - 0.1783\, paidwork$$

$$R^2 = 0.2539 \qquad n = 62$$

a) Interprete el coeficiente de *paidwork.*
b)¿Cuál es el incremento previsto de sueño, en promedio, si el tiempo dedicado al trabajo remunerado disminuye en una hora por día?
c) ¿Que parte de la variación en el sueño se explica por el tiempo dedicado a trabajo remunerado?

Ejercicio 2.20 La cuantificación de la felicidad no es una tarea fácil. Los investigadores de la Encuesta Mundial de Gallup investigaron sobre este tema mediante encuestas a miles de participantes en 155 países, entre 2006 y 2009, con el fin de medir dos tipos de bienestar. Se preguntó a los encuestados sobre la satisfacción general en su vida, utilizando una escala de puntuación de 1 a 10. Para explicar la satisfacción general (*stsfglo*) se formuló el siguiente modelo en el que cada observación se refiere a las medias obtenidas en los distintos países:

$$stsfglo = \beta_1 + \beta_2 lifexpec + u$$

donde *lifexpec* es la esperanza de vida al nacer, es decir, el número de años que se espera que viva un recién nacido.

Utilizando el archivo *HDR2010,* se obtiene el siguiente modelo ajustado:

$$\widehat{stsfglo} = -1.499 + 0.1062 lifexpec$$

$$R^2 = 0.6135 \qquad n = 144$$

a) Interprete el coeficiente de *lifexpec*.
b) ¿Cuál sería la media de satisfacción global en un país con una esperanza de vida al nacer de 80 años?

c) ¿Cuál debe ser la esperanza de vida al nacer para obtener una satisfacción global igual a 6?

Ejercicio 2.21 En economía se denomina intensidad en la actividad en investigación y desarrollo, o simplemente I+D, a la relación entre la inversión de una empresa en investigación y desarrollo y las ventas de dicha empresa.

Para la estimación un modelo que explique la intensidad en I+D es necesario contar con una base de datos apropiada. En España se puede utilizar la Encuesta sobre Estrategias Empresariales realizada por el Ministerio de Industria. Esta encuesta, con periodicidad anual, proporciona un profundo conocimiento de la evolución del sector industrial a través del tiempo, ya que ofrece múltiples datos relativos al desarrollo empresarial y a las decisiones de la empresa. Esta encuesta también está diseñada para generar información microeconómica que permite especificar y contrastar modelos econométricos. En cuanto a su cobertura, la población de referencia de esta encuesta son empresas con diez o más trabajadores de la industria manufacturera. El área geográfica de referencia es España, y los datos son anuales. Una de las características más destacadas de esta encuesta es su alto grado de representatividad.

Utilizando el fichero *rdspain*, que es una base de datos de las empresas españolas desde 1983 a 2006, se estimó la siguiente ecuación para explicar los gastos en investigación y desarrollo (*rdintens*):

$$\widehat{rdintens} = -2.639 + 0.2123\ln(sales)$$

$$R^2 = 0.0350 \qquad n = 1983$$

donde *rdintens* se expresa como un porcentaje de las ventas, y las ventas se miden en millones de euros.

a) Interprete el coeficiente de ln(*sales*).
b) Si las ventas aumentan en un 50%, ¿cuál es el cambio estimado en puntos porcentuales de *rdintens*?
c) ¿Qué porcentaje de la variación de *rdintens* se explica por las ventas? ¿Es elevado? Justifique su respuesta.

Ejercicio 2.22 El siguiente modelo se formuló para explicar el salario de un graduado MBA (*salMBAgr*) en función de las tasas de matrícula (*tuition*)

$$salMBAgr = \beta_1 + \beta_2 tuition + u$$

donde *salMBApr* es el salario medio anual en dólares para los estudiantes matriculados en el año 2010 de las 50 mejores escuelas de negocios americanas y *tuition* son los derechos de matrícula, incluyendo todos los gastos necesarios para el programa completo (con exclusión de los gastos de subsistencia).

Utilizando los datos de *MBAtui10*, se obtuvo el siguiente modelo ajustado:

$$\widehat{salMBAgr}_i = 54242 + 0.4313 tuition_i$$

$$R^2 = 0.4275 \qquad n = 50$$

a) ¿Cuál es la interpretación del término independiente?
b) ¿Cuál es la interpretación del coeficiente de la pendiente?
c) ¿Cuál es el valor predicho de *salMBAgr* para un estudiante de posgrado que pagó 110000 dólares por los derechos de matrícula en un MBA de 2 años?

Ejercicio 2.23 Usando una submuestra de la Encuesta Estructural de Salarios para España en 2006 (*wage06sp*), se estimó el siguiente modelo para explicar los salarios:

$$\ln(wage) = 1.919 + 0.0527 educ$$

$$R^2=0.2445 \qquad n=50$$

donde *educ* (educación) se mide en años y el salario (*wage*) en euros por hora.

a) ¿Cuál es la interpretación del coeficiente *educ*?

b) ¿Cuántos años de educación más se requieren para obtener un salario un 10% más elevado?

c) Sabiendo que $\overline{educ} = 10.2$, calcule la elasticidad salario/educación.

Ejercicio 2.24 Utilizando datos de la economía española para el período 1954-2010 (fichero *consump*), se estimó la función de consumo keynesiana:

$$conspc_t = -288 + 0.9416 incpc_t$$

$$R^2=0.994 \qquad n=57$$

donde el consumo (*conspc*) y la renta disponible (*incpc*) se expresan en euros constantes per cápita, tomando 2008 como año de referencia.

a) ¿Cuál es la interpretación del término independiente? Opine sobre el signo y magnitud del término independiente.

b) Interprete el coeficiente de *incpc*. ¿Cuál es el significado económico de este coeficiente?

c) Compare la propensión marginal a consumir con la propensión media al consumo para el punto de la media muestral ($\overline{conspc} = 8084$, $\overline{incpc} = 8896$). Comente el resultado obtenido.

d) Calcule la elasticidad consumo/renta para la media muestral.

Anexo 2.1 Un caso de estudio: Curvas de Engel para la demanda de productos lácteos

La curva de Engel muestra la relación entre las diversas cantidades de un bien que el consumidor está dispuesto a comprar para diferentes niveles de renta.

En una encuesta realizada a 40 familias se han obtenido datos de gasto anual en productos lácteos y de renta disponible que aparecen en el cuadro 2.6. Para evitar distorsiones debidas al diferente tamaño de los hogares, tanto el consumo como la renta se han expresado en términos *per capita*. Los datos vienen expresados en miles de euros al mes.

Antes de proceder a su estimación con los datos del cuadro 2.6, vamos exponer varios tipos de modelos que se utilizan en los estudios de demanda, analizando las propiedades de cada uno de ellos. Los modelos que se van examinar son los siguientes: lineal, inverso, semilogarítmico, potencial, exponencial y exponencial inverso. En los tres primeros modelos, el regresando de la ecuación a estimar es directamente la variable endógena, mientras que en los tres últimos, después de realizar las transformaciones adecuadas, el regresando es el logaritmo neperiano de la variable endógena.

En todos los modelos se calculará la propensión marginal, así como la elasticidad de la demanda.

CUADRO 2.6 Gasto en productos lácteos (*dairy*), renta disponible (*inc*) en términos *per capita*. (Unidad: euros por mes)

familia	*dairy*	*inc*	*familia*	*dairy*	*inc*
1	8.87	1.250	21	16.20	2.100
2	6.59	985	22	10.39	1.470
3	11.46	2.175	23	13.50	1.225
4	15.07	1.025	24	8.50	1.380
5	15.60	1.690	25	19.77	2.450
6	6.71	670	26	9.69	910
7	10.02	1.600	27	7.90	690
8	7.41	940	28	10.15	1.450
9	11.52	1.730	29	13.82	2.275
10	7.47	640	30	13.74	1.620
11	6.73	860	31	4.91	740
12	8.05	960	32	20.99	1.125
13	11.03	1.575	33	20.06	1.335
14	10.11	1.230	34	18.93	2.875
15	18.65	2.190	35	13.19	1.680
16	10.30	1.580	36	5.86	870
17	15.30	2.300	37	7.43	1.620
18	13.75	1.720	38	7.15	960
19	11.49	850	39	9.10	1.125
20	6.69	780	40	15.31	1.875

Modelo lineal

El modelo lineal de la demanda de productos lácteos es el siguiente:

$$dairy = \beta_1 + \beta_2 inc + u \tag{2-73}$$

Como sabemos la propensión marginal del gasto nos indica cómo cambia el gasto al variar la renta, y se obtiene derivando el gasto con respecto a la renta en la ecuación de demanda. En el modelo lineal la propensión marginal del gasto en productos lácteos viene dada por

$$\frac{d\ dairy}{d\ inc} = \beta_2 \tag{2-74}$$

Es decir, en el modelo lineal la propensión marginal se mantiene constante y, por lo tanto, es independiente del valor que tome la renta. El hecho de que sea constante es una ventaja, pero al mismo tiempo tiene el inconveniente de que puede no ser adecuada para describir el comportamiento de los consumidores, especialmente cuando existan diferencias importantes en la renta de las familias analizadas. Así, no parece plausible que una familia con unos ingresos mensuales de 700 euros dedique al consumo de productos lácteos de cada euro adicional de que disponga una proporción igual que la que dedicaría una familia con ingresos de 20000 euros. Ahora bien, si la variación de la renta no es muy elevada un modelo lineal puede ser adecuado para describir la demanda de ciertos bienes.

La propensión marginal mide el cambio absoluto que se produce en el gasto al variar la renta. En muchas ocasiones, sin embargo, el investigador está más interesado en conocer cuál es la tasa de variación del gasto ante una variación de la renta medida en porcentaje. Así, en este caso en concreto el investigador puede tener un especial interés, por ejemplo, en conocer el porcentaje de variación del gasto en productos lácteos al incrementarse la renta en un 1%. Este tipo de aproximación requiere que se calcule la elasticidad gasto/renta.

En términos matemáticos, la elasticidad gasto/renta viene dada por

$$\varepsilon_{lacteos/rendis}^{linear} = \frac{d\ dairy}{d\ inc} inc = \beta_2 \frac{inc}{dairy} \tag{2-75}$$

Estimando el modelo (2-73) con los datos del cuadro 2.6, obtenemos

$$dairy = 4.012 + 0.005288 \times inc \qquad R^2 = 0.4584 \tag{2-76}$$

Modelo inverso

En el modelo inverso se establece una relación lineal entre el gasto y la inversa de la renta. Por lo tanto, este modelo es directamente lineal en los parámetros. Su expresión es la siguiente:

$$dairy = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-77}$$

El signo del coeficiente β_2 será negativo en el caso normal de que la renta esté correlacionada positivamente con el gasto en el bien. Como puede comprobarse fácilmente, cuando la renta tiende hacia infinito, el gasto tiende a un límite que es igual a β_1. Es decir, β_1 representa el máximo consumo que puede haber de ese bien.

En la figura 2.14 puede verse la representación de la parte sistemática de este modelo. En la primera figura se ha representado la relación entre la variable dependiente y la variable explicativa. En la segunda se ha representado la relación entre la regresando y regresor. La segunda función es lineal como se puede ver en la figura.

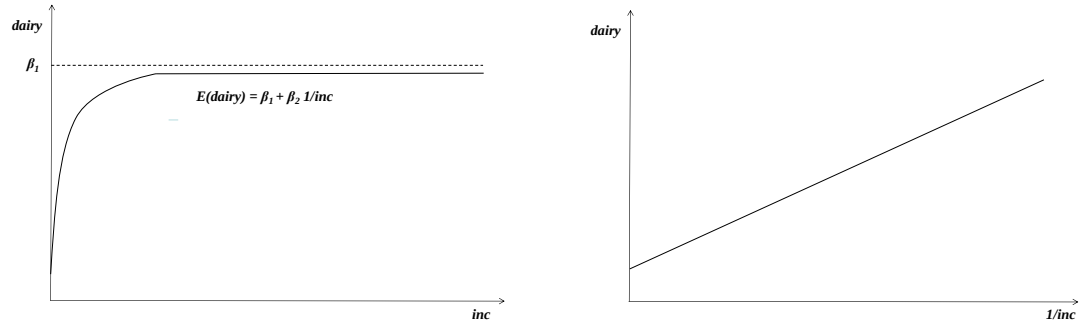


FIGURA 2.14. El modelo inverso.

En el modelo inverso la propensión marginal del gasto viene dada por

$$\frac{d\ dairy}{d\ inc} = -\beta_2 \frac{1}{(inc)^2} \tag{2-78}$$

De acuerdo con (2-78), la propensión marginal al consumo va disminuyendo de forma inversamente proporcional al cuadrado del nivel de renta.

Por otra parte, la elasticidad disminuye, según puede verse en la siguiente expresión, de forma inversamente proporcional al producto del gasto por la renta:

$$\varepsilon_{dairy/inc}^{inv} = \frac{d\ dairy}{d\ inc} \frac{inc}{dairy} = -\beta_2 \frac{1}{inc \times dairy} \tag{2-79}$$

Estimando el modelo (2-77) con los datos del cuadro 2.6, se obtiene

$$dairy = 18.652 - 8702 \frac{1}{inc} \qquad R^2 = 0.4281 \tag{2-80}$$

En este caso, el coeficiente $\hat{\beta}_2$ no tiene un significado económico.

Modelo lineal logarítmico

Este modelo recibe la denominación de lineal logarítmico por ser el gasto una función lineal del logaritmo de la renta, es decir,

$$dairy = \beta_1 + \beta_2 \ln(inc) + u \tag{2-81}$$

En este modelo, la propensión marginal al gasto viene dada por

$$\frac{d\ dairy}{d\ inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{inc} = \frac{d\ dairy}{d\ln(inc)}\frac{1}{inc} = \beta_2 \frac{1}{inc} \tag{2-82}$$

y la elasticidad gasto/renta viene dada por

$$\varepsilon^{lin\text{-}\log}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ dairy}{d\ \ln(inc)}\frac{1}{dairy} = \beta_2 \frac{1}{dairy} \tag{2-83}$$

La propensión marginal es inversamente proporcional al nivel de renta en el modelo lineal logarítmico, mientras que la elasticidad es inversamente proporcional al nivel de gasto en productos lácteos.

En la figura 2.15, podemos ver a una doble representación de la función poblacional correspondiente a este modelo.

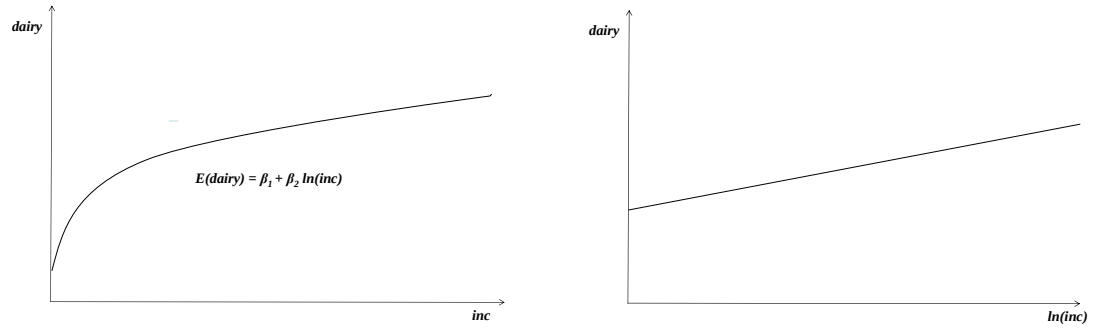


FIGURA 2.15. El modelo lineal logarítmico.

Estimando el modelo (2-81) con los datos del cuadro 2.6, se obtiene

$$dairy = -41.623 + 7.399 \times \ln(inc) \quad R^2 = 0.4567 \tag{2-84}$$

La interpretación de $\hat{\beta}_2$ es la siguiente: si la renta aumenta en un 1%, la demanda de productos lácteos se incrementará en 0.07399 euros.

Modelo potencial o doblemente logarítmico

El modelo potencial se define de la siguiente manera:

$$dairy = e^{\beta_1} inc^{\beta_2} e^u \tag{2-85}$$

Este modelo no es lineal en los parámetros, pero es linealizable, ya que al tomar logaritmos neperianos se obtiene el modelo:

$$\ln(dairy) = \beta_1 + \beta_2 \ln(inc) + u \tag{2-86}$$

A este modelo se le denomina también doblemente logarítmico, ya que ésta es la estructura de la versión linealizada.

En el modelo potencial la propensión marginal de la demanda viene dada por

$$\frac{d\ dairy}{d\ inc} = \beta_2 \frac{dairy}{inc} \tag{2-87}$$

En el modelo potencial la elasticidad es constante. Por lo tanto, ante una variación dada de la renta, el gasto se incrementará en el mismo porcentaje con independencia de cuál sea el nivel de renta y gasto a que se aplique. La expresión de la elasticidad es la siguiente:

$$\varepsilon_{dairy/inc}^{log\text{-}log} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ \ln(inc)} = \beta_2 \tag{2-88}$$

En la figura 2.16 puede verse una doble representación de la función poblacional correspondiente a este modelo.

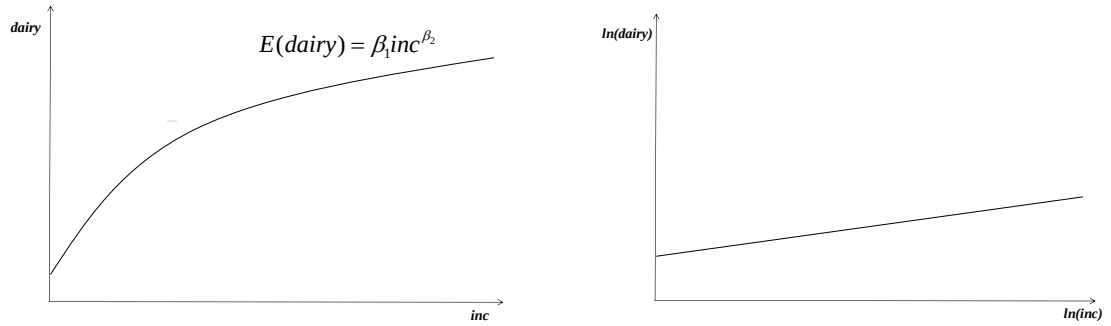


FIGURA 2.16. Modelo doblemente logarítmico

Estimando el modelo (2-86) con los datos del cuadro 2.6, se obtiene

$$\ln(dairy) = -2.556 + 0.6866 \times \ln(inc) \quad R^2 = 0.5190 \tag{2-89}$$

En este caso $\hat{\beta}_2$ es la elasticidad del gasto/renta. Su interpretación es la siguiente: si el ingreso aumenta en un 1%, la demanda de productos lácteos se incrementará en un 0,68%.

Modelo exponencial

El modelo exponencial se define del siguiente modo:

$$dairy = \exp(\beta_1 + \beta_2 inc + u) \tag{2-90}$$

Tomando logaritmos neperianos en ambos miembros de (2-90), se obtiene el siguiente modelo que es lineal en los parámetros:

$$\ln(dairy) = \beta_1 + \beta_2 inc + u \tag{2-91}$$

En el modelo exponencial la propensión marginal del gasto viene dada por

$$\frac{d\ dairy}{d\ inc} = \beta_2 dairy \tag{2-92}$$

En el modelo exponencial, a diferencia de otros modelos vistos anteriormente, la propensión marginal aumenta cuando el nivel de gasto lo hace. Por esta razón, este

modelo es adecuado para describir la demanda de productos de lujo. Por otro lado, la elasticidad es proporcional al nivel de renta:

$$\varepsilon_{dairy/inc}^{exp} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc}inc = \beta_2 inc \tag{2-93}$$

En la figura 2.17, podemos ver a una doble representación de la función poblacional correspondiente a este modelo.

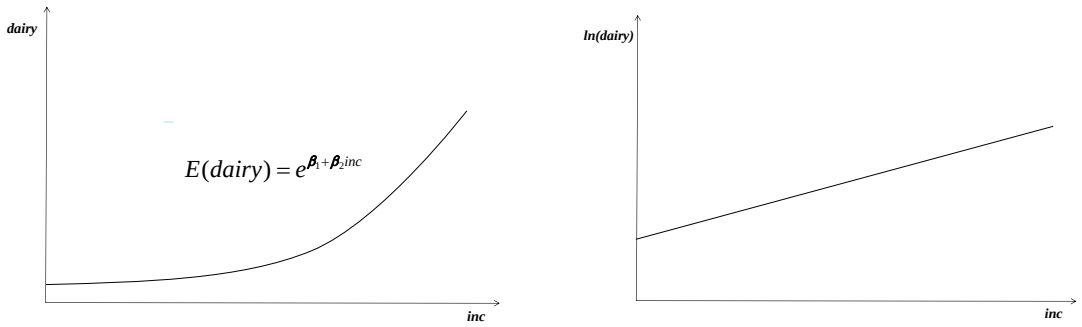


FIGURA 2.17. El modelo exponencial.

Estimando el modelo (2-91) con los datos del cuadro 2.6 se obtiene

$$\ln(dairy) = 1.694 + 0.00048 \times inc \quad R^2 = 0.4978 \tag{2-94}$$

La interpretación de $\hat{\beta}_2$ es la siguiente: si la renta se incrementa en 1 euro la demanda de productos lácteos se incrementará en un 0.048%.

Modelo exponencial inverso

El modelo exponencial inverso es una mezcla del modelo exponencial y del modelo inverso, teniendo propiedades que lo hacen adecuado para determinar la demanda de productos en los que hay un punto de saturación. Su expresión es la siguiente:

$$dairy = \exp(\beta_1 + \beta_2 \frac{1}{inc} + u) \tag{2-95}$$

Tomando logaritmos neperianos en ambos miembros de (2-95) se obtiene el siguiente modelo que es lineal en los parámetros:

$$\ln(dairy) = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-96}$$

En el modelo exponencial inverso la propensión marginal del gasto viene dada por

$$\cdot\ \frac{d\ dairy}{d\ inc} = -\beta_2 \frac{dairy}{(inc)^2} \tag{2-97}$$

y la elasticidad por

$$\varepsilon_{dairy/inc}^{invexp} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc}inc = -\beta_2 \frac{1}{inc} \tag{2-98}$$

Estimando el modelo (2-96) con los datos de la tabla 2.6 se obtiene

$$\ln(dairy) = 3.049 - 822.02\frac{1}{inc} \qquad R^2 = 0.5040 \tag{2-99}$$

En este caso, como en el modelo inverso, el coeficiente $\hat{\beta}_2$ no tiene un significado económico.

En la tabla 2.7, se muestran los resultados de la propensión marginal, la elasticidad del gasto/renta y el R^2 en los seis modelos ajustados.

TABLA 2.7. Propensión marginal, elasticidad gasto/renta y R^2 en los modelos estimados para analizar la demanda de productos lácteos.

Modelo	*Propensión marginal*	*Elasticidad*	R^2
Lineal	$\hat{\beta}_2$ =0.0053	$\hat{\beta}_2 \frac{\overline{inc}}{\overline{dairy}}$ =0.6505	0.4440
Inverso	$-\hat{\beta}_2 \frac{1}{\left[\overline{inc}\right]^2}$ =0.0044	$-\hat{\beta}_2 \frac{1}{\overline{dairy} \times \overline{inc}}$ =0.5361	0.4279
Lineal logarítmico	$\hat{\beta}_2 \frac{1}{\overline{inc}}$ =0.0052	$\hat{\beta}_2 \frac{1}{\overline{dairy}}$ =0.6441	0.4566
Doblemente logarítmico	$\hat{\beta}_2 \frac{\overline{dairy}}{\overline{inc}}$ =0.0056	$\hat{\beta}_2$ =0.6864	0.5188
Logarítmico lineal	$\hat{\beta}_2 \times \overline{dairy}$ =0.0055	$\hat{\beta}_2 \times \overline{inc}$ =0.6783	0.4976
Logarítmico inverso	$-\hat{\beta}_2 \frac{\overline{dairy}}{\left[\overline{inc}\right]^2}$ =0.0047	$-\hat{\beta}_2 \frac{1}{\overline{inc}}$ =0.5815	0.5038

El R^2 obtenido en los tres primeros modelos no es comparable con el R^2 obtenido en los tres últimos porque la forma funcional del regresando es diferente: y en los tres primeros modelos y $\ln(y)$ en los tres últimos.

Comparando los tres primeros modelos entre sí, el mejor ajuste se obtiene con el modelo lineal logarítmico si utilizamos R^2 como medida de bondad de ajuste. Comparando los tres últimos modelos el mejor ajuste corresponde al modelo doblemente logarítmico. Si se hubiera utilizado el Criterio de Información de Akaike (AIC), que permite comparar los modelos con diferentes formas funcionales para el regresando, entonces el modelo doblemente logarítmico habría sido el mejor entre los seis modelos estimados. La medida AIC será estudiada en el capítulo 3.

Apéndices

Apéndice 2.1: Dos formas alternativas de expresar $\hat{\beta}_2$

Es fácil ver que

$$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=\sum_{i=1}^{n}(y_ix_i-\bar{x}y_i-\bar{y}x_i+\overline{yx})=\sum_{i=1}^{n}y_ix_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+n\overline{yx}$$

$$=\sum_{i=1}^{n}y_ix_i-n\overline{xy}-\bar{y}\sum_{i=1}^{n}x_i+n\overline{yx}=\sum_{i=1}^{n}y_ix_i-\bar{y}\sum_{i=1}^{n}x_i$$

Por otro lado, tenemos que

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\overline{xx})^2=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\overline{xx}$$

$$=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i$$

Por lo tanto, (2-17) se puede expresar de la siguiente manera:

$$\hat{\beta}_2=\frac{\sum_{i=1}^{n}y_ix_i-\bar{y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Apéndice 2.2. Demostración de que $r_{xy}^2=R^2$

En primer lugar vamos a estudiar una equivalencia que se va a utilizar en la demostración. Por definición,

$$\hat{y}_i=\hat{\beta}_1+\hat{\beta}_2x_i$$

De la primera ecuación normal, tenemos que

$$\bar{y}=\hat{\beta}_1+\hat{\beta}_2\bar{x}$$

Restando la segunda ecuación de la primera:

$$\hat{y}_i-\bar{y}=\hat{\beta}_2(x_i-\bar{x})$$

Elevando al cuadrado ambos miembros

$$(\hat{y}_i-\bar{y})^2=\hat{\beta}_2^2(x_i-\bar{x})^2$$

y sumando para todo *i*, tenemos

$$\sum(\hat{y}_i-\bar{y})^2=\hat{\beta}_2^2\sum(x_i-\bar{x})^2$$

Teniendo en cuenta la anterior equivalencia, tenemos que

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = r_{xy}^2$$

Apéndice 2.3. Cambio proporcional versus cambio en logaritmos

El cambio en logaritmos es una tasa de variación, que se utiliza en la investigación económica. La relación entre el cambio proporcional y el cambio en logaritmos puede verse si se aplica un desarrollo en serie de Taylor a (2-45):

$$\ln(x_1) - \ln(x_0) = \ln\left[\frac{x_1}{x_0}\right]$$

$$= \ln(1) + \left[\frac{x_1}{x_0} - 1\right]\left[\frac{1}{\frac{x_1}{x_0}}\right]_{\frac{x_1}{x_0}=1} + \frac{1}{2}\left[\frac{x_1}{x_0} - 1\right]^2\left[-\frac{1}{\left[\frac{x_1}{x_0}\right]}\right]_{\frac{x_1}{x_0}=1}$$

$$+ \frac{1}{3\times 2}\left[\frac{x_1}{x_0} - 1\right]^3\left[\frac{2}{\left[\frac{x_1}{x_0}\right]^3}\right]_{\frac{x_1}{x_0}=1} + \cdots\cdots \qquad (2\text{-}100)$$

$$= \left[\frac{x_1}{x_0} - 1\right] - \frac{1}{2}\left[\frac{x_1}{x_0} - 1\right]^2 + \frac{1}{3}\left[\frac{x_1}{x_0} - 1\right]^3 + \cdots$$

$$= \frac{\Delta x_1}{x_0} - \frac{1}{2}\left[\frac{\Delta x_1}{x_0}\right]^2 + \frac{1}{3}\left[\frac{\Delta x_1}{x_0}\right]^3 + \cdots$$

Por lo tanto, si tomamos la aproximación lineal en este desarrollo, tenemos que

$$\Delta \ln(x) = \ln(x_1) - \ln(x_0) = \ln\left[\frac{x_1}{x_0}\right] \approx \frac{\Delta x_1}{x_0} \qquad (2\text{-}101)$$

Apéndice 2.4. Demostración de que los *estimadores MCO son lineales e insesgados*

Solo demostraremos la insesgadez del estimador $\hat{\beta}_2$ que es el más relevante. Para demostrarlo, debemos expresar nuestro estimador en términos del parámetro poblacional. La fórmula (2-18) se puede expresar como

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-102}$$

ya que $\sum_{i=1}^{n}(x_i - \bar{x})\,\bar{y} = \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = \bar{y} \times 0 = 0$

Ahora vamos a expresar (2-102) de la siguiente manera:

$$\hat{\beta}_2 = \sum_{i=1}^{n} c_i y_i \tag{2-103}$$

donde

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-104}$$

Los coeficientes c_i tienen las siguientes propiedades

$$\sum_{i=1}^{n} c_i = 0 \tag{2-105}$$

$$\sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-106}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1 \tag{2-107}$$

Ahora, si sustituimos $y = \beta_1 + \beta_2 x + u$ (supuesto 1) en (2-102), tenemos que

$$\begin{aligned} \hat{\beta}_2 &= \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i(\beta_1 + \beta_2 x_i + u_i) \\ &= \beta_1 \sum_{i=1}^{n} c_i + \beta_2 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i u_i = \beta_2 + \sum_{i=1}^{n} c_i u_i \end{aligned} \tag{2-108}$$

Asumiendo que los regresores son no estocásticos (supuesto 2), c_i será también no estocástico. Por lo tanto, $\hat{\beta}_2$ es un estimador que es función lineal de u.

Tomando esperanzas en (2-108) y teniendo en cuenta el supuesto 6, e implícitamente los supuestos del 3 al 5, se obtiene

$$E(\hat{\beta}_2) = \beta_2 + \sum_{i=1}^{n} c_i E(u_i) = \beta_2 \tag{2-109}$$

Por lo tanto, $\hat{\beta}_2$ es un estimador insesgado de β_2

Apéndice 2.5. Cálculo de la varianza de $\hat{\beta}_2$:

$$E\left[\hat{\beta}_2 - \beta_2\right]^2 = \left[\sum_{i=1}^{n} c_i u_i\right]^2 = \sum_{i=1}^{n} c_i^2 E(u_i^2) + \sum_{i \neq j}\sum_{i=1}^{n} c_i c_j E(u_i u_j)$$
$$= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{nS_X^2} \qquad (2\text{-}110)$$

En la demostración anterior, al pasar de la segunda a la tercera igualdad, se han tenido en cuenta los supuestos 6 y 7.

Apéndice 2.6. Demostración del teorema de Gauss-Markov para la pendiente en la regresión simple

El procedimiento que vamos a seguir para la demostración es el siguiente. En primer lugar, vamos a definir un estimador arbitrario, $\tilde{\beta}_2$, que es lineal en *y*. En segundo lugar, vamos a imponer las restricciones que se requieren para que sea insesgado. En tercer lugar, se mostrará que la varianza de este estimador arbitrario debe ser mayor, o por lo menos igual, que la varianza de $\hat{\beta}_2$.

Así pues, vamos a definir un estimador arbitrario, $\tilde{\beta}_2$, que es lineal en *y*:

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i \qquad (2\text{-}111)$$

Ahora, sustituimos y_i por su valor en el modelo poblacional (supuesto 1):

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i = \sum_{i=1}^{n} h_i(\beta_1 + \beta_2 x_i + u_i) = \beta_1 \sum_{i=1}^{n} h_i + \beta_2 \sum_{i=1}^{n} h_i x_i + \sum_{i=1}^{n} h_i u_i \qquad (2\text{-}112)$$

Para que el estimador $\tilde{\beta}_2$ sea insesgado es necesario que las restricciones siguientes se cumplan:

$$\sum_{i=1}^{n} h_i = 0 \qquad \sum_{i=1}^{n} h_i x_i = 1 \qquad (2\text{-}113)$$

Por lo tanto,

$$\tilde{\beta}_2 = \beta_2 + \sum_{i=1}^{n} h_i u_i \qquad (2\text{-}114)$$

La varianza de este estimador es la siguiente:

$$E\left[\tilde{\beta}_2-\beta_2\right]^2=\left[\sum_{i=1}^{n}h_iu_i\right]^2=\sigma^2\sum_{i=1}^{n}h_i^2=$$

$$\sigma^2\sum_{i=1}^{n}\left[h_i-\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}+\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]^2=\sigma^2\sum_{i=1}^{n}\left[h_i-\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]^2 \qquad (2\text{-}115)$$

$$+\sigma^2\sum_{i=1}^{n}\left[\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]^2+2\sigma^2\sum_{i=1}^{n}\left[h_i-\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

El tercer término de la última igualdad es 0, como se muestra a continuación:

$$2\sigma^2\sum_{i=1}^{n}\left[h_i-\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$=2\sigma^2\sum_{i=1}^{n}\left[h_i\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]-2\sigma^2\sum_{i=1}^{n}\left[\frac{(x_i-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]=2\sigma^2\times1-2\sigma^2\times1=0 \qquad (2\text{-}116)$$

Por lo tanto, teniendo en cuenta (2-116) y operando, tenemos que

$$E\left[\tilde{\beta}_2-\beta_2\right]^2=\sigma^2\sum_{i=1}^{n}\left[h_i-c_i\right]^2+\sigma^2\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad (2\text{-}117)$$

donde $c_i=\dfrac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$

El segundo término de la última igualdad es la varianza de $\hat{\beta}_2$, mientras que el primer término es siempre positivo, ya que es una suma de cuadrados, excepto que se cumpla que $h_i=c_i$, para todo i, en cuyo caso será igual a 0, y entonces $\tilde{\beta}_2=\hat{\beta}_2$. Así pues,

$$E\left[\tilde{\beta}_2-\beta_2\right]^2\geq E\left[\hat{\beta}_2-\beta_2\right]^2 \qquad (2\text{-}118)$$

Apéndice 2.7. Demostración de que $\hat{\sigma}^2$ es un estimador insesgado de la varianza de las perturbaciones

El modelo poblacional es, por definición:

$$y_i=\beta_1+\beta_2x_i+u_i \qquad (2\text{-}119)$$

Si sumamos ambos miembros de para todo *i* y dividimos por *n*, tenemos

$$\bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{u} \tag{2-120}$$

Restando (2-120) de (2-119), tenemos que

$$y_i - \bar{y} = \beta_2 (x_i - \bar{x}) + (u_i - \bar{u}) \tag{2-121}$$

Por otra parte, $\hat{u}_i$ es por definición:

$$\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \tag{.2-122}$$

Si sumamos ambos miembros de (2-122) y dividimos por *n*, tenemos

$$\bar{\hat{u}} = \bar{y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{x} \tag{2-123}$$

Restando (2-123 de (2-122), y teniendo en cuenta que $\bar{\hat{u}}$ =0,

$$\hat{u}_i = (y_i - \bar{y}) - \hat{\beta}_2 (x_1 - \bar{x}) \tag{2-124}$$

Sustituyendo (2-121) en (2-124), tenemos que

$$\begin{aligned} \hat{u}_i &= \beta_2 (x_i - \bar{x}) + (u_i - \bar{u}) - \hat{\beta}_2 (x_1 - \bar{x}) \\ &= -(\hat{\beta}_2 - \beta_2)(x_1 - \bar{x}) + (u_i - \bar{u}) \end{aligned} \tag{2-125}$$

Elevando al cuadrado y sumando en ambos miembros de (2-125), se tiene que

$$\begin{aligned} \sum_{i=1}^{n} \hat{u}_i^2 = \left[\tilde{\beta}_2 - \beta_2\right]^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 \\ -2\left[\tilde{\beta}_2 - \beta_2\right] \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u}) \end{aligned} \tag{2-126}$$

Tomando las esperanzas en (2-126), se obtiene que

$$\begin{aligned} E\left[\sum_{i=1}^{n} \hat{u}_i^2\right] &= \sum_{i=1}^{n} (x_i - \bar{x})^2 E\left[\tilde{\beta}_2 - \beta_2\right]^2 + E\left[\sum_{i=1}^{n} (u_i - \bar{u})^2\right] \\ &\quad -2E\left[(\tilde{\beta}_2 - \beta_2) \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u})\right] \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned} \tag{2-127}$$

Para obtener el primer término de la última igualdad de (2-127), se ha utilizado (2-64). Para obtener el segundo y el tercer término de la última igualdad de (2-127) se

han utilizado los desarrollos que se hacen en (2-128) y (2-129) respectivamente. En ambos casos se han tenido en cuenta los supuestos 7 y 8.

$$E\left[\sum_{i=1}^{n}(u_i-\bar{u})^2\right]=E\left[\sum_{i=1}^{n}u_i^2-n\bar{u}^2\right]=E\left[\sum_{i=1}^{n}u_i^2-n\left(\frac{\sum_{i=1}^{n}u_i}{n}\right)^2\right] \tag{2-128}$$

$$=E\left[\sum_{i=1}^{n}u_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}u_i^2+\sum_{i\neq j}u_iu_j\right)\right]=n\sigma^2-\frac{n}{n}\sigma^2=(n-1)\sigma^2$$

$$E\left[\left(\tilde{\beta}_2-\beta_2\right)\sum_{i=1}^{n}(x_i-\bar{x})(u_i-\bar{u})\right]=E\left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})u_i\sum_{i=1}^{n}(x_i-\bar{x})u_i\right]$$

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

$$=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2E(u_i)^2+\sum_{i\neq j}\sum(x_i-\bar{x})(x_i-\bar{x})E(u_iu_j)\right]=\sigma^2$$

129)

De acuerdo con (2-127), se tiene que

$$E\left[\sum_{i=1}^{n}\hat{u}_i^2\right]=(n-2)\sigma^2 \tag{2-130}$$

Por lo tanto, un estimador insesgado viene dado por

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2} \tag{2-131}$$

ya que

$$E\left(\hat{\sigma}^2\right)=\frac{1}{n-2}E\left(\sum_{i=1}^{n}\hat{u}_i^2\right)=\sigma^2 \tag{2-132}$$

Apéndice 2.8. Consistencia de los estimadores de *MCO*

El operador plim tiene la propiedad de invarianza (propiedad de Slutsky). Es decir, si $\hat{\theta}$ es un estimador consistente de θ y $g\left(\hat{\theta}\right)$ es cualquier función continua de $\hat{\theta}$, entonces

$$\operatorname{plim}_{n\to\infty} g(\hat{\theta}) = g(\theta) \tag{2-133}$$

Esto significa que si $\hat{\theta}$ es un estimador consistente de θ, entonces $1/\hat{\theta}$ y $\ln(\hat{\theta})$ son también estimadores consistentes de $1/\theta$ y $\ln(\theta)$, respectivamente. Hay que tener en cuenta que estas propiedades no son válidas para el operador esperanza E; por ejemplo, si $\hat{\theta}$ es un estimador insesgado de θ [es decir, $E(\hat{\theta})=\theta$], no es cierto que $1/\theta$ sea un estimador insesgado de una $1/\theta$, es decir, $E(1/\hat{\theta}) \neq 1/E(\hat{\theta}) \neq 1/\theta$. Esto es debido al hecho de que el operador esperanza únicamente puede ser aplicado a funciones *lineales* de variables aleatorias. Por otra parte, el operador plim es aplicable a cualquier función continua.

Bajo los supuestos del 1 al 6, los estimadores de *MCO*, $\hat{\beta}_1$ y $\hat{\beta}_2$ son consistentes.

Ahora vamos a demostrar en particular que $\hat{\beta}_2$ es un estimador consistente. En primer lugar, $\hat{\beta}_2$ se puede expresar como:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_1+\beta_2 x_i+u_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ &= \frac{\beta_1\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \beta_2\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_2 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\end{aligned} \tag{2-134}$$

Con el fin de comprobar su consistencia necesitamos tomar plim en (2-134) y aplicar la *Ley de los Grandes Números*. Esta ley establece que, en condiciones generales, los momentos muestrales convergen a sus correspondientes momentos poblacionales. Por lo tanto, tomando plim en (2-134):

$$\operatorname{plim}_{n\to\infty}\hat{\beta}_2 = \operatorname{plim}_{n\to\infty}\left[\beta_2 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \beta_2 + \frac{\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2-135}$$

En esta última igualdad hemos dividido el numerador y el denominador del segundo término por n, porque, si no lo hacemos, ambos sumatorios tenderán a infinito cuando n tiende a infinito.

Si aplicamos la ley de grandes números al numerador y denominador del segundo término de (2-135), convergerán en probabilidad a las cantidades poblacionales $cov(x,u)$ y $var(x)$ respectivamente. Siempre que $var(x)\neq 0$ (supuesto 4), podemos utilizar las propiedades del *límite de probabilidad* para obtener

$$\operatorname{plim}\hat{\beta}_2 = \beta_2 + \frac{cov(x,u)}{var(x)} = \beta_2 \tag{2-136}$$

Para alcanzar la última igualdad, utilizando los supuestos 2 y 6, obtenemos que

$$cov(x,u) = E\left[(x-\bar{x})u\right] = (x-\bar{x})E\left[u\right] = (x-\bar{x})\times 0 = 0 \tag{2-137}$$

Por lo tanto, $\hat{\beta}_2$ es un estimador consistente.

Apéndice 2.9 Estimación por máxima verosimilitud

Teniendo en cuenta los supuestos del 1 al 6 la esperanza de y_i es la siguiente:

$$E(y_i) = \beta_1 + \beta_2 x_i \tag{2-138}$$

Si tenemos en cuenta el supuesto 7, la varianza de y_i es igual a

$$\text{var}(y_i) = E\left[y_i - E(y_i)\right]^2 = E\left[y_i - \beta_1 + \beta_2 x_i\right]^2 = E\left[u_i\right]^2 = \sigma^2 \quad \forall i \tag{2-139}$$

De acuerdo con el supuesto 1 y_i es una función lineal de u_i, y si u_i tiene una distribución normal (supuesto 9), entonces y_i será normal e independientemente distribuida (supuesto 8) con media $\beta_1 + \beta_2 x_i$ y varianza σ^2.

Entonces, la función de densidad de probabilidad conjunta de $y_1, y_2, \cdots, y_n$ se puede expresar como un producto de *n* funciones de densidad individuales:

$$\begin{aligned} &f\left(y_1,\, y_2,\, \ldots, y_n \,|\, \beta_1 + \beta_2 x_i,\, \sigma^2\right) \\ &= f\left(y_1 \,|\, \beta_1 + \beta_2 x_i,\, \sigma^2\right) f\left(y_2 \,|\, \beta_1 + \beta_2 x_i,\, \sigma^2\right) \cdots f\left(y_n \,|\, \beta_1 + \beta_2 x_i,\, \sigma^2\right) \end{aligned} \tag{2-140}$$

donde

$$f\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \tag{2-141}$$

que es la función de densidad de una variable distribuida normalmente con la media y la varianza dada.

Sustituyendo (2-141) en (2-140)para cada y_i, se obtiene

$$\begin{aligned} &f\left(y_1,\, y_2, \cdots,\, y_n\right) = f\left(y_1\right) f\left(y_2\right) \cdots f\left(y_n\right) \\ &= \frac{1}{\sigma^n \left(\sqrt{2\pi}\right)^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \end{aligned} \tag{2-142}$$

Si se conocen $y_1, y_2, \cdots, y_n$, pero β_2, β_3, y σ^2 son desconocidos, a la función en (2-142) se denomina *función de verosimilitud*, y se denota por $L(\beta_2, \beta_3, \sigma^2)$ o simplemente L. Si se toman logaritmos en (2-142), se obtiene

$$\begin{aligned}\ln L &= -n\ln\sigma - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i-\beta_1-\beta_2 x_i\right]^2}{\sigma^2} \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i-\beta_1-\beta_2 x_i\right]^2}{\sigma^2}\end{aligned} \tag{2-143}$$

El método de *máxima* verosimilitud (*MV*), como su nombre sugiere, consiste en estimar los parámetros desconocidos de tal manera que la probabilidad de observar las y_i dadas sea tan alta como sea posible. Por lo tanto, tenemos para encontrar el máximo de la función (2-143). Para maximizar (2-143). hay que derivar parcialmente con respecto a β_2, β_3, y σ^2 e igualar a 0. Denominando $\tilde{\beta}_1$, $\tilde{\beta}_2$ y $\tilde{\sigma}^2$ a los estimadores de *MV*, obtenemos que:

$$\begin{aligned}\frac{\partial \ln L}{\partial \tilde{\beta}_1} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y-\tilde{\beta}_1-\tilde{\beta}_2 x_i\right)(-1) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\beta}_2} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y-\tilde{\beta}_1-\tilde{\beta}_2 x_i\right)(-x_i) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\sigma}^2} &= -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum\left(y-\tilde{\beta}_1-\tilde{\beta}_2 x_i\right)^2 = 0\end{aligned} \tag{2-144}$$

Si tomamos las dos primeras ecuaciones de (2-144) y operamos, tenemos que

$$\sum y_i = n\tilde{\beta}_1 + \tilde{\beta}_2\sum x_i \tag{2-145}$$

$$\sum y_i x_i = \tilde{\beta}_1\sum x_i + \tilde{\beta}_2\sum x_i^2 \tag{2-146}$$

Como puede verse, (2-145) y (2-146) son iguales a (2-13) y (2-14), es decir, los estimadores de *MV*, bajo los supuestos del *MLC*, son iguales a los estimadores de *MCO*.

Sustituyendo $\tilde{\beta}_1$ y $\tilde{\beta}_2$, -obtenidos al resolver (2-145) y (2-146)- en la tercera ecuación de (2-144) se tiene que

$$\tilde{\sigma}^2 = \frac{1}{n}\sum\left(y_i-\tilde{\beta}_1-\tilde{\beta}_2 x_i\right)^2 = \frac{1}{n}\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i\right)^2 = \frac{1}{n}\sum\hat{u}_i^2 \tag{2-147}$$

El estimador de *MV* de $\tilde{\sigma}^2$ es sesgado, ya que, de acuerdo con (2-131),

$$E(\tilde{\sigma}^2) = \frac{1}{n}E\left[\sum_{i=1}^{n}\hat{u}_i^2\right] = \frac{n-2}{n}\sigma^2 \tag{2-148}$$

En cualquier caso, $\tilde{\sigma}^2$ es un estimador consistente porque

$$\lim_{n\to 0}\frac{n-2}{n} = 1 \tag{2-149}$$

3 EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE: ESTIMACIÓN Y PROPIEDADES

3.1 El modelo de regresión lineal múltiple

El modelo de regresión lineal simple no es adecuado para modelizar muchos fenómenos económicos, ya que para explicar una variable económica se requiere en general tener en cuenta más de un factor. Veamos algunos ejemplos.

En la función keynesiana clásica el consumo se hace depender de la renta disponible como única variable relevante:

$$cons = \beta_1 + \beta_2 renta + u \tag{3-1}$$

Sin embargo, hay otros factores que pueden considerarse relevantes en el comportamiento del consumidor. Uno de esos factores podría ser la riqueza. Con la inclusión de este factor se tendrá un modelo con dos variables explicativas:

$$cons = \beta_1 + \beta_2 inc + \beta_3 riqueza + u \tag{3-2}$$

En el análisis de la producción se utilizan a menudo las funciones potenciales, que con una especificación adecuada pueden ser transformadas (tomando logaritmos naturales, en este caso) en modelos lineales en los parámetros. Utilizando un solo input (*trabajo*), un modelo para explicar el *output* se especifica del siguiente modo:

$$\ln(output) = \beta_1 + \beta_2 \ln(trabajo) + u \tag{3-3}$$

El modelo anterior es claramente insuficiente para el análisis económico. Sería mejor utilizar el conocido modelo de Cobb-Douglas, en el que se consideran dos inputs primarios (*trabajo* y *capita*l):

$$\ln(output) = \beta_1 + \beta_2 \ln(trabajo) + \beta_3 \ln(capital) + u \tag{3-4}$$

De acuerdo con la teoría microeconómica, los costes totales (*costot*) se expresan como una función de la cantidad producida (*cantprod*). Una primera aproximación para explicar el coste total podría ser un modelo con un único regresor:

$$costot = \beta_1 + \beta_2 cantprod + u \tag{3-5}$$

Sin embargo, es muy restrictivo considerar que, como sería el caso del modelo anterior, el coste marginal permanece constante, independientemente de la cantidad producida. En la teoría económica se propone, una función cúbica, lo que conduce al siguiente modelo econométrico:

$$costot = \beta_1 + \beta_2 cantprod + \beta_3 cantprod^2 + \beta_4 cantprod^3 + u \tag{3-6}$$

En este caso, a diferencia de los anteriores, en el modelo sólo hay una variable explicativa, pero que da lugar a tres regresores.

Los salarios se determinan por diferentes factores. Un modelo relativamente simple para explicar los salarios en función de los años de educación y de los años de experiencia es el siguiente:

$$salarios = \beta_1 + \beta_2 educ + \beta_3 exper + u \qquad (3\text{-}7)$$

De todos modos, otros factores importantes para explicar los salarios pueden ser variables cuantitativas tales como el tiempo de formación y la edad, o variables cualitativas, como el sexo, la rama de actividad, etc.

Por último, para explicar los gastos en consumo de pescado los factores relevantes pueden ser su precio, el precio de un producto sustitutivo como la carne, y la renta disponible. Es decir:

$$gastopescado = \beta_1 + \beta_2 preciopescado + \beta_3 preciocarne + \beta_4 renta + u \qquad (3\text{-}08)$$

Por lo tanto, los ejemplos anteriores han puesto de relieve la necesidad de utilizar modelos de regresión múltiple. El tratamiento econométrico del modelo de regresión lineal simple se hizo utilizando álgebra ordinaria. El tratamiento de un modelo econométrico de dos variables explicativas mediante el uso de álgebra ordinaria es tedioso y engorroso; por otra parte, un modelo con tres variables explicativas es prácticamente intratable con esta herramienta. Por esta razón, el modelo de regresión se va a presentar utilizando álgebra matricial.

3.1.1 Modelo de regresión poblacional y función de regresión poblacional

En el modelo de regresión lineal múltiple, el regresando -que puede ser la variable endógena o una transformación de las variables endógenas-, es una función lineal de *k* regresores correspondientes a las variables explicativas -o a transformaciones de las mismas- y una perturbación aleatoria o error. El modelo también incluye un término independiente. Si designamos por *y* al regresando, por x_2, x_3,..., x_k a los regresores y por *u* al error o perturbación aleatoria, el modelo poblacional de regresión lineal múltiple vendrá dado por la siguiente expresión:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (3\text{-}9)$$

Los parámetros $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ son fijos y desconocidos.

En el segundo miembro de (3-9) se pueden distinguir dos componentes: un componente sistemático $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$ y la perturbación aleatoria *u*. Llamando μ_y al componente sistemático, podemos escribir:

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \qquad (3\text{-}10)$$

Esta ecuación es conocida como función de regresión poblacional (*FRP*) o *hiperplano poblacional*. Cuando *k*=2, la *FRP* es específicamente una línea recta, cuando *k*=3, la *FRP* es específicamente un plano y, por último, cuando *k*>3, la *FRP* es denominada genéricamente hiperplano, que no es susceptible de ser representado físicamente.

De acuerdo con (3-10), μ_y es una función lineal en los parámetros $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$. Ahora, supongamos que tenemos una muestra aleatoria de tamaño *n*,

$\{(y_i, x_{2i}, x_{31}, \cdots, x_{ki}): i = 1, 2, \cdots, n\}$, extraída de la población estudiada. Si expresamos el modelo poblacional para todas las observaciones de la muestra, se obtiene el siguiente el sistema:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \cdots + \beta_k x_{k1} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \cdots + \beta_k x_{k2} + u_2 \\ \cdots & \quad \cdots \quad \cdots \quad \cdots \\ y_n &= \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \cdots + \beta_k x_{kn} + u_n \end{aligned} \tag{3-11}$$

El anterior sistema de ecuaciones puede expresarse de una forma más compacta usando la notación matricial. Así, vamos a denominar

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{22} & x_{32} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix}$$

La matriz **X** es la matriz de regresores. Entre los regresores también se incluye el regresor correspondiente al término independiente. Este regresor, que a menudo se denomina regresor *ficticio*, toma el valor 1 para todas las observaciones.

El modelo de regresión lineal múltiple (3-11) expresado en notación matricial es el siguiente:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{22} & x_{32} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \tag{3-12}$$

Si se tiene en cuenta las denominaciones dadas a vectores y matrices, el modelo de regresión lineal múltiple puede ser expresado de forma compacta de la siguiente manera:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3-13}$$

donde, de acuerdo con la notación utilizada, **y** es un vector $n \times 1$, **X** es una matriz $n \times k$, $\boldsymbol{\beta}$ es un vector $k \times 1$ y **u** es un vector $n \times 1$.

3.1.2 Función de regresión muestral

La idea básica de la regresión consiste en estimar los parámetros poblacionales $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$, a partir de una muestra dada.

La *función de regresión muestral* (*FRM*) es la contrapartida de la función de regresión poblacional (*FRP*). Dado que *FRM* se obtiene de una muestra dada, una nueva muestra generará diferentes estimaciones.

La *FRM*, que es una estimación de la *FRP*, que viene dada por

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \tag{3-14}$$

nos permite calcular el *valor ajustado* ($\hat{y}_i$) correspondiente a cada y_i. En la *FRM*, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ son los estimadores de los parámetros $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$.

Se denomina residuo a la diferencia entre y_i e $\hat{y}_i$. Esto es

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \tag{3-15}$$

En otras palabras, el residuo $\hat{u}_i$ es la diferencia entre un valor muestral y su correspondiente valor ajustado.

El sistema de ecuaciones (3-14) puede expresarse de una forma más compacta utilizando notación matricial. Así, vamos a denotar

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad \hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \cdots \\ \hat{u}_n \end{bmatrix}$$

El modelo ajustado correspondiente, para todas las observaciones de la muestra, será el siguiente:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-16}$$

El vector de los residuos es igual a la diferencia entre el vector de valores observados y el vector de valores ajustados, es decir,

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-17}$$

3.2 Obtención de estimaciones de mínimos cuadrados, interpretación de los coeficientes, y otras características

3.2.1 Obtención de estimadores *MCO*

Denominando *S* a la suma de los cuadrados de los residuos,

$$S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]^2 \tag{3-18}$$

para aplicar el criterio de mínimos cuadrados en el modelo de regresión lineal múltiple, calculamos la primera derivada de *S* con respecto a cada $\hat{\beta}_j$ en la expresión (3-18):

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_1} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-1] \\
\frac{\partial S}{\partial \hat{\beta}_2} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{2i}] \\
\frac{\partial S}{\partial \hat{\beta}_3} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{3i}] \\
\cdots \quad & \quad \cdots \quad \quad \cdots \quad \quad \cdots \\
\frac{\partial S}{\partial \hat{\beta}_k} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{ki}]
\end{aligned} \tag{3-19}$$

Los estimadores de mínimos cuadráticos se obtienen al igualar a 0 las derivadas anteriores:

$$\begin{aligned}
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] = 0 \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{2i} = 0 \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{3i} = 0 \\
&\cdots \quad \cdots \quad \quad \cdots \quad \quad \cdots \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{ki} = 0
\end{aligned} \tag{3-20}$$

o, con notación matricial,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{3-21}$$

Al sistema anterior se le denomina genéricamente sistema de ecuaciones normales del *hiperplano*.

En notación matricial ampliada, el sistema de ecuaciones normales es el siguiente:

$$\begin{bmatrix}
n & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\
\sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{ki} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2
\end{bmatrix}
\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} =
\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix} \tag{3-22}$$

Obsérvese que:

a) a) $\mathbf{X}'\mathbf{X}/n$ es la matriz de momentos muestrales de segundo orden, con respecto al origen, de los regresores, entre los cuales se incluye el regresor ficticio (x_{1i}) asociado al término independiente, que toma el valor x_{1i}=1 para todo i.

b) $\mathbf{X}'\mathbf{y}/n$ es el vector de momentos muestrales de segundo orden, con respecto al origen, entre el regresando y los regresores.

En este sistema hay *k* ecuaciones y *k* incógnitas $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k)$. Este sistema puede resolverse fácilmente utilizando álgebra matricial. Con el fin de resolver unívocamente el sistema (3-21) con respecto a $\hat{\boldsymbol{\beta}}$, es preciso que el rango de la matriz $\mathbf{X}'\mathbf{X}$ sea igual a *k*. Si esto se cumple, ambos miembros de (3-21) pueden ser premultiplicados por $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$:

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$$

obteniéndose la expresión del vector de estimadores de mínimos cuadrados, o más exactamente, el vector de estimadores de mínimos cuadrados ordinarios (*MCO*), porque $\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Por lo tanto, la solución es la siguiente:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} \tag{3-23}$$

Como la matriz de segundas derivadas, $2\mathbf{X}'\mathbf{X}$, es una matriz definida positiva, la conclusión es que *S* presenta un mínimo en $\hat{\boldsymbol{\beta}}$.

3.2.2 Interpretación de los coeficientes

El coeficiente $\hat{\beta}_j$ mide el efecto *parcial* del regresor x_i, manteniendo los otros regresores fijos. Vamos a ver el significado de esta expresión.

El modelo estimado para la observación *i*-ésima viene dado por

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_k x_{ki} \tag{3-24}$$

Consideremos ahora el modelo estimado para la observación *h*-ésima, en el que los valores de las variables explicativas y, en consecuencia, de *y* habrán cambiado con respecto a la (3-24):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_{2h} + \hat{\beta}_3 x_{3h} + \cdots + \hat{\beta}_j x_{jh} + \cdots + \hat{\beta}_k x_{kh} \tag{3-25}$$

Restando (3-25) de ((3-24), tenemos que

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2 + \hat{\beta}_3 \Delta x_3 + \cdots + \hat{\beta}_j \Delta x_j + \cdots + \hat{\beta}_k \Delta x_k \tag{3-26}$$

donde $\Delta\hat{y} = \hat{y}_i - \hat{y}_h, \Delta x_2 = x_{2i} - x_{2h}, \Delta x_3 = x_{3i} - x_{3h}, \cdots \Delta x_k = x_{ki} - x_{kh}$.

La expresión anterior capta la variación de $\hat{y}$ debida a cambios en todos los regresores. Si sólo cambia x_j, tendremos que

$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j \tag{3-27}$$

Si x_k se incrementa en una unidad, tenemos

$$\Delta\hat{y} = \hat{\beta}_j \quad \text{for } \Delta x_j = 1 \tag{3-28}$$

En consecuencia, el coeficiente $\hat{\beta}_j$ mide el cambio en *y* cuando x_j aumenta en 1 unidad, *manteniendo fijos los regresores* $x_2, x_3, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k$. Es muy importante en la interpretación de los coeficientes tener en cuenta esta cláusula *ceteris paribus*.

Esta interpretación no es válida, por supuesto, para el término independiente.

EJEMPLO 3.1 Cuantificando la influencia de la edad y del salario sobre el absentismo en la empresa Buenosaires

Buenosaires es una empresa dedicada a la fabricación de ventiladores, habiendo tenido resultados relativamente aceptables en los últimos años. Los directivos consideran, sin embargo, que los resultados habrían sido mejores si el absentismo en la empresa no fuera tan alto. Para este propósito, el modelo que se propone es el siguiente:

$$absent = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4 wage + u$$

donde la ausencia, *absent,* se mide en días por año, el salario, *wage,* en miles de euros al año; los años trabajados en la empresa, *tenure,* y la edad, *age*, se expresan en años.

Utilizando una muestra de tamaño 48 (fichero *absent*), se ha estimado la siguiente ecuación:

$$\underset{}{absent} = \underset{(1.603)}{14.413} - \underset{(0.048)}{0.096}\, age - \underset{(0.067)}{0.078}\, tenure - \underset{(0.007)}{0.036}\, wage$$

$$R^2 = 0.694 \qquad n = 48$$

La interpretación de $\hat{\beta}_2$ es la siguiente: manteniendo fijo el salario y los años trabajados en la empresa, si la edad se incrementa en un año, el absentismo laboral se reducirá en 0.096 días al año. La interpretación de $\hat{\beta}_3$ es como sigue: manteniendo fijo el salario y la edad, si los años trabajados en la empresa se incrementan en un año, el absentismo laboral se reducirá en 0.078 días al año. Finalmente, la interpretación de $\hat{\beta}_4$ es la siguiente: manteniendo fija la edad y los años trabajados en la empresa, si el salario se incrementa en 1000 euros al año, el absentismo laboral se reducirá en 0.036 días por año.

EJEMPLO 3.2 Demanda de servicios hoteleros

Para explicar la demanda de servicios hoteleros se formuló el siguiente modelo:

$$\ln\ hostel\ = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + u \tag{3-29}$$

donde *hostel* es el gasto en servicios hoteleros e *inc* es la renta disponible; ambas variables están expresadas en euros por mes. La variable *hhize* es el número de miembros del hogar.

La ecuación estimada con una muestra de 40 hogares, utilizando el fichero *hostel*, es la siguiente:

$$\ln(hostel_i) = -27.36 + 4.442 \ln(inc_i) - 0.523 hhsize_i$$

$$R^2 = 0.738 \qquad n = 40$$

A la vista de estos resultados, podemos decir que los servicios hoteleros son un bien de lujo, ya que la elasticidad de la demanda/renta para este bien es muy alta (4.44). Esto significa que si la renta se incrementa en un 1%, el gasto en servicios hoteleros se incrementará un 4.44%, manteniendo fijo el tamaño de la familia. Por otro lado, si el tamaño del hogar aumenta en un miembro, entonces el gasto en servicios hoteleros disminuirá en un 52%.

EJEMPLO 3.3 Una regresión hedónica para coches

El modelo hedónico de medición de precios se basa en el supuesto de que el valor de un bien depende del valor de sus diferentes características. Así, el precio de un coche dependerá del valor que el comprador asigne a sus atributos: cualitativos (por ejemplo, cambio automático, potencia, diesel, dirección asistida, aire acondicionado), y cuantitativos (por ejemplo, consumo de combustible, peso, etc.). La base de datos para este ejercicio es el fichero *hedcarsp* (precios hedónicos de los coches en España) y cubre los años 2004 y 2005. Un primer modelo basado sólo en atributos cuantitativos es el siguiente:

$$\ln(price) = \beta_1 + \beta_2 volume + \beta_3 fueleff + u$$

donde *volume* es longitud×anchura×altura en m^3 y *fueleff* es la *ratio* litros por 100 km/caballos de vapor expresada en porcentaje.

La ecuación estimada con una muestra de 214 observaciones es la siguiente:

$$\ln(price)_i = \underset{(0.151)}{14.97} + \underset{(0.009)}{0.0956}\, volume_i - \underset{(0.010)}{0.1608}\, fueleff_i$$

$$R^2=0.765 \qquad n=214$$

La interpretación de $\hat{\beta}_2$ y $\hat{\beta}_3$ es la siguiente. Manteniendo *fueleff* fijo, si aumenta *volume* en 1 m^3, el precio de los coches se incrementará en un 9.56%. Manteniendo fijo *volume*, si la ratio litros por 100 km/caballos de vapor aumenta en un punto porcentual, el precio de los automóviles se reducirá en un 16,08%.

EJEMPLO 3.4. Ventas y publicidad: el caso de Lydia E. Pinkham

En este caso se va a estimar un modelo con datos de series temporales con objeto de medir el efecto que puedan tener los gastos de publicidad, realizados a lo largo de distintos períodos de tiempo, sobre las ventas del momento actual. Designando por V_t y P_t a las ventas y a los gastos en publicidad, realizados en el momento *t*, el modelo planteado inicialmente para explicar las ventas, en función de los gastos en publicidad presentes y pasados, es el siguiente:

$$V_t = \alpha + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 P_{t-2} + \cdots + u_t \tag{3-30}$$

En la expresión anterior los puntos suspensivos indican que los gastos en publicidad realizados en el pasado siguen ejerciendo influencia de forma indefinida, aunque, se supone, que con un impacto decreciente sobre las ventas del momento actual. Naturalmente, el modelo anterior no es operativo, ya que tiene un número indefinido de coeficientes. Para solucionar el problema se pueden adoptar, en principio, dos enfoques. El primer enfoque consiste en fijar *a priori* el número máximo de períodos durante los cuales la publicidad mantiene sus efectos sobre las ventas. En el segundo enfoque, se postula que los coeficientes se comportan de acuerdo con alguna ley que permite determinar su valor en función de un número reducido de parámetros, posibilitando además una ulterior simplificación.

En el primer enfoque el problema que surge es que en general no existen criterios precisos e información suficiente que permitan la fijación *a priori* del número máximo de períodos. Por esta razón, vamos a ver un caso particular del segundo enfoque que tiene gran interés por la plausibilidad del supuesto y su fácil aplicación. El caso que vamos a considerar consiste en establecer que los coeficientes β_i disminuyen de forma geométrica a medida que nos alejamos hacia atrás en el tiempo según el esquema:

$$\beta_i = \beta_1 \lambda^i \quad \forall i \qquad 0 < \lambda < 1 \tag{3-31}$$

A la anterior transformación se le denomina transformación de Koyck, ya que fue este autor el que la introdujo en 1954 para el estudio de la inversión.

Sustituyendo (3-31) en (3-30), se obtiene que

$$V_t = \alpha + \beta_1 P_t + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \cdots + u_t \tag{3-32}$$

El modelo anterior sigue teniendo infinitos términos, pero sólo tres parámetros y además se puede simplificar. En efecto, si expresamos la ecuación (3-32) para el período *t*-1 y multiplicamos ambos miembros por λ se obtiene que

$$\lambda V_{t-1} = \alpha\lambda + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \beta_1 \lambda^3 P_{t-3} + \cdots + \lambda u_{t-1} \tag{3-33}$$

Restando miembro a miembro (3-33) de (3-32), y teniendo en cuenta que los factores λ^i tienden a 0 al tender *i* a infinito, se llega al siguiente resultado:

$$V_t = \alpha(1-\lambda) + \beta_1 P_t + \lambda V_{t-1} + u_t - \lambda u_{t-1} \tag{3.34}$$

El modelo ha quedado simplificado de forma que solamente tiene tres regresores, aunque, a cambio, se ha pasado a un término de perturbación compuesto. Antes de ver la aplicación de este modelo se va analizar el significado del coeficiente λ y el problema de la duración de los efectos de los gastos en publicidad sobre las ventas. El parámetro λ es la tasa a que decaen los efectos de los gastos en publicidad

presentes sobre las ventas presentes y futuras. Los efectos acumulados del gasto en publicidad de una unidad monetaria sobre las ventas después de transcurridos *m* períodos vienen dados por

$$\beta_1(1+\lambda+\lambda^2+\lambda^3+\cdots+\lambda^m) \tag{3-35}$$

Para calcular la suma acumulada de los efectos, dada en (3-35), vamos a tener en cuenta que esta expresión es la suma de los términos de una progresión geométrica[2], con lo que se puede expresar de la siguiente forma:

$$\frac{\beta_1(1-\lambda^m)}{1-\lambda} \tag{3-36}$$

Cuando *m* tiende a infinito, entonces la suma de los efectos acumulados viene dada por

$$\frac{\beta_1}{1-\lambda} \tag{3-37}$$

Una cuestión interesante es determinar cuántos períodos de tiempo se requieren para que se obtenga el *p*% (por ejemplo, el 50%) del efecto total. Designando por *h* el número de períodos requeridos para obtener dicho porcentaje, se puede establecer que

$$p = \frac{\text{Efecto en } h \text{ periodos}}{\text{Efecto total}} = \frac{\dfrac{\beta_1(1-\lambda^h)}{1-\lambda}}{\dfrac{\beta_1}{1-\lambda}} = 1-\lambda^h \tag{3-38}$$

Fijado *p* se puede calcular *h* de acuerdo con (3-38). En efecto, despejando *h* en esta expresión se obtiene que

$$h = \frac{\ln(1-p)}{\ln\lambda} \tag{3-39}$$

Este modelo lo utilizó Kristian S. Palda en su tesis doctoral publicada en 1964, titulada *The Measurement of Cumulative Advertising Effects*, para analizar los efectos acumulados de los gastos en publicidad, en el caso de la compañía Lydia E. Pinkham. Este caso, así como el estudio de Palda, han sido la base a partir de la cual se ha desarrollado la investigación de los efectos de los gastos en publicidad. Vamos a ver a continuación algunas características de este caso:

1) La Lydia E. Pinkham Medicine Company fabricaba desde 1873 un extracto de hierbas diluido en una solución alcohólica. Este producto se anunciaba inicialmente como un analgésico y también como un remedio para una enorme variedad de enfermedades.

2) En general, en los distintos tipos de productos suele haber competencia entre distintas marcas, como pueda ser el caso paradigmático de la Coca-Cola y la Pepsi-Cola en el campo de las colas. Cuando esto ocurre, para analizar los efectos de los gastos en publicidad hay tener en cuenta el comportamiento de los principales competidores. Lydia E. Pinkham tenía la ventaja de no tener competidores, y de que, en su línea de producto, actuaba en la práctica como monopolista.

3) Otra característica del caso Lydia E. Pinkham era que la mayor parte de los gastos de distribución se asignaban a la publicidad, ya que la compañía no tenía agentes comerciales, siendo muy elevada la relación gastos en publicidad/ventas.

4) El producto pasó a lo largo del tiempo por distintos avatares. Así, en 1914 la Food and Drug Administration (organismo de los Estados Unidos que establece los controles para los productos alimenticios y los medicamentos) le acusó de publicidad engañosa por lo que tuvo que cambiar sus mensajes publicitarios. También la Internal Revenue (oficina de impuestos) le amenazó con aplicarle una

[2] Designando por a_p, a_u y r al primer término, al último término y a la razón respectivamente, la suma de los términos de una progresión geométrica convergente viene dada por

$$\frac{a_p - a_u}{1-r}$$

tasa sobre bebidas alcohólicas, ya que el contenido alcohólico del producto era del 18%. Por todos estos motivos se produjeron cambios en la presentación y contenido durante el período 1915-1925. En 1925 la Food and Drug Administration prohibió que el producto se anunciara como medicina, pasando a distribuirse como bebida tónica. En el período 1926-1940 se incrementaron considerablemente los gastos en publicidad para después decaer.

La estimación del modelo (3-34) con datos desde 1907 a 1960, recogidos en el fichero *pinkham*, es la siguiente:

$$ventas_t = \underset{(95.7)}{138.7} + \underset{(0.156)}{0.3288}\, gpub_t + \underset{(0.0915)}{0.7593}\, ventas_{t-1}$$

$$R^2=0.877 \qquad n=53$$

La suma de los efectos acumulados de los gastos en publicidad sobre las ventas se obtiene aplicando la fórmula (3-37):

$$\frac{\hat{\beta}_1}{1-\hat{\lambda}} = \frac{0.3288}{1-0.7593} = 1.3660$$

De acuerdo con dicho resultado, por cada unidad monetaria adicional gastada en publicidad se obtiene un efecto acumulado total en las ventas de 1.366 unidades monetarias. Dado que es importante no solo determinar el efecto total, sino también como se distribuyen estos efectos a lo largo del tiempo, vamos a contestar ahora a la siguiente pregunta: ¿Cuantos períodos de tiempo se requieren para alcanzar la mitad de los efectos totales? Aplicando la fórmula (3-39) para el caso de p=0,5, se obtiene el siguiente resultado:

$$\hat{h}(0.5) = \frac{\ln(1-0.5)}{\ln(0.7593)} = 2.5172$$

3.2.3 Implicaciones algebraicas de la estimación

Las implicaciones algebraicas de la estimación se derivan exclusivamente de la aplicación del método de *MCO* al modelo de regresión lineal múltiple:

1. *La suma de los residuos de MCO es igual a 0*:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{3-40}$$

De la definición de residuos

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \tag{3-41}$$

Si sumamos para las n observaciones, obtenemos:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \tag{3-42}$$

Por otro lado, la primera ecuación del sistema de ecuaciones normales (3-20) es igual a

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = 0 \tag{3-43}$$

Si comparamos (3-42) y (3.43), llegamos a la conclusión de que (3-40) se cumple.

Tenga en cuenta que, si (3-40) se cumple, esto implica que

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} \hat{y}_i \tag{3-44}$$

y, dividiendo (3-40) y (3-44) por *n*, obtenemos

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \tag{3-45}$$

2. *El hiperplano MCO pasa siempre a través del punto de medias muestrales* $(\bar{y}, \bar{x}_2, \cdots, \bar{x}_k)$.

En efecto, dividiendo la ecuación (3-43) por *n* se tiene que:

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k \tag{3-46}$$

3. El producto cruzado muestral entre cada uno de los regresores y los residuos MCO es cero.

Es decir,

$$\sum_{i=1}^{n} x_{ji} \hat{u}_i = 0 \qquad\qquad j = 2, 3, \cdots, k \tag{3-47}$$

Utilizando las últimas *k* ecuaciones normales (3-20) y teniendo en cuenta que, por definición $\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}$, podemos ver que

$$\begin{aligned} &\sum_{i=1}^{n} \hat{u}_i x_{2i} = 0 \\ &\sum_{i=1}^{n} \hat{u}_i x_{3i} = 0 \\ &\cdots \qquad\qquad \cdots \\ &\sum_{i=1}^{n} \hat{u}_i x_{ki} = 0 \end{aligned} \tag{3-48}$$

4. *El producto cruzado muestral entre los valores ajustados ($\hat{y}$) y los residuos MCO es cero.*

Es decir,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{3-49}$$

Teniendo en cuenta (3-40) y (3-48), obtenemos

$$\begin{aligned} &\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}) \hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} \hat{u}_i + \cdots \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \hat{u}_i \\ &= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 + \cdots \hat{\beta}_k \times 0 = 0 \end{aligned} \tag{3-50}$$

3.3 Supuestos y propiedades estadísticas de los estimadores de *MCO*

Antes de estudiar las propiedades estadísticas de los estimadores de *MCO* en el modelo de regresión lineal múltiple, necesitamos formular un conjunto de supuestos estadísticos. Específicamente, al conjunto de supuestos que vamos a formular se les denomina *supuestos del modelo lineal clásico* (*MLC*). Es importante destacar que los

supuestos estadísticos del *MLC* son muy simples, y que los estimadores de *MCO* tienen, bajo estos supuestos, muy buenas propiedades.

3.3.1 Supuestos estadísticos del *MLC* en la regresión lineal múltiple

a) Supuesto sobre la forma funcional

1) La relación entre el regresando, los regresores y el error es lineal en los parámetros:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{3-51}$$

o, alternativamente, para todas las observaciones,

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \tag{3-52}$$

b) Supuestos sobre los regresores

2) Los valores de $x_2, x_3, \cdots, x_k$ *son fijos en repetidas muestras, o la matriz* **X** *es fija en repetidas muestras*:

Este es un supuesto fuerte en el caso de las ciencias sociales, donde, en general, no es posible experimentar. Una hipótesis alternativa puede formularse así:

2) Los regresores* $x_2, x_3, \cdots, x_k$ *se distribuyen independientemente de la perturbación aleatoria.* Formulado de otra manera, **X** se distribuye de forma independiente del vector de perturbaciones aleatorias, lo que implica que $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$

Como hicimos en el capítulo 2, vamos a adoptar también el supuesto 2).

3) La matriz de regresores, **X**, *no contiene errores de medición.*

4) La matriz de regresores, **X**, *tiene rango igual a k*:

$$\rho(\mathbf{X}) = k \tag{3-53}$$

Recordemos que la matriz de regresores contiene *k* columnas, correspondientes a los *k* regresores del modelo, y *n* filas, correspondientes al número de observaciones. El supuesto 4 tiene dos implicaciones:

1. El número de observaciones, *n*, debe ser igual o mayor que el número de regresores, *k*. Intuitivamente, esto tiene sentido: para estimar *k* parámetros, necesitamos al menos *k* observaciones
2. Cada regresor debe ser linealmente independiente, lo que implica que no existen relaciones lineales exactas entre los regresores. Si un regresor es una combinación lineal exacta de otros regresores, entonces se dice que hay *multicolinealidad perfecta*, y el modelo no puede estimarse.

Si existe una relación lineal aproximada - es decir, no una relación exacta -, entonces se pueden estimar los parámetros, aunque su fiabilidad se verá afectada. En este caso, se dice que existe *multicolinealidad no perfecta*.

c) Supuesto sobre los parámetros

5) Los parámetros $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ *son constantes, o* $\boldsymbol{\beta}$ *es un vector constante*

d) Supuestos sobre la perturbación aleatoria

6) La media de las perturbaciones es cero,

$$E(u_i) = 0, \quad i = 1,2,3,\ldots,n \quad \text{o} \quad E(\mathbf{u}) = \mathbf{0} \tag{3-54}$$

7) Las perturbaciones tienen una varianza constante (supuesto de homoscedasticidad):

$$var(u_i) = \sigma^2 \qquad i = 1,2,\cdots n \tag{3-55}$$

8) Las perturbaciones con diferentes subíndices no están correlacionadas entre sí (supuesto de no autocorrelación):

$$E(u_i u_j) = 0 \qquad i \neq j \tag{3-56}$$

La formulación de los supuestos de homoscedasticidad y no autocorrelación permite especificar la matriz de covarianzas del vector de perturbaciones:

$$\begin{aligned} E\left[\left[\mathbf{u} - E(\mathbf{u})\right]\left[\mathbf{u} - E(\mathbf{u})\right]'\right] &= E\left[\left[\mathbf{u} - \mathbf{0}\right]\left[\mathbf{u} - \mathbf{0}\right]'\right] = E\left[\left[\mathbf{u}\right]\left[\mathbf{u}\right]'\right] \\ = E\left[\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}\right] &= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix} \\ = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{bmatrix} &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \end{aligned} \tag{3-57}$$

Para obtener la última igualdad se ha tenido en cuenta que la varianza de cada elemento es constante e igual a σ^2, de acuerdo con (3-55), y que la covarianza entre cada par de elementos es 0, de acuerdo con (3-56).

El resultado anterior puede expresarse de forma compacta del siguiente modo:

$$E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I} \tag{3-58}$$

A la matriz dada en (3-58) se le denomina matriz e*scalar*, puesto que es un escalar (σ^2, en este caso) multiplicado por la matriz identidad.

9) La perturbación u tiene una distribución normal

Teniendo en cuenta los supuestos 6 a 9, tenemos que

$$u_i \sim NID(0, \sigma^2) \quad i = 1,2,\cdots,n \qquad \text{o} \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3-59}$$

donde el *NID* significa que la perturbación está *normal e independiente distribuida*.

3.3.2 Propiedades estadísticas del estimador de *MCO*

Bajo los supuestos del *MLC*, el estimador de *MCO* poseen buenas propiedades. En las demostraciones de este apartado implícitamente se tendrán en cuenta siempre los supuestos 3, 4 y 5.

Linealidad e insesgadez del estimador de MCO

Ahora, vamos a demostrar que el estimador de *MCO* es linealmente insesgado. En primer lugar expresaremos $\hat{\boldsymbol{\beta}}$ como una función del vector **u**, utilizando el supuesto 1, de acuerdo con (3-52):

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\left[\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right] = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u} \tag{3-60}$$

El estimador de *MCO* puede expresarse del siguiente modo con el fin de ver de forma más clara la propiedad de linealidad:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta}+\mathbf{A}\mathbf{u} \tag{3-61}$$

donde $\mathbf{A} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'$ es fija bajo el supuesto 2. Así pues, $\hat{\boldsymbol{\beta}}$ es una función lineal de **u** y, consecuentemente, es un estimador *lineal*.

Tomando las esperanzas en (3-60) y aplicando el supuesto 6, se obtiene

$$E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E\left[\mathbf{u}\right] = \boldsymbol{\beta} \tag{3-62}$$

Por lo tanto, $\hat{\boldsymbol{\beta}}$ es un estimador *insesgado*.

Varianza del estimador de MCO

Para calcular la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$ son necesarios los supuestos 7 y 8, además de los 6 primeros:

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= E\left[\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})\right]' = E\left[\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right]\left[\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right]' \\ &= E\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} \\ &= \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\sigma^2\mathbf{I})\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}\end{aligned} \tag{3-63}$$

En el tercer paso de la demostración anterior se ha tenido en cuenta que, de acuerdo con (3-60), $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}$. El supuesto 2 se ha tenido en cuenta en el cuarto paso. Finalmente, los supuestos 7 y 8 se han utilizado en el último paso.

Por lo tanto, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}$ es la matriz de covarianzas del vector $\hat{\boldsymbol{\beta}}$. En esta matriz de covarianzas, la varianza de cada elemento $\hat{\beta}_j$ aparece en la diagonal principal, mientras que las covarianzas entre cada par de elementos se encuentran fuera de la diagonal principal. Específicamente, la varianza de $\hat{\beta}_j$ (para j=2,3,…,k) es igual a σ^2 multiplicada por el elemento correspondiente de la diagonal principal de $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$. Después de operar, la varianza de $\hat{\beta}_j$ puede expresarse como

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{nS_j^2(1-R_j^2)} \tag{3-64}$$

donde R_j^2 es el *R* cuadrado de la regresión de cada x_j sobre el resto de regresores, *n* es el tamaño de la muestra y S_j^2 es la varianza muestral del regresor x_j.

La fórmula (3-64) es válida para todos los coeficientes de pendiente, pero no para el término independiente.

A la raíz cuadrada de (3-64) se le denomina *desviación estándar* (*de*) de $\hat{\beta}_j$:

$$de(\hat{\beta}_j) = \frac{\sigma}{\sqrt{nS_j^2(1-R_j^2)}} \tag{3-65}$$

Los estimadores de MCO son ELIO

Bajo los supuestos 1 a 8 del *MLC*, que son denominados supuestos de Gauss-Markov, los estimadores de *MCO* son *estimadores lineales insesgados y óptimos (ELIO).*

El teorema de Gauss Markov establece que los estimadores *MCO* son estimadores óptimos dentro la clase de los estimadores lineales insesgados. En este contexto *óptimo*, significa que es un estimador con la varianza más pequeña para un determinado tamaño de muestra. Vamos ahora a comparar la varianza de un elemento de $\hat{\boldsymbol{\beta}}$ ($\hat{\beta}_j$), con cualquier otro estimador $\tilde{\beta}_j$ que sea lineal (tal que $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$) e insesgado (de forma que los pesos, w_j, deben cumplir algunas restricciones). La propiedad de que $\hat{\beta}_j$ es un estimador *ELIO* tiene las siguientes implicaciones al comparar su varianza con la varianza de $\tilde{\beta}_j$:

1) La varianza de un coeficiente $\tilde{\beta}_j$ o es mayor que, o igual a, la varianza de $\hat{\beta}_j$ obtenido por *MCO*:

$$\text{var}(\tilde{\beta}_j) \geq \text{var}(\hat{\beta}_j) \qquad \text{j} = 1, 2, \ldots, k \tag{3-66}$$

2) La varianza de cualquier combinación lineal de $\tilde{\beta}_j$ es mayor que, o igual a, la varianza de la correspondiente combinación lineal de $\hat{\beta}_j$.

En el apéndice 3.1 puede verse la demostración del teorema de Gauss-Markov.

Estimador de la varianza de la perturbación

Teniendo en cuenta el sistema de ecuaciones normales (3-20), si conocemos *n–k* residuos, podemos obtener los otros k residuos utilizando las restricciones que impone ese sistema a los residuos.

Por ejemplo, la primera ecuación normal nos permite obtener el valor de $\hat{u}_n$ en función de los residuos restantes:

$$\hat{u}_n = -\hat{u}_1 - \hat{u}_2 - \cdots - \hat{u}_{n-1}$$

Por lo tanto, sólo hay *n-k* grados de libertad en los residuos de *MCO*, a diferencia de los *n* grados de libertad en las perturbaciones. Recuerde que los grados de libertad se definen como la diferencia entre el número de observaciones y el número de parámetros estimados.

El estimador insesgado de σ^2 se ajusta teniendo en cuenta los grados de libertad:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k} \tag{3-67}$$

Bajo los supuestos 1 a 8, se obtiene que

$$E(\hat{\sigma}^2) = \sigma^2 \tag{3-68}$$

Véase el apéndice 3.2 para la demostración.

A la raíz cuadrada de (3-67), $\hat{\sigma}$ se le denomina *error estándar de la regresión* y es un estimador de σ.

Estimadores de la varianzas de $\hat{\boldsymbol{\beta}}$ **y *del coeficiente de pendiente*** $\hat{\beta}_j$

El estimador de la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$ viene dado por

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_1) & Cov(\hat{\beta}_1,\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_k) \\ Cov(\hat{\beta}_2,\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_k) \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ Cov(\hat{\beta}_j,\hat{\beta}_1) & Cov(\hat{\beta}_j,\hat{\beta}_2) & \cdots & \text{var}(\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_j,\hat{\beta}_k) \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ Cov(\hat{\beta}_k,\hat{\beta}_1) & Cov(\hat{\beta}_k,\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_k,\hat{\beta}_j) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix} \tag{3-69}$$

La varianza del coeficiente de la pendiente $\hat{\beta}_j$, dada en (3-64), es una función del parámetro desconocido σ^2. Cuando σ^2 se sustituye por su estimador $\hat{\sigma}^2$, se obtiene un estimador de la varianza de $\hat{\beta}_j$:

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{3-70}$$

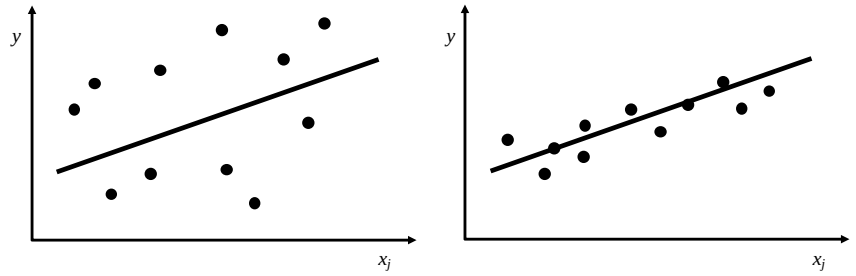
De acuerdo con la expresión anterior, el estimador de la varianza de $\hat{\beta}_j$ viene afectado por los siguientes factores:

a) Cuanto mayor es $\hat{\sigma}^2$, mayor es la varianza del estimador. Esto no es sorprendente en absoluto: cuanto más "ruido" exista en la ecuación, y, en consecuencia, más grande será $\hat{\sigma}^2$, con lo que será más difícil estimar con precisión el efecto parcial de cualquier regresor sobre *y*. (Véase figura 3.1).

b) A medida que se incrementa el tamaño de la muestra, la varianza del estimador se reduce.

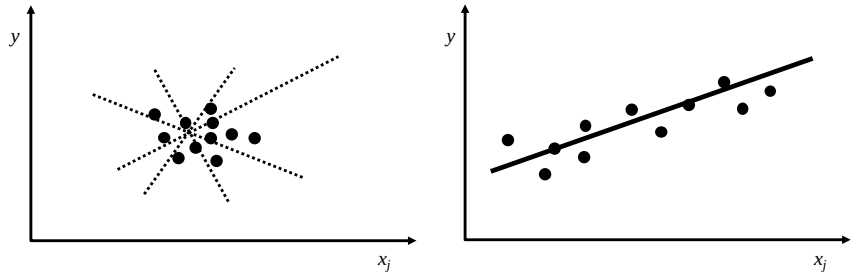
c) Cuanto más pequeña sea la varianza muestral de un regresor, mayor es la variación del coeficiente correspondiente. Manteniendo los demás factores igual, para estimar β_j es preferible que la variación muestral de x_j sea lo más grande posible, tal como se ilustra en la figura 3.2. Como se puede ver hay muchas líneas hipotéticas que podrían ajustarse a los datos cuando la varianza muestral de x_j, (S_j^2), es pequeña como puede verse en la parte a) de la figura. En cualquier caso, no está permitido por el supuesto 4 que S_j^2=0.

d) Cuanto mayor sea R_j^2, (es decir, cuanto mayor sea la correlación del regresor *j*-ésimo con el resto de los regresores), mayor será la varianza de $\hat{\beta}_j$.



a) $\hat{\sigma}^2$ grande b) $\hat{\sigma}^2$ pequeña

FIGURA 3.1. Influencia de $\hat{\sigma}^2$ sobre el estimador de la varianza.



a) S_j^2 pequeño b) S_j^2 grande

FIGURA 3.2. Influencia de S_j^2 sobre el estimador de la varianza.

A la raíz cuadrada de (3-70) se le denomina *error estándar* (*ee*) de $\hat{\beta}_j$:

$$ee(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{nS_j^2(1-R_j^2)}} \tag{3-71}$$

Otras propiedades de los estimadores MCO

Bajo los supuestos 1 a 6 del *MLC*, el estimador de *MCO*, $\hat{\boldsymbol{\beta}}$, es consistente, como puede verse en el apéndice 3.3, *asintótica y normalmente distribuido*, y también *asintóticamente eficiente* dentro de la clase de los estimadores consistentes y asintóticamente normales.

Bajo los supuestos 1 a 9 del *MLC*, el estimador *MCO* es también el estimador de *máxima verosimilitud* (*MV*), como se prueba en el apéndice 3.4, y es el *estimador de*

mínima varianza insesgado (*EMVI*). Esto último significa que el estimador de *MCO* tiene la menor varianza entre todos los estimadores insesgados, sean lineales o no.

3.4 Más sobre formas funcionales

En este apartado vamos a examinar dos temas sobre formas funcionales: el uso de los logaritmos en modelos econométricos y las funciones polinomiales.

3.4.1 Utilización de logaritmos en los modelos econométricos

Algunas variables se utilizan a menudo en forma logarítmica. Así es en el caso de las variables monetarias que, en general, son positivas o de otras variables con valores elevados. La utilización de modelos con transformaciones logarítmicas tiene además sus ventajas. Una de ellas es que los coeficientes tienen interpretaciones atractivas (elasticidades o semi-elasticidades). Otra es la invariancia de los coeficientes de pendiente cuando hay cambios de escala en las variables. Tomar logaritmos puede ser conveniente debido a que reduce el rango de las variables, lo que hace que las estimaciones sean menos sensibles a los valores extremos de las variables. Los supuestos del *MLC* se satisfacen más a menudo en modelos que aplican logaritmos a la variable endógena, que en los modelos que no aplican ninguna transformación. Así, sucede que la distribución condicional de y es frecuentemente heteroscedástica, mientras que $\ln(y)$ puede ser homoscedástica.

Una limitación de la transformación logarítmica es que no se puede utilizar cuando la variable original toma valores cero o negativos. Por otro lado, algunas variables que se miden en años y en otras variables que son una proporción o un porcentaje, se utiliza la variable original sin ninguna transformación.

3.4.2 Funciones polinomiales

Las funciones polinomiales se han utilizado ampliamente en la investigación econométrica. Cuando en el modelo solo hay regresores correspondientes a una función polinomial decimos que es un *modelo polinomial*. La forma general del *modelo polinomial de grado k* puede expresarse como

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_k x^k + u \tag{3-72}$$

Funciones cuadráticas

Un caso interesante de funciones polinomiales es la *función cuadrática*, que es una *función polinomial de segundo grado*. Cuando hay sólo regresores correspondientes a la función cuadrática, tenemos un *modelo cuadrático*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + u \tag{3-73}$$

Las funciones cuadráticas se utilizan muy a menudo en economía aplicada para captar la disminución o el aumento de los efectos marginales. Es importante observar que, en tal caso, β_2 no mide el cambio en y con respecto a x, porque no tiene sentido mantener x^2 fijo, mientras cambia x. El efecto marginal de x sobre y, que depende linealmente del valor de x, es el siguiente:

$$em = \frac{dy}{dx} = \beta_2 + 2\beta_3 x \tag{3-74}$$

En una aplicación específica, este efecto marginal se evaluará para valores específicos de *x*. Si β_2 y β_3 tienen signos opuestos el punto de cambio está situado en

$$x^* = -\frac{\beta_2}{2\beta_3} \tag{3-75}$$

Si $\beta_2>0$ y $\beta_3<0$, el efecto marginal de *x* sobre y es positivo al principio, pero será negativo cuando *x* sea mayor que x^*. Si $\beta_2<0$ y $\beta_3>0$, el efecto marginal de *x* sobre *y* es negativo al principio, pero será positivo para *x* mayor que x^*.

Ejemplo 3.5 Salarios y años de antigüedad en la empresa

Utilizando los datos de *ceosal2* para estudiar el tipo de relación entre el salario (*salary)* de los consejeros delegados (CEO) en Estados Unidos y los años de permanencia en la empresa como CEO de la compañía (*ceoten*), se estimó el siguiente modelo:

$$\ln(salary) = \underset{(0.086)}{6.246} + \underset{(0.0001)}{0.0006}\, profits + \underset{(0.0156)}{0.0440}\, ceoten - \underset{(0.00052)}{0.0012}\, ceoten^2$$

$$R^2=0.1976 \qquad n=177$$

donde los beneficios de las compañías (*profits*) están expresados en millones de dólares y el salario es la remuneración anual expresada en miles de dólares.

El efecto marginal de *ceoten* sobre *salary* expresado en porcentaje es el siguiente:

$$em_{salario/ceoten}\,\% = 4.40 - 2\times 0.12 ceoten$$

Así, para un consejero delegado con 10 años en su compañía, si está un año más en la empresa, su salario se incrementará en un 2%. Igualando a cero la expresión anterior y despejando *ceoten*, nos encontramos con que el efecto máximo de permanencia como consejero delegado sobre el salario se alcanza a los 18 años. Es decir, hasta los 18 años como CEO el efecto marginal del salario con respecto a los años de permanencia en la compañía es positivo. Por el contrario, desde los 18 años en adelante, este efecto marginal es negativo.

Funciones cúbicas

Otro caso interesante es la *función cúbica* o *función polinomial de tercer grado*. Si en el modelo hay sólo regresores correspondientes a la función cúbica, tenemos un *modelo cúbico*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + u \tag{3-76}$$

Los modelos cúbicos se utilizan muy a menudo en economía aplicada para captar variaciones en los efectos marginales, particularmente en las funciones de costes. El efecto marginal (*em*) de *x* sobre *y*, que depende, según una forma cuadrática, del valor de *x*, será el siguiente:

$$em = \frac{dy}{dx} = \beta_2 + 2\beta_3 x + 3\beta_4 x^2 \tag{3-77}$$

El mínimo de *em* se producirá cuando

$$\frac{dem}{dx} = 2\beta_3 + 6\beta_4 x = 0 \tag{3-78}$$

Por lo tanto,

$$em_{\min} = \frac{-\beta_3}{3\beta_4} \tag{3-79}$$

En un modelo cúbico de una función de costes, debe cumplirse la restricción $\beta_3^2 < 3\beta_4\beta_2$ para garantizar que em_{mi} sea positivo. Otras restricciones que la función de costes debe cumplir son las siguientes: β_1, β_2, and $\beta_4>0$; y $\beta_3<0$.

Ejemplo 3.6 Efecto marginal en una función de costes

Utilizando los datos de 11 empresas de plantas de celulosa (fichero *costfunc*) para estudiar la función de costes, se estimó el siguiente modelo:

$$cost = \underset{(1.602)}{29.16} + \underset{(0.2167)}{2.316}\,output - \underset{(0.0081)}{0.0914}\,output^2 + \underset{(0.000086)}{0.0013}\,output^3$$

$$R^2=0.9984 \quad n=11$$

donde *output* es la producción de pasta de papel en miles de toneladas y *cost* es el coste total en millones de euros.

El *coste marginal* es el siguiente:

$$marcost = 2.316 - 2\times 0.0914 output + 3\times 0.0013 output^2$$

Por lo tanto, en una empresa con una producción de 30 mil toneladas de pasta de papel, si la empresa aumenta la producción de celulosa en mil de toneladas, el coste se incrementará en 0.754 millones de euros. Calculando el mínimo de la expresión anterior y resolviendo para el *output*, nos encontramos con que el coste marginal mínimo es igual a una producción de 23222 toneladas de pasta de papel.

3.5 Bondad del ajuste y selección de regresores

Una vez que se han aplicado los mínimos cuadrados, es conveniente tener alguna medida de la bondad del ajuste del modelo a los datos. En el caso de que se hayan estimado varios modelos alternativos, las medidas de la bondad del ajuste podrían ser utilizadas para seleccionar el modelo más apropiado.

En la literatura econométrica existen numerosas medidas de bondad del ajuste. La más popular es el coeficiente de determinación, que se designa por R^2 o *R*-cuadrado, y el coeficiente de determinación ajustado, que se designa por $\bar{R}^2$ o *R*-cuadrado ajustado. Dado que estas medidas tienen algunas limitaciones, nos referiremos también al criterio de información de Akaike (*AIC*) y al criterio de Schwarz (*SC*).

3.5.1 Coeficiente de determinación

Como vimos en el capítulo 2, el coeficiente de determinación se basa en la siguiente descomposición:

$$SCT = SCE + SCR \tag{3-80}$$

donde *SCT* es la *suma de cuadrados totales*, *SCE* es la *suma de cuadrados* explicados y *SCR* es la *suma de cuadrados residual*.

Basándose en esta ecuación, el coeficiente de determinación se define como:

$$R^2 = \frac{SCE}{SCT} \tag{3-81}$$

Alternativamente, y de una forma equivalente, el coeficiente de determinación se puede definir como

$$R^2 = 1 - \frac{SCR}{SCT} \tag{3-82}$$

Los valores extremos del coeficiente de determinación son: 0, cuando la varianza explicada es cero, y 1, cuando la varianza residual es cero, es decir, cuando el ajuste es perfecto. Por lo tanto,

$$0 \leq R^2 \leq 1 \tag{3-83}$$

Un R^2 pequeño implica que la varianza de la perturbación (σ^2) es grande en relación a la variación de *y*, lo que significa que β_j no puede ser estimada con precisión. Pero hay que recordar, que una varianza de la perturbación grande puede compensarse con un tamaño muestral elevado, de forma que si *n* es suficientemente grande, podemos ser capaces de estimar los coeficientes con precisión a pesar de que no se hayan controlado muchos de los factores no observados.

Para interpretar el coeficiente de determinación adecuadamente, se deben tener en cuenta las siguientes cautelas:

a) Cuando se añaden nuevas variables explicativas, el coeficiente de determinación aumenta su valor o, al menos, mantiene el mismo valor. Esto sucede a pesar de que la variable o variables añadidas no tengan relación con la variable endógena. Así pues, siempre se verifica que

$$R_j^2 \geq R_{j-1}^2 \tag{3-84}$$

donde R_{j-1}^2 es el *R* cuadrado en un modelo con *j*-1 regresores, y R_j^2 es el *R* cuadrado en un modelo con un regresor adicional. Es decir, si se añaden variables a un modelo determinado, R^2 nunca disminuirá, incluso si estas variables no tienen una influencia significativa.

b) Si el modelo no tiene término independiente, el coeficiente de determinación no tiene una interpretación clara, porque la descomposición dada (3-80) no se cumple. Además, las dos formas de cálculo mencionadas -(3-81) y (3-82)- por lo general conducen a resultados diferentes, que en algunos casos pueden quedar fuera del intervalo [0,1].

c) El coeficiente de determinación no se puede utilizar para comparar modelos en los que la forma funcional de la variable endógena es diferente. Por ejemplo, R^2 no se puede aplicar para comparar dos modelos en los que el regresando es la variable original en uno, *y*, y ln(*y*) en el otro.

3.5.2 R cuadrado ajustado

Para superar una de las limitaciones del R^2, este coeficiente se puede "ajustar" de manera que tenga en cuenta el número de variables incluidas en un modelo dado. Para ver cómo el R^2 usual podría ajustarse, es útil expresarlo como

$$R^2 = 1 - \frac{SCR / n}{SCT / n} \tag{3-85}$$

donde, en el segundo término del segundo miembro, aparece la varianza residual dividida por la varianza del regresando.

El R^2, tal como está definido en (3-85), es una medida *muestral*. Ahora bien, si deseamos una medida poblacional (R_{POB}^2), ésta se podría definir como

$$R^2_{POB} = 1 - \frac{\sigma_u^2}{\sigma_y^2} \tag{3-86}$$

Sin embargo, hay que tener en cuenta que se dispone de una mejor estimación de las varianzas, σ_u^2 y σ_y^2, que las utilizadas en (3-85). En su lugar, vamos a utilizar estimaciones insesgadas de estas varianzas:

$$\bar{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k} \tag{3-87}$$

Esta medida se denomina *R cuadrado ajustado*, o $\bar{R}^2$. El principal atractivo del $\bar{R}^2$ es que impone una penalización al añadir otros regresores a un modelo. Si se añade un regresor al modelo la *SCR* decrece o, en el peor de los casos queda, igual. Por otra parte, los *grados de libertad* de la regresión (*n–k*) siempre disminuyen. Por ello, el $\bar{R}^2$ puede crecer o decrecer cuando se añade un nuevo regresor al modelo. Es decir:

$$\bar{R}^2_j \geq \bar{R}^2_{j-1} \quad \text{o} \quad \bar{R}^2_j \leq \bar{R}^2_{j-1} \tag{3-88}$$

Un resultado algebraico interesante es el hecho de que si añadimos un nuevo regresor a un modelo, el $\bar{R}^2$ se incrementa si, y solo si, el estadístico *t* del nuevo regresor es mayor que uno en valor absoluto. Así, vemos inmediatamente que $\bar{R}^2$ podría ser utilizado para decidir si un determinado regresor adicional debe ser incluido en el modelo. El $\bar{R}^2$ tiene una cota superior que es igual a 1, pero estrictamente no tiene una cota inferior, ya que puede tomar un valor negativo, aunque muy cerca de 0.

Las observaciones b) y c) hechas para el *R* cuadrado siguen siendo válidas para el *R* cuadrado ajustado.

3.5.3 Criterio de información de Akaike (*AIC*) y criterio de Schwarz (*SC*)

Estos dos criterios -criterio de información de Akaike (*AIC*) y criterio de Schwarz (*SC*)- tienen una estructura muy similar. Por esta razón, se examinarán conjuntamente.

El estadístico *AIC*, propuesto por Akaike (1974) y basado en la teoría de la información, tiene la siguiente expresión:

$$AIC = -\frac{2l}{n} + \frac{2k}{n} \tag{3-89}$$

donde *l* es el logaritmo de la función de verosimilitud (suponiendo que las perturbaciones tengan una distribución normal) evaluada para los valores estimados de los coeficientes.

El estadístico *SC* propuesto por Schwarz (1978), tiene la siguiente expresión:

$$SC = -\frac{2l}{n} + \frac{k\ln(n)}{n} \tag{3-90}$$

Los estadísticos *AIC* y *SC*, a diferencia de los coeficientes de determinación (R^2 y $\bar{R}^2$), indican mejores ajustes cuanto más bajos sean sus valores. Es importante destacar que los estadísticos *AIC* y *SC* no tienen cotas, a diferencia del R^2.

a) Los estadísticos *AIC* y *SC* penalizan la introducción de nuevos regresores. En el caso de *AIC*, como puede verse en el segundo término del segundo miembro de (3-89), el número de regresores *k* aparece en el numerador. Por lo tanto, el crecimiento de *k* incrementará el valor del *AIC* y por lo tanto empeorará la bondad del ajuste, si no se ve compensado por un crecimiento suficiente de *l*. En el caso del *SC*, como puede verse en el segundo término del segundo miembro de (3-90), el numerador es $k\ln(n)$. Para $n>7$, ocurre lo siguiente: $k\ln(n)>2k$. Por lo tanto, el *SC* impone una penalización mayor a la introducción de regresores que el *AIC* cuando el tamaño de la muestra es mayor de 7.

b) Los estadísticos *AIC* y *SC* se puede aplicar a modelos estadísticos sin término independiente.

c) Los estadísticos *AIC* y *SC* no son medidas relativas como lo son los coeficientes de determinación. Por lo tanto, su magnitud, en sí misma, no ofrece ninguna información.

d) Los estadísticos *AIC* y *SC* se puede aplicar para comparar modelos en los que las variables endógenas tienen diferentes formas funcionales. En particular, vamos a comparar dos modelos en los que los regresandos son *y* y $\ln(y)$. Cuando el regresando es *y*, se aplica la fórmula (3-89) en el caso del *AIC*, o (3-90) en el caso del *SC*. Cuando el regresando es $\ln(y)$, y además queremos comparar con otro modelo en el que el regresando es *y*, hay que corregir esos estadísticos de la siguiente manera:

$$AIC_C = AIC + 2\overline{\ln(Y)} \tag{3-91}$$

$$SC_C = SC + 2\overline{\ln(Y)} \tag{3-92}$$

donde AIC_C y SC_C son los estadísticos corregidos, y *AIC* y *SC* son los estadísticos que suministra cualquier paquete econométrico como, por ejemplo, el E-views.

Ejemplo 3.7 Selección del mejor modelo

Para analizar los determinantes del gasto en productos lácteos, se han considerados los siguientes modelos alternativos:

1) $dairy = \beta_1 + \beta_2 inc + u$
2) $dairy = \beta_1 + \beta_2 \ln(inc) + u$
3) $dairy = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
4) $dairy = \beta_2 inc + \beta_3 punder5 + u$
5) $dairy = \beta_1 + \beta_2 inc + \beta_3 hhsize + u$
6) $\ln(dairy) = \beta_1 + \beta_2 inc + u$
7) $\ln(dairy) = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
8) $\ln(dairy) = \beta_2 inc + \beta_3 punder5 + u$

donde *inc* es la renta disponible de los hogares, *hhsize* es el número de miembros del hogar y *punder5* es la proporción de niños menores de cinco años en el hogar.

Utilizando una muestra de 40 hogares (fichero *demand*), y teniendo en cuenta que $\overline{\ln(dairy)} = 2.3719$, los estadísticos de bondad del ajuste obtenidos para los 8 modelos se muestran en el cuadro 3.1. En particular, el estadístico *AIC* corregido para el modelo 6) se ha calculado como sigue:

$$AIC_C = AIC + 2\overline{\ln(Y)} = 0.2794 + 2 \times 2.3719 = 5.0232$$

Conclusiones

a) El *R*-cuadrado puede ser utilizado para comparar los siguientes pares de modelos: 1) con 2), y 3) con 5).

b) El *R*-cuadrado ajustado sólo se puede utilizar para comparar los modelos 1) con 2), 3) y 5); y 6) con 7.

c) El mejor de los ocho modelos es el modelo de 7) de acuerdo con los criterios *AIC* y *SC*.

CUADRO 3.1. Medidas de bondad de ajuste de ocho modelos.

Número de modelo	1	2	3	4	5	6	7	8
Regresando	*dairy*	*dairy*	*dairy*	*dairy*	*dairy*	ln(*dairy*)	ln(*dairy*)	ln(*dairy*)
Regresores	*intercept* *inc*	*intercept* ln(*inc*)	*intercept* *inc* *punder5*	*inc* *punder5*	*intercept* *Inc* *househsize*	*intercept* *inc*	*intercept* *inc* *punder5*	*inc* *punder5*
R-cuadrado	0.4584	0.4567	0.5599	0.5531	0.4598	0.4978	0.5986	-0.6813
R-cuadrado ajustado	0.4441	0.4424	0.5361	0.5413	0.4306	0.4846	0.5769	-0.7255
Criterio de información de Akaike	5.2374	5.2404	5.0798	5.0452	5.2847	0.2794	0.1052	1.4877
Criterio de Schwarz	5.3219	5.3249	5.2065	5.1296	5.4113	0.3638	0.2319	1.5721
Criterio de información de Akaike corregido						5.0232	4.8490	6.2314
Criterio de Schwarz corregido						5.1076	4.9756	6.3159

Ejercicios

Ejercicio 3.1 Considere el modelo de regresión lineal $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, donde $\mathbf{X}$ es una matriz 50×5.

Conteste de forma razonada a las siguientes cuestiones:

a) ¿Cuáles son las dimensiones de los vectores $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$?

b) ¿Cuántas ecuaciones hay en el sistema de ecuaciones normales $\mathbf{X'X\hat{\beta}} = \mathbf{X'y}$?

c) ¿Qué condición debe cumplirse para poder obtener $\hat{\boldsymbol{\beta}}$?

Ejercicio 3.2 Dado el modelo

$$y_i=\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+u_i$$

y los siguientes datos:

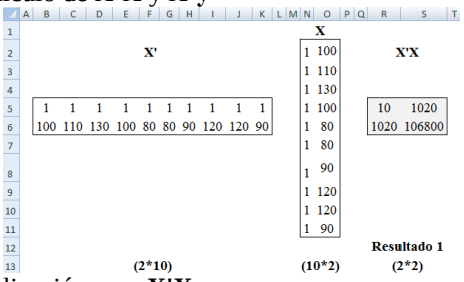
y	x_2	x_3
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

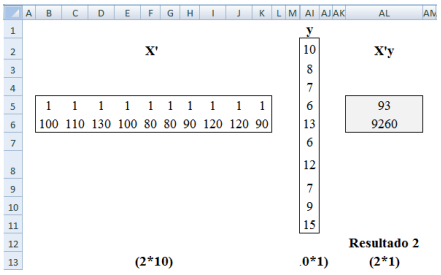
a) Estime β_1, β_2 y β_3 por *MCO*.
b) Calcule la suma de los cuadrados de los residuos.
c) Obtenga la varianza residual.
d) Obtenga la varianza explicada por la regresión.
e) Obtenga la varianza de la variable endógena.
f) Calcule el coeficiente de determinación.
g) Obtenga una estimación insesgada de σ^2.
h) Estime la varianza de $\hat{\beta}_2$.

Para responder a estas preguntas puede utilizar Excel. Véase el recuadro 3.1.

Recuadro 3.1

a) Cálculo de **X'X** y **X'y**





Explicación para **X'X**

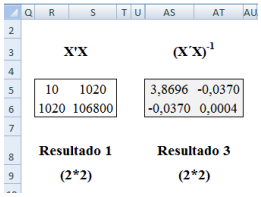
a) Introduzca las matrices **X'** y **X**: B5:K6 y N2:O11 en Excel

b) El producto **X'X** se calcula seleccionando previamente las celdas donde desea colocar la matriz resultante (R5:S6).

c) Una vez seleccionadas las celdas para la matriz resultante, y mientras aún está resaltada, escriba la fórmula siguiente: = MMULT (B5:K6;N2:O11)

d) Cuando la fórmula se haya introducido, pulse la *tecla Ctrl* y la *tecla Shift* simultáneamente, entonces, teniendo presionadas estas dos teclas, pulse la *tecla Enter* también.

2) Cálculo de $\mathbf{(X'X)^{-1}}$



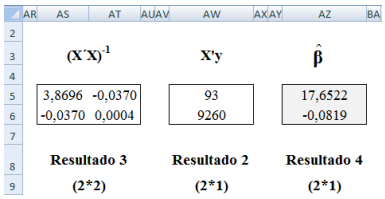
a) Introduzca la matriz **X'X** en Excel: R5:S6

b) Encontramos la inversa de la matriz **X'X**, seleccionando previamente las celdas donde deseamos colocar la matriz resultante (AS5:AT6)

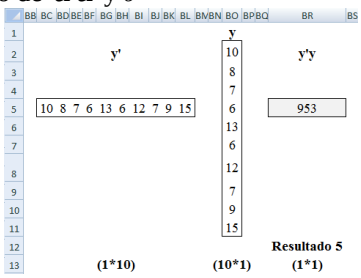
c) Una vez seleccionadas las celdas para la matriz resultante, y mientras aún está resaltada, escriba la fórmula siguiente: = MINVERSA (AO5:AP6).

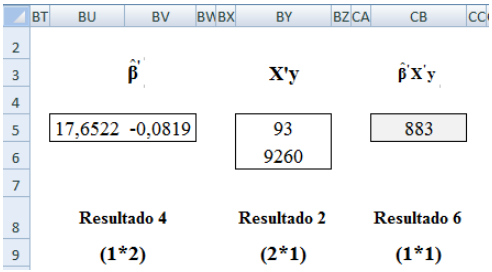
d) Cuando la fórmula se haya introducido, pulse la *tecla Ctrl* y la *tecla Shift* simultáneamente, entonces, teniendo presionadas estas dos teclas, pulse la *tecla Enter* también.

3) Cálculo del vector $\hat{\boldsymbol{\beta}}$



4) Cálculo de $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ y σ^2





$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = R.5 - R.6 = 953 - 883 = 70$$

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-2} = \frac{70}{8} = 8.6993$$

5) Cálculo de la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} = 8.6993\begin{pmatrix} 3.8696 & -0.0370 \\ -0.0370 & 0.0004 \end{pmatrix} = \begin{pmatrix} 33.6624 & -0.3215 \\ -0.3215 & 0.0032 \end{pmatrix}$$

Ejercicio 3.3 El siguiente modelo ha sido estimado para explicar las *ventas* anuales de empresas fabricantes de productos de limpieza doméstica en función de un índice de precios relativo (*ipr*) y de gastos de publicidad (*gpub*):

$$ventas = \beta_1 + \beta_2 ipr + \beta_3 gpub + u$$

donde las *ventas* están expresadas en millones de euros, *ip* es un índice de precios relativos (precios de la empresa/precios de la empresa 1 de la muestra) y *gpub* son los gastos anuales realizados en publicidad y campañas de promoción y difusión, expresados también en millones de euros.

Para ello se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

firm	*ventas*	*ipr*	*gpub*
1	10	100	300
2	8	110	400
3	7	130	600
4	6	100	100
5	13	80	300
6	6	80	100
7	12	90	600
8	7	120	200
9	9	120	400
10	15	90	700

Utilizando una hoja excel:

a) Estime los parámetros del modelo propuesto.
b) Estime la matriz de covarianzas.
c) Calcule el coeficiente de determinación.

Nota: En el recuadro 3.1 se estima el modelo $ventas = \beta_1 + \beta_2 rpi + u$ utilizando Excel. Allí también pueden verse las instrucciones para hacerlo.

Ejercicio 3.4 Un investigador, que está elaborando un modelo econométrico con el que desea explicar el comportamiento de la renta, formula la siguiente especificación:

$$renta = \alpha + \beta cons + \gamma ahorro + u \qquad [1]$$

donde *renta* es la renta disponible de las familias, *cons* es el consumo total y *ahorro* es el ahorro total de las familias.

El investigador no tuvo en cuenta que las tres magnitudes anteriores están ligadas por la identidad

$$renta = cons + ahorro \qquad [2]$$

La equivalencia entre los modelos [1] y [2] exige que, además de desaparecer el término de perturbación, los parámetros del modelo [1] tomen los siguientes valores:

$$\alpha = 0,\ \beta = 1,\ \gamma = 1$$

Si se emplean los datos de un país para ajustar la ecuación [1] por MCO, ¿Se puede esperar, *en general*, que las estimaciones obtenidas tomen los valores

$$\hat{\alpha} = 0, \hat{\beta} = 1, \hat{\gamma} = 0\,?$$

Justifíquese la respuesta, utilizando notación matemática.

Ejercicio 3.5 Un investigador plantea el siguiente modelo econométrico para explicar los ingresos totales por turismo en un país determinado (*ingtotur*):

$$gastotur = \beta_1 + \beta_2 gasmetur + \beta_3 numtur + u$$

donde *gasmetur* es el gasto medio por turista y *numtur* es el número total de turistas.

a) Es obvio que *gastotur*, *gasmetur* y *numtur* están ligados también por la relación *gastotur*= *gasmetur*×*numtur*; ¿afectará este hecho de alguna forma a las estimaciones de los parámetros del modelo propuesto?

b) ¿Existe otra forma funcional del modelo que implique restricciones más fuertes sobre los parámetros? Si la hubiera, indíquela.

c) ¿Le parece razonable utilizar el modelo indicado para explicar el comportamiento de los ingresos por turismo?

Ejercicio 3.6 Supongamos que se tiene que estimar el modelo

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + u$$

utilizando las siguientes observaciones:

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

¿Qué problemas puede plantear la estimación de este modelo?

Ejercicio 3.7 Conteste a las siguientes preguntas:

a) Explique que miden los coeficientes de determinación (R^2) y de determinación corregido ($\overline{R}^2$). ¿Para qué se pueden utilizar? Razone la respuesta.

b) Dados los modelos

$$\ln(y)=\beta_1+\beta_2\ln(x)+u \quad (1)$$
$$\ln(y)=\beta_1+\beta_2\ln(x)+\beta_3\ln(z)+u \quad (2)$$
$$\ln(y)=\beta_1+\beta_2\ln(z)+u \quad (3)$$
$$y=\beta_1+\beta_2 z+u \quad (4)$$

indique qué medida de bondad del ajuste es adecuada para comparar los siguientes pares de modelos: (1)-(2); (1)-(3); y (1)-(4). Razone su respuesta.

Ejercicio 3.8 Se estima por *MCO* el siguiente modelo:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + \beta_3 \ln(z) + u$$

a) ¿Pueden ser todos los residuos mínimo cuadráticos positivos? Razone la respuesta.

b) Bajo la hipótesis básica de no autocorrelación de las perturbaciones, ¿son independientes los residuos minimocuadráticos? Razone la respuesta.

c) Suponiendo que las perturbaciones no tengan distribución normal, ¿los estimadores minimocuadráticos son insesgados? Razone la respuesta.

Ejercicio 3.9 Considere el modelo de regresión

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$$

donde **y** y **u** son vectores 8×1, **X** es una matriz 8×3 y **β** es un vector 3×1 de parámetros desconocidos. Además, se dispone de la siguiente información:

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}2 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 3\end{bmatrix} \qquad \hat{\mathbf{u}}'\hat{\mathbf{u}}=22$$

Responda a las siguientes preguntas, justificando la respuesta:

a) Indique el tamaño de la muestra, el número de regresores, el número de parámetros y los grados de libertad de la suma de los cuadrados de los residuos.

b) Deduzca la matriz de covarianzas del vector $\hat{\boldsymbol{\beta}}$, explicitando los supuestos utilizados. Estime las varianzas de los estimadores de los parámetros del modelo.

c) ¿Contiene el modelo de regresión término constante? ¿Qué implicaciones tiene la contestación a esta pregunta en el significado del R^2 en este modelo?

Ejercicio 3.10 Argumente la veracidad o falsedad de las siguientes afirmaciones:

a) En un modelo de regresión lineal, la suma de los residuos es cero.

b) El coeficiente de determinación (R^2) es siempre una buena medida de la calidad del modelo.

c) El estimador por mínimos cuadrados es un estimador sesgado.

Ejercicio 3.11 El siguiente modelo se formula para explicar el tiempo empleado en dormir:

$$sleep=\beta_1+\beta_2\ totalwrk+\beta_3 leisure+u$$

donde el tiempo dedicado a dormir (*sleep),* al trabajo -remunerado y no remunerado- (*totalwrk*), y al ocio (*leisure*) (tiempo no dedicado a dormir o trabajar) están medidos en minutos por día.

La ecuación estimada con una muestra de 200 observaciones, utilizando el fichero *timuse03,* es la siguiente:

$$sleep=1440-1\times\ total\_work-1\times leisure$$

$$R^2=1.000 \qquad n=1000$$

a) ¿Cuál es su opinión acerca de estos resultados?

b) ¿Cuál es el significado del término independiente estimado?

Ejercicio 3.12 Utilizando una submuestra de la Encuesta de Estructura Salarial para España en 2006 (archivo *wage06sp*) se estimó el siguiente modelo para explicar el salario (*wage*):

$$\ln(wage)=1.565+0.0448educ+0.0177tenure+0.0065age$$

$$R^2=0.337 \qquad n=800$$

donde educación (*educ*), permanencia en la empresa (*tenure)* y edad (*age*) están medidos en años y el salario en euros por hora.

a) ¿Cuál es la interpretación de los coeficientes *educ*, *tenure* y *age*?
b) ¿Cuántos años tiene que aumentar la edad para que tenga un efecto similar al incremento de 1 año en la educación, manteniendo fijos en cada caso, los otros dos regresores?
c) Sabiendo que $\overline{educ}=10.2$, $\overline{tenure}=7.2$ y $\overline{age}=42.0$, calcule las elasticidades de los salarios con respecto a la educación, permanencia en la empresa y edad, manteniendo fijos los otros regresores. ¿Considera usted que estas elasticidades son altas o bajas?

Ejercicio 3.13 La siguiente ecuación describe el precio de la vivienda en términos del número de dormitorios de la casa (*bedrooms*), del número de baños completos (*bathrms*) y del tamaño de la parcela en pies cuadrados (*lotsize*):

$$price = \beta_1 + \beta_2 bedrooms + \beta_3 bathrms + \beta_4 lotsize + u$$

donde el precio (*price*) de la vivienda se mide en dólares.

Utilizando los datos de la ciudad de Windsor contenidos en el fichero *housecan*, se estima el siguiente modelo:

$$price = -2418 + 5827 bedrooms + 19750 bathrms + 5.411 lotsize$$

$$R^2=0.486 \qquad n=546$$

a) ¿Cuál es el aumento estimado en el precio de una casa con un dormitorio y un baño adicionales, manteniendo *lotsize* constante?
b) ¿Qué porcentaje de variación en el precio se explica por el número de dormitorios, el número de baños completos y el tamaño de la vivienda en conjunto?
c) Determine el precio de venta predicho para una casa de la muestra con *bedrooms*=3, *bathrms*=2 y *lotsize*=3880.
d) El precio de venta real de la casa en *c*) fue de 66000$. Encuentre el valor del residuo para esta casa. A la vista de este resultado, ¿el comprador pagó de más o de menos por la casa?

Ejercicio 3.14 Para examinar los efectos de los rendimientos de las empresas sobre los salarios de sus consejeros delegados se formuló el siguiente modelo:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + u$$

donde *roa*, es la ratio beneficios/activos expresada en porcentaje y *tenure* es el número de años en la empresa como consejero delegado (=0 si es menor de 6 meses). El salario (*salary*) está expresado en miles de dólares, mientras que las ventas (*sales*) y los beneficios (*profits*) están en millones de dólares.

Se ha utilizado el fichero *ceoforbes* para la estimación del modelo. Este archivo contiene datos sobre 447 ejecutivos de las 500 empresas más grandes de EE.UU. (52 de las 500 empresas fueron excluidas por falta de datos sobre una o más variables. Apple Computer también fue excluido porque Steve Jobs, consejero delegado de Apple en 1999, no recibió ninguna compensación durante ese período.) Los datos de las empresas provienen de la revista Fortune y se refieren a 1999, los datos de los consejeros delegados provienen de la revista Forbes y se refieren también a 1999. Los resultados obtenidos fueron los siguientes:

$$\ln(salary) = 4.641 + 0.0054 roa + 0.2893 \ln(sales) + 0.0000564 profits + 0.0122 tenure$$

$$R^2=0.232 \qquad n=447$$

a) Interprete el coeficiente del regresor *roa*.
b) Interprete el coeficiente del regresor ln(*sales*). ¿Cuál es su opinión sobre la magnitud de la elasticidad del *salary/sales*?
c) Interprete el coeficiente del regresor *profits*.
d) ¿Cuál es la elasticidad de *salary/profits* para el punto de las medias muestrales ($\overline{salary}$ =2028 y $\overline{profits}$ =700).

Ejercicio 3.15 (Continuación del ejercicio 2.21) Utilizando una base de datos de 1983 empresas encuestada en el año 2006 (fichero *rdspain*), se estimó la siguiente ecuación:

$$rdintens = -1.8168 + 0.1482\ln(sales) + 0.0110 exponsal$$

$$R^2 = 0.048 \quad n = 1983$$

donde *rdintens* es el gasto en investigación y desarrollo (I+D) expresado como porcentaje de las ventas, las ventas (*sales*) se miden en millones de euros, y *exponsal* son las exportaciones tomadas como porcentaje de las ventas.

a) Interprete el coeficiente de ln(*sales*). En particular, si las ventas aumentan en un 100%, ¿cuál es el porcentaje de cambio estimado de *rdintens*? ¿Es éste un efecto económico grande?
b) Interprete el coeficiente de *exponsal*. ¿Es grande este coeficiente desde un punto de vista económico?
c) ¿Qué porcentaje de la variación en *rdintens* se explica por las ventas y por las exportaciones tomadas como porcentaje de las ventas?
d) ¿Cuál es la elasticidad *rdintens*/*sales* para la media muestral ($\overline{rdintens}$ = 0.732 y $\overline{sales}$ =63544960). Comente el resultado.
e) ¿Cuál es la elasticidad *rdintens*/*exponsal* para la media muestral ($\overline{rdintens}$ = 0.732 y $\overline{exponsal}$ 17.657)? Comente el resultado.

Ejercicio 3.16 La siguiente regresión hedónica se formuló para explicar los precios de los coches (véase ejemplo 3.3):

$$\ln(price) = \beta_1 + \beta_2 cid + \beta_3 hpweight + \beta_4 fueleff + u$$

donde *cid*, es el desplazamiento en pulgadas cúbicas, *hpweight* es la *ratio* potencia/ peso en kg, expresada en porcentaje y *fueleff* es la ratio litros por 100 km/caballos de vapor expresada en porcentaje.

a) ¿Cuáles son los signos probables de β_2, β_3 y β_4? Explíquelo.
b) Estime el modelo utilizando el fichero *hedcarsp* y exprese los resultados en forma de ecuación.
c) Interprete el coeficiente del regresor *cid*.
d) Interprete el coeficiente del regresor *hpweight*.
e) Expanda el modelo, introduciendo un regresor relativo al tamaño de coche, como el volumen o el peso. ¿Qué pasa si se introducen los dos en la regresión? ¿Cuál cree que es la relación entre el peso y el volumen?

Ejercicio 3.17 El concepto de trabajo cubre un amplio espectro de actividades posibles en la economía productiva. Una parte importante del trabajo es no remunerado, no pasa por el mercado y, por lo tanto, no tiene precio. El trabajo no remunerado más importante es el trabajo realizado en el hogar (*houswork*) llevado a cabo principalmente por mujeres. Con el fin de analizar los factores que influyen en el trabajo del hogar, se ha formulado el siguiente modelo:

$$housework = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

donde *educ* son los años de educación alcanzados, *hhinc* son los ingresos de los hogares en euros por mes, *age* es la edad de la persona entrevistada y *paidwork* es el trabajo remunerado. Las variables *housework* y *paidwork* están medidas en minutos por día.

Utilice los datos contenidos en el fichero *timuse03* para estimar el modelo. Este archivo contiene 1000 observaciones correspondientes a una submuestra aleatoria extraída de la encuesta de *Empleo del Tiempo* en España que se llevó a cabo en el período 2002-2003.

a) ¿Qué signos esperaría para β_2, β_3, β_4 y β_5? Explíquelo.
b) Exprese los resultados en forma de ecuación
c) ¿Cree usted que hay factores relevantes omitidos en la ecuación anterior? Explíquelo.
d) Interprete los coeficientes de los regresores *educ*, *hhinc*, *age* y *paidwork*.

Ejercicio 3.18 (Continuación del ejercicio 2.20) Para explicar la satisfacción general de las personas (*stsfglo*) se formula el siguiente modelo:

$$stsfglo = \beta_1 + \beta_2 gnipc + \beta_3 lifexpec + u$$

donde *gnipc* es la renta nacional bruta per cápita expresada en dólares (USA) PPA (paridad del poder adquisitivo) a precios de 2008 y *lifexpec* es la esperanza de vida al nacer, es decir, el número de años que un recién nacido puede esperar vivir. Cuando una magnitud se expresa en dólares estadounidenses PPA, eso significa que se ha convertido a dólares internacionales usando tasas PPA. (Un dólar internacional tiene el mismo poder adquisitivo que un dólar de los EE.UU. en los Estados Unidos.)

Utilice el fichero *HDR2010* para la estimación del modelo.

a) ¿Qué signos esperaría para β_2 y β_3? Explíquelo.
b) ¿Cuál sería la de satisfacción global media de un país cuyos habitantes tienen una esperanza de vida al nacer de 80 años y que tienen una renta nacional bruta per cápita de 30.000 dólares PPA expresados en dólares de 2008 de Estados Unidos?
c) Interprete los coeficientes de *gnipc* y *lifexpe*.
d) Teniendo en cuenta un país cuya esperanza de vida al nacer es igual a 50 años, ¿Cual debería ser la renta nacional bruta per cápita para obtener una satisfacción global igual a 5?

Ejercicio 3.19 (Continuación del ejercicio 2.24) Debido a los problemas surgidos en el modelo keynesiano, Brown introdujo en la función de consumo, además de la renta, el consumo retardado para reflejar la persistencia de hábitos del consumidor:

$$conspc = \beta_1 + \beta_2 incpc + \beta_3 conspc(-1) + u$$

Como en este modelo se incluye el consumo retardado, hay que distinguir entre propensión marginal al consumo a corto plazo y a largo plazo. La propensión marginal a corto plazo se calcula de igual forma que en la función de consumo de Keynes. Para calcular la propensión marginal a largo plazo se debe considerar una situación de equilibrio, en la que no hay variaciones en las variables. Designando por $conspc^e$ y $incpc^e$ al consumo y a la renta de equilibrio y prescindiendo de la perturbación aleatoria, el modelo anterior en situación de equilibrio viene dado por

$$conspc^e = \beta_1 + \beta_2 incpc^e + \beta_3 conspc^e$$

La función de consumo de Brown se estimó con datos de la economía española para el periodo 1954-2010 (fichero *consumsp*), obteniéndose los siguientes resultados:

$$\widehat{conspc}_t = -7.156 + 0.3965 incpc_t + 0.5771 conspc_{t-1}$$

$$R^2 = 0.997 \qquad n = 56$$

a) Interprete el coeficiente de *incpc*. En su interpretación, ¿incluiría la clausula "mantenido fijo el otro regresor? Justifique la respuesta.

b) Calcule la elasticidad a corto plazo para las medias muestrales ($\overline{conspc}$ =8084, $\overline{incpc}$ =8896).

c) Calcule la elasticidad a largo plazo para las medias muestrales.

d) Comente la diferencia entre los valores obtenidos para los dos tipos de elasticidad.

Ejercicio 3.20 Para explicar la influencia de los incentivos y los gastos de publicidad en las ventas, se han formulado los modelos alternativos siguientes:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u \tag{1}$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(advert) + \beta_3 \ln(incent) + u \tag{2}$$

$$\ln(sales) = \beta_1 + \beta_2 advert + \beta_3 incent + u \tag{3}$$

$$sales = \beta_2 advert + \beta_3 incent + u \tag{4}$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(incent) + u \tag{5}$$

$$sales = \beta_1 + \beta_2 incent + u \tag{6}$$

a) Utilizando una muestra de 18 áreas de venta (fichero *advincen*), estime los modelos anteriores:

b) Dentro de cada uno de los siguientes grupos seleccione el mejor modelo, indicando cuáles han sido los criterios que se han utilizado. Justifique su respuesta.

*b*1*)* (1) y (6)
*b*2*)* (2) y (3)
*b*3*)* (1) y (4)
*b*4*)* (2), (3) y (5)
*b*5*)* (1), (4) y (6)
*b*6*)* (1), (2), (3), (4), (5) y (6)

Apéndices

Apéndice 3.1 Demostración del Teorema de Gauss-Markov

Para demostrar este teorema, se utilizan los supuestos 1 a 8 del *MLC*.

Consideremos otro estimador $\tilde{\boldsymbol{\beta}}$ que es una función de **y** (recuerde que $\hat{\boldsymbol{\beta}}$ es también una función de **y**), dado por

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X} \right]^{-1} \mathbf{X}' + \mathbf{A} \right] \mathbf{y} \tag{3-93}$$

donde **A** es una matriz arbitraria, $k \times n$, que es función de **X** y/o otras variables no estocásticas, pero no es función de **y**. Para que $\tilde{\boldsymbol{\beta}}$ sea insesgado, se han de cumplir ciertas condiciones.

Teniendo en cuenta (3-52), tenemos que

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \tag{3-94}$$

Tomando las esperanzas en ambos miembros de (3-94) tenemos que

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]E(\mathbf{u}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} \tag{3-95}$$

Para que $\tilde{\boldsymbol{\beta}}$ sea insesgado, es decir, $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, se debe cumplir lo siguiente:

$$\mathbf{A}\mathbf{X} = \mathbf{I} \tag{3-96}$$

Consecuentemente,

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \tag{3-97}$$

Teniendo en cuenta los supuestos 7 y 8, y (3-96), la $Var(\tilde{\boldsymbol{\beta}})$ es igual a

$$\begin{aligned} Var(\tilde{\boldsymbol{\beta}}) &= E((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})') = E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}'\right]\right] \\ &= E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] + \mathbf{A}\mathbf{A}'\right] = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}'\right] \end{aligned} \tag{3-98}$$

La diferencia entre ambas varianzas es la siguiente:

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}' - \left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \sigma^2\mathbf{A}\mathbf{A}' \tag{3-99}$$

El producto de una matriz por su transpuesta es siempre una matriz semidefinida positiva. Por lo tanto,

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}\mathbf{A}' \geq 0 \tag{3-100}$$

La diferencia entre la varianza de un estimador $\tilde{\boldsymbol{\beta}}$-arbitrario, pero lineal e insesgado- y la varianza del estimador $\hat{\boldsymbol{\beta}}$ es una matriz semidefinida positiva. En consecuencia, $\hat{\boldsymbol{\beta}}$ es un Estimador Lineal Insesgado Óptimo, es decir, es un estimador *ELIO*.

Apéndice 3.2 Demostración: $\hat{\sigma}^2$ es un estimador insesgado de la varianza de la perturbación

Con el fin de ver cuál es el estimador más conveniente de σ^2, se van a analizar primero las propiedades de la suma de los cuadrados de los residuos. Este es precisamente el numerador de la varianza residual.

Teniendo en cuenta (3-17) y (3-23), vamos a expresar el vector de residuos como una función del regresando

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{M}\mathbf{y} \tag{3-101}$$

donde **M** es la matriz idempotente.

Alternativamente, el vector de residuos se puede expresar como una función del vector de las perturbaciones:

$$\begin{aligned}\hat{\mathbf{u}} &= \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\boldsymbol{\beta}\mathbf{u} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} \\ &= \mathbf{M}\mathbf{u}\end{aligned} \tag{3-102}$$

Teniendo en cuenta (3-102), la suma de los cuadrados de los residuos (*SRC*) se puede expresar en la forma siguiente:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u} \tag{3-103}$$

Ahora bien, teniendo en cuenta que estamos buscando un estimador insesgado de σ^2, vamos a calcular la esperanza de la expresión anterior:

$$\begin{aligned}E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] &= E\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = trE\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = E\left[tr\mathbf{u}'\mathbf{M}\mathbf{u}\right] \\ &= E\left[tr\mathbf{M}\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}E\left[\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}\sigma^2\mathbf{I} \\ &= \sigma^2 tr\mathbf{M} = \sigma^2(n-k)\end{aligned} \tag{3-104}$$

En la deducción de (3-104), hemos utilizado la propiedad de la traza de que $tr(\mathbf{AB}) = tr(\mathbf{BA})$). Teniendo en cuenta esta propiedad de la traza, se obtiene el valor de $tr\mathbf{M}$:

$$\begin{aligned}tr\mathbf{M} &= tr\left[\mathbf{I}_{n\times n} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right] = tr\mathbf{I}_{n\times n} - tr\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' \\ &= tr\mathbf{I}_{n\times n} - tr\mathbf{I}_{k\times k} = n-k\end{aligned}$$

De acuerdo con (3-104), se cumple que

$$\sigma^2 = \frac{E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right]}{n-k} \tag{3-105}$$

A la vista de (3-105), un estimador de la varianza insesgado vendrá dado por:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k} \tag{3-106}$$

puesto que, de acuerdo con (3-104),

$$E(\hat{\sigma}^2) = E\left[\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}\right] = \frac{E(\hat{\mathbf{u}}'\hat{\mathbf{u}})}{n-k} = \frac{\sigma^2(n-k)}{n-k} = \sigma^2 \tag{3-107}$$

El denominador de (3-106) son los grados de libertad que corresponden a la *SCR* que aparece en el numerador. Este resultado se justifica por el hecho de que las ecuaciones normales del hiperplano imponen *k* restricciones sobre los residuos. Por lo tanto, el número de grados de libertad de la *SCR* es igual al número de observaciones (*n*) menos el número de restricciones *k*.

Apéndice 3.3 La consistencia del estimador de *MCO*

En el apéndice 2.8 hemos probado en el modelo de regresión simple la consistencia del estimador *MCO*. Ahora vamos a probar la consistencia del vector $\hat{\boldsymbol{\beta}}$ obtenido por *MCO*.

En primer lugar, el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$, dado en (3-23) puede expresarse así

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\left(\frac{1}{n}\mathbf{X'u}\right) \tag{3-108}$$

Ahora, tomamos límites al último factor de **¡Error! No se encuentra el origen de la referencia.** y llamamos **Q** al resultado:

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{X'X} = \mathbf{Q} \tag{3-109}$$

Si **X** es fija en muestras repetidas, de acuerdo con el supuesto 2, entonces (3-109) implica que $\mathbf{Q}=(1/n)\mathbf{X'X}$. De acuerdo con el supuesto 3, y debido a que la matriz inversa es una función continua de la matriz original, existe $\mathbf{Q}^{-1}$. Por lo tanto, podemos escribir que

$$\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim}\left[\frac{1}{n}\mathbf{X'u}\right]$$

El último término de (3.108) se puede expresar como

$$\begin{aligned}\frac{1}{n}\mathbf{X'u} &= \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{ji} & \cdots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{ki} & \cdots & x_{kn} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_i & \cdots & \mathbf{x}_n \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i = \overline{\mathbf{x}_i u_i}\end{aligned} \tag{3-110}$$

donde $\mathbf{x}_i$ es el vector de la columna correspondiente a la observación *i-ésima*.

Ahora, vamos a calcular la esperanza y la varianza de (3-110),

$$E\left[\overline{\mathbf{x}_i u_i}\right] = \frac{1}{n}\sum_{i=1}^{n}E\ \mathbf{x}_i u_i\ = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E\ u_i\ = \frac{1}{n}\mathbf{X'}E\ \mathbf{u}\ = \mathbf{0} \tag{3-111}$$

$$var\left[\overline{\mathbf{x}_i u_i}\right] = E\left[\overline{\mathbf{x}_i u_i}(\overline{\mathbf{x}_i u_i})'\right] = \frac{1}{n}\mathbf{X}'E\ \mathbf{uu}'\ \mathbf{X} = \frac{1}{n}\frac{\sigma^2}{n}\frac{\mathbf{X}'\mathbf{X}}{n} = \frac{\sigma^2}{n^2}\mathbf{Q} \qquad (3\text{-}112)$$

ya que $E\ \mathbf{uu}' = \sigma^2\mathbf{I}$ de acuerdo con los supuestos 7 y 8.

Tomando límites en (3-112), se sigue entonces que

$$\lim_{n\to\infty} \text{var}\left[\overline{\mathbf{x}_i \mathbf{u}_i}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n^2}\mathbf{Q} = 0(\mathbf{Q}) = \mathbf{0} \qquad (3\text{-}113)$$

Dado que la esperanza de $\overline{\mathbf{x}_i \mathbf{u}_i}$ es idéntica a cero y su varianza converge a cero, $\overline{\mathbf{x}_i \mathbf{u}_i}$ converge en media cuadrática a cero. La convergencia en media cuadrática implica la convergencia en probabilidad, por lo que plim($\overline{\mathbf{x}_i \mathbf{u}_i}$)=0. Por lo tanto,

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}(\mathbf{x}_i\mathbf{u}_i) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}\left[\frac{1}{n}\mathbf{X}'\mathbf{u}\right] = \boldsymbol{\beta} + \mathbf{Q}^{-1}\times 0 = \boldsymbol{\beta} \qquad (3\text{-}114)$$

En consecuencia, $\hat{\boldsymbol{\beta}}$ es un estimador consistente.

Apéndice 3.4 Estimador de máxima verosimilitud

El método de máxima verosimilitud es un método ampliamente utilizado en econometría. Este método propone como estimadores de los parámetros aquellos valores para los que la probabilidad de obtener las observaciones dadas es máxima. En la estimación de mínimos cuadrados no se adoptó *a priori* ningún supuesto; por el contrario, la estimación por máxima verosimilitud requiere establecer *a priori* supuestos estadísticos sobre los diversos elementos del modelo. Así, en la estimación por máxima verosimilitud vamos a adoptar todos los supuestos del modelo lineal clásico (*MLC*).

Por lo tanto, en la estimación por máxima verosimilitud de $\boldsymbol{\beta}$ y σ^2 en el modelo (3-52), se toman como estimadores a aquellos valores que maximizan la probabilidad de obtener las observaciones de una muestra dada.

Vamos a ver el procedimiento para obtener los estimadores de máxima verosimilitud $\boldsymbol{\beta}$ y σ^2 . De acuerdo con los supuestos del *MLC*

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad (3\text{-}115)$$

La esperanza y la varianza de la distribución de **y** están dadas por

$$E(\mathbf{y}) = E\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} \qquad (3\text{-}116)$$

$$\text{var}(\mathbf{y}) = E\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\right] = E\left[\mathbf{uu}'\right] = \sigma^2\mathbf{I} \qquad (3\text{-}117)$$

Por lo tanto,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \qquad (3\text{-}118)$$

La función de densidad de **y** (o función de verosimilitud), considerando **X** e **y** fijas y $\boldsymbol{\beta}$ y σ^2 variables, será de acuerdo con (3-118) igual a

$$L = f(y \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\left(1/2\sigma^2\right)(\mathbf{y - X\beta})'(\mathbf{y - X\beta})\right) \tag{3-119}$$

Dado que el máximo para L se alcanza en el mismo punto que ln(L), por ser la función logaritmo monótona, podemos, a efectos de maximización, trabajar con ln(L) en lugar de con L. Entonces,

$$\ln(L) = -\frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - \frac{1}{2\sigma^2} - \mathbf{(y - X\beta)'(y - X\beta)} \tag{3-120}$$

Para maximizar ln(L), hay que derivar con respecto a $\boldsymbol{\beta}$ y σ^2:

$$\frac{\delta \ln(L)}{\delta \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X'y} + 2\mathbf{X'X\beta}) \tag{3-121}$$

$$\frac{\delta \ln(L)}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\mathbf{(y - X\beta)'(y - X\beta)}}{2\sigma^4} \tag{3-122}$$

Igualando (3-121) a cero, vemos que el estimador de máxima verosimilitud de $\boldsymbol{\beta}$, denotado por $\tilde{\boldsymbol{\beta}}$, satisface que

$$\mathbf{X'X}\tilde{\beta} = \mathbf{X'y} \tag{3-123}$$

Como se supone que $\mathbf{X'X}$ es invertible, tenemos que

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X'X}\right]^{-1} = \mathbf{X'y} \tag{3-124}$$

En consecuencia, el estimador de máxima verosimilitud de $\boldsymbol{\beta}$, bajo los supuestos del *MLC*, coincide con el estimador *MCO*, es decir,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \tag{3-125}$$

Por lo tanto,

$$\mathbf{(y - X\tilde{\beta})'(y - X\tilde{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}) = \hat{u}'\hat{u}} \tag{3-126}$$

Igualando (3-122) a cero y sustituyendo $\boldsymbol{\beta}$ por $\tilde{\boldsymbol{\beta}}$, obtenemos:

$$-\frac{n}{2\tilde{\sigma}^2} + \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{2\tilde{\sigma}^4} = 0 \tag{3-127}$$

donde hemos designado por $\tilde{\sigma}^2$ al estimador de máxima verosimilitud de la varianza de las perturbaciones aleatorias. De (3-127) se deduce que

$$\tilde{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n} \tag{3-128}$$

Como puede verse el estimador de máxima verosimilitud no es igual al estimador insesgado que se ha obtenido en (3-106). De hecho, si tomamos esperanzas en (3-128),

$$E\left[\tilde{\sigma}^2\right] = \frac{1}{n}E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] = \frac{n-k}{n}\sigma^2 \tag{3-129}$$

Es decir, el estimador de máxima verosimilitud, $\tilde{\sigma}^2$, es un estimador sesgado, aunque su sesgo tiende a cero cuando n tiende a infinito, ya que

$$\lim_{n\to\infty} \frac{n-k}{n} = 1 \qquad (3\text{-}130)$$

4 CONTRASTE DE HIPÓTESIS EN EL MODELO DE REGRESIÓN MÚLTIPLE

4.1 El contraste de hipótesis: una panorámica

Antes de contrastar las hipótesis en el modelo de regresión múltiple, vamos a ofrecer una panorámica sobre el contraste de hipótesis.

El contraste de hipótesis permite realizar inferencias acerca de parámetros poblacionales utilizando datos provenientes de una muestra. Para realizar el contraste de hipótesis estadístico, en general, hay que realizar los siguientes pasos:

1) Establecer una hipótesis nula y una hipótesis alternativa relativas a los parámetros de la población.

2) Construir un estadístico para contrastar las hipótesis formuladas.

3) Definir una regla de decisión para determinar si la hipótesis nula debe ser, o no, rechazada en función del valor que tome el estadístico construido.

Vamos a examinar a continuación cada uno de estos pasos.

4.1.1 Formulación de la hipótesis nula y de la hipótesis alternativa

Antes de referirnos al modo de formular la hipótesis nula y alternativa, distinguiremos entre *hipótesis simples* e *hipótesis compuestas*. Las hipótesis que se formulan mediante una o más igualdades se denominan hipótesis simples. Cuando para formular una hipótesis se utilizan los operadores "desigualdad", "mayor que" y "menor que", entonces a dicha hipótesis se le denomina compuesta.

Es importante señalar que el contraste de hipótesis se refiere siempre a los parámetros *poblacionales*. El contraste de hipótesis implica tomar la decisión, sobre la base de los datos muestrales, de rechazar o no que ciertas restricciones sean satisfechas por el modelo básico asumido. Las restricciones que se desean contrastar se conocen como la *hipótesis nula*, a la que designa H_0. Así pues, una hipótesis nula es una declaración sobre los parámetros poblacionales.

Aunque es posible formular hipótesis nulas compuestas en el contexto del modelo de regresión, siempre consideraremos que la hipótesis nula es una hipótesis simple. Es decir, para formular una hipótesis nula, utilizaremos siempre el operador "igualdad". Veamos a continuación algunos ejemplos de hipótesis nulas referidas al modelo de regresión:

a) $H_0 : \beta_1 = 0$

b) $H_0 : \beta_1 + \beta_2 = 0$

c) $H_0 : \beta_1=\beta_2=0$

d) $H_0 : \beta_2+\beta_3=1$

También vamos a definir una *hipótesis alternativa*, designada por H_1, que representa nuestra conclusión en el caso de que el contraste concluya que H_0 es falsa.

Aunque las hipótesis alternativas pueden ser también simples o compuestas, en el modelo de regresión, tomaremos siempre, como hipótesis alternativa, una hipótesis compuesta. La hipótesis alternativa, a la que se designa H_1, se formula mediante el operador "desigualdad" en la mayor parte de los casos. Así, por ejemplo, dada la H_0:

$$H_0 : \beta_j = 1 \tag{4-1}$$

podemos formular la siguiente H_1 :

$$H_1 : \beta_j \neq 1 \tag{4-2}$$

que es una hipótesis "alternativa de dos colas".

Las siguientes hipótesis se llaman “alternativas de una cola”:

$$H_1 : \beta_j < 1 \tag{4-3}$$

$$H_1 : \beta_j > 1 \tag{4-4}$$

4.1.2 Estadístico de contraste

Un *estadístico de contraste* es una función de una muestra aleatoria, por lo que también es una variable aleatoria. Cuando se calcula el estadístico de contraste para una muestra dada, se obtiene un resultado, es decir, un número. Al realiza un contraste estadístico sería conveniente conocer la distribución del estadístico de contraste bajo la hipótesis nula. Esta distribución depende en gran medida de las hipótesis formuladas en el modelo. Si en la especificación del modelo se asume el supuesto de normalidad, entonces la distribución estadística apropiada será la distribución normal o alguna de las distribuciones asociadas a la misma, como son la *Chi-cuadrado*, la *t* de Student, o la *F* de Snedecor.

En el cuadro 4.1 se muestran algunas distribuciones, que son apropiadas en diferentes situaciones, bajo el supuesto de normalidad de las perturbaciones del modelo.

CUADRO 4.1. Algunas distribuciones utilizadas en el contraste de hipótesis.

	1 restricción	*1 o más restricciones*
σ^2 *conocida*	*N*	*Chi-square*
σ^2 *desconocida*	*t* de Student	*F* de Snedecor

El estadístico utilizado para el contraste se construye teniendo en cuenta la H_0 y los datos muestrales. En la práctica, como σ^2 es siempre desconocida, se utilizarán siempre las distribuciones *t* y *F*.

4.1.3 Regla de decisión

Para el contraste de hipótesis vamos a considerar dos enfoques: el enfoque clásico y un enfoque alternativo basado en los valores-*p*. Pero antes de ver la manera de aplicar la regla de decisión, examinaremos los tipos de error que se pueden cometer en el contraste de hipótesis.

Tipos de errores en el contraste de hipótesis

En el contraste de hipótesis, *podemos* cometer dos tipos de errores: *error de tipo* I y *error de tipo* II.

Error de tipo I

Nosotros podemos rechazar la H_0 cuando en realidad es cierta. A este error se le denomina *error de tipo* I. Generalmente, se define el *nivel de significación* (α) de un contraste como la probabilidad de cometer un *error de tipo* I. Simbólicamente,

$$\alpha = \Pr(\textit{rechazar } H_0 \mid H_0) \tag{4-5}$$

Dicho de otro modo, el nivel de significación es la probabilidad de rechazar la H_0 cuando la H_0 es cierta. Las reglas para el contraste de hipótesis se construyen haciendo que la probabilidad de un *error de tipo* I sea suficientemente pequeña. Los niveles de significación usuales para α son 0.10, 0.05 y 0.01. Algunas veces también se utiliza 0.001.

Después de haber tomado la decisión de rechazar o no la H_0, la decisión puede haber sido la correcta o puede que se haya cometido un error. Nunca sabremos con certeza si se cometió un error. Sin embargo, podemos calcular la *probabilidad* de haber cometido un *error de tipo* I o un *error de tipo* II.

Error de tipo II

Podemos fracasar en rechazar la H_0 cuando en realidad es falsa. Si esto sucede se ha cometido un *error de tipo* II.

$$\beta = \Pr(\textit{No rechazar } H_0 \mid H_1) \tag{4-6}$$

En otras palabras, β es la probabilidad de no rechazar H_0 cuando H_1 es cierta.

Es importante señalar que no es posible minimizar ambos tipos de error de forma simultánea. En la práctica lo que hacemos es seleccionar un nivel de significación bajo.

Enfoque clásico: Aplicación de la regla de decisión

El método clásico implica los siguientes pasos:

a) *Elección de* α. El contraste de hipótesis clásico requiere que inicialmente se especifique un *nivel de significación*. Cuando se especifica un valor para α, esencialmente lo que estamos cuantificando es nuestra tolerancia para un *error de tipo* I. Si α=0.05, entonces el investigador está dispuesto a rechazar H_0 falsamente en un 5% de los casos.

b) *Obtención de c, valor crítico*, utilizando tablas estadísticas. El valor *c* se determina por el valor de α.

El valor crítico en un contraste de hipótesis es un umbral con el cual se compara el estadístico de contraste para determinar si la hipótesis nula se rechaza o no.

c) *Comparando el resultado del estadístico de contraste, s, con c*, la H_0 se rechaza o no para un valor dado de α.

La región de rechazo (*RR*), delimitada por el valor crítico (*c*), es un conjunto de valores del estadístico de contraste para los cuales se rechaza la hipótesis nula. (Véase figura 4.1). Es decir, el espacio muestral del estadístico de contraste se divide en dos regiones: una región (la región de rechazo) nos lleva a rechazar la hipótesis nula H_0, mientras que la otra no nos deja rechazar la hipótesis nula. Por lo tanto, si el valor observado del estadístico de contraste *s* se encuentra en la región crítica, rechazamos la H_0; en el caso de que no se encuentre en la región de rechazo llegamos a la conclusión, de *no rechazar* la H_0 o de *fracasar en rechazar* la H_0.

Simbólicamente,

$$\begin{array}{llll} \text{Si} & s \geq c & \text{se rechaza} & H_0 \\ \text{Si} & s < c & \text{no se rechaza} & H_0 \end{array} \qquad (4\text{-}7)$$

Si la hipótesis nula se rechaza con la evidencia de la muestra, ésta es una conclusión *fuerte*. Sin embargo, la aceptación de la hipótesis nula es una conclusión *débil* porque no conocemos cuál es la probabilidad de no rechazar la hipótesis nula cuando debe ser rechazada. Es decir, no conocemos la probabilidad de cometer un e*rror de tipo* II. Por lo tanto, en lugar de utilizar la expresión de la aceptación de la hipótesis nula, es más correcto decir *fracasar en rechazar* la hipótesis nula, o *no rechazar*, ya que lo que realmente sucede es que no tenemos suficiente evidencia empírica para rechazar la hipótesis nula.

En el proceso de contrastación, la parte más subjetiva es la determinación *a priori* del nivel de significación. ¿Qué criterios se pueden utilizar para determinar α? En general, se trata de una decisión arbitraria, aunque como ya hemos dicho, los niveles de 1%, 5% y 10% para α son los más utilizados en la práctica. A veces se efectúa el contraste condicionado a distintos niveles de significación.

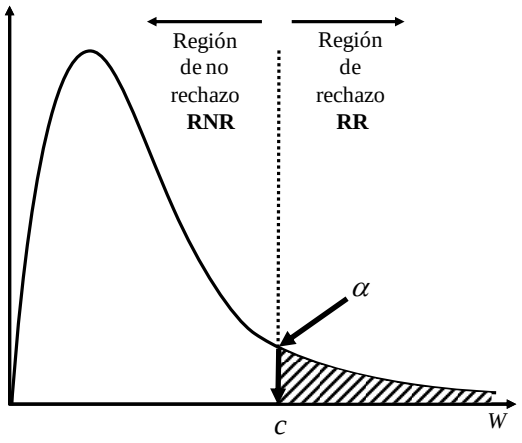


FIGURA 4.1. Contraste de hipótesis: enfoque clásico.

Un enfoque alternativo: el valor-p

Con la utilización de ordenadores, el contraste de hipótesis puede contemplarse desde otra perspectiva mucho más racional. Así, los programas de ordenador suelen ofrecer, junto al estadístico de contraste una probabilidad. Esta probabilidad, a la cual se le denomina *valor-p* (*p*-value) -es decir, valor de probabilidad-, también es conocida como nivel de significación crítico o exacto, o probabilidad exacta de cometer un *error de tipo* I. Más técnicamente, el *valor-p* se define como el más bajo nivel de significación al que puede ser rechazada una hipótesis nula.

Una vez que el *valor-p* ha sido determinado, sabemos que la hipótesis nula se rechaza para cualquier nivel de significación $\alpha \geq$ *valor-p*; por el contrario, la hipótesis nula no se rechaza cuando $\alpha <$ *valor-p*. Por lo tanto, el *valor-p* es un indicador del nivel de admisibilidad de la hipótesis nula: cuanto mayor sea el *valor-p*, más confianza podemos tener en la hipótesis nula. El uso de *valor-p* cambia por completo el enfoque en el contraste de hipótesis. Así, en lugar de fijar *a priori* el nivel de significación, se calcula el *valor-p*, que nos permite determinar los niveles de significación para los que se rechaza la hipótesis nula.

En las secciones siguientes vamos a ver en la práctica el uso de *valor-p* en el contraste de hipótesis.

4.2 Contraste de hipótesis utilizando el estadístico *t*

4.2.1 Contraste de un solo parámetro

El estadístico t

Bajo los supuestos del *MCL* del 1 al 9,

$$\hat{\beta}_j \sim N\left[\beta_j, \text{var}(\hat{\beta}_j)\right] \qquad j = 1, 2, 3, \cdots, k \tag{4-8}$$

Si tipificamos

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N\left[0,1\right] \qquad j = 1, 2, 3, \cdots, k \tag{4-9}$$

El supuesto de normalidad se apoya en el Teorema del Límite Central (*TCL*), pero este teorema es restrictivo en algunos casos. Es decir, la normalidad no siempre se puede asumir. En cualquier aplicación, asumir o no el supuesto de normalidad de *u* es realmente una cuestión empírica. A menudo, mediante una transformación, por ejemplo, tomando logaritmos, se obtiene una distribución que está más cercana a la normalidad y que es más fácil de manejar desde un punto de vista matemático. Las muestras grandes nos permiten prescindir del supuesto de normalidad sin afectar demasiado a los resultados.

Bajo los supuestos del *MLC* del 1 al 9, se obtiene una distribución *t* de Student

$$\frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \sim t_{n-k} \tag{4-10}$$

donde *k* es el número de parámetros desconocidos en el modelo poblacional (*k*-1 parámetros de pendiente y el término independiente, β_1). La expresión (4-10) es importante porque nos permite contrastar la hipótesis sobre β_j.

Si comparamos (4-10) con (4-9), vemos que la distribución de *t* de Student deriva del hecho de que el parámetro σ de $ee(\hat{\beta}_j)$ ha sido reemplazado por su estimador $\hat{\sigma}$, que es una variable aleatoria. Así pues, los grados de libertad de la *t* son *n-k*, correspondientes a los grados de libertad utilizados en la estimación de $\hat{\sigma}^2$.

Cuando una distribución *t* tiene muchos grados de libertad (*gl*) se aproxima a una distribución normal estándar. En la figura 4.2 se ha representado la función de densidad normal y la función de densidad de la *t* para diferentes grados de libertad. Como puede verse, las funciones de densidad de la *t* son más aplanadas (platicúrticas) y con las colas más anchas que la función de densidad normal, pero a medida que aumentan los *gl*, la función de densidad de la *t* está más próxima de la función de densidad normal. De hecho, lo que pasa es que la distribución *t* tiene en cuenta que σ^2 se ha estimado por ser desconocida. Dada esta incertidumbre, la distribución de la *t* se extiende más que la de la normal. Sin embargo, cuando los *gl* crecen, la distribución *t* está más cerca de la distribución normal porque la incertidumbre de no conocer σ^2 disminuye.

Por lo tanto, debería tenerse en mente la siguiente convergencia en distribución:

$$t_n \xrightarrow[n\to\infty]{} N(0,1) \tag{4-11}$$

Así pues, cuando el número de grados de libertad de una *t* de Student tiende hacia infinito converge hacia una distribución *N*(0,1). En el contexto del contraste de hipótesis, si crece el tamaño de la muestra, también lo harán los grados de libertad. Esto implica que para tamaños grandes se puede utilizar, de forma prácticamente equivalente, la distribución normal para contrastar hipótesis con una sola restricción, aun cuando no se conozca la varianza poblacional. Como regla práctica, cuando los *gl* son mayores que 120, pueden tomarse valores críticos de la distribución normal.

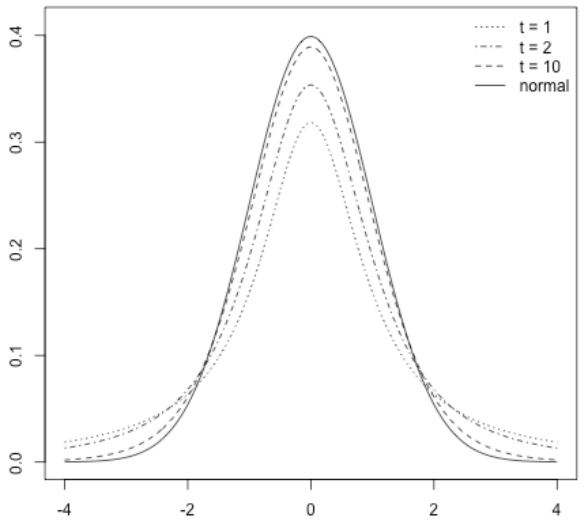


FIGURA 4.2. Funciones de densidad: normal y *t* para diferentes grados de libertad

Considere la hipótesis nula,

$$H_0 : \beta_j = 0$$

Puesto que β_j mide el efecto parcial de x_j sobre y, después de controlar para todas las otras variables independientes, $H_0 : \beta_j = 0$ significa que, una vez que x_2, x_3, …,x_{j-1}, x_{j+1},…, x_k han sido tenidos en cuenta, *x no tiene efecto* sobre *y*. Esta H_0 corresponde al denominado *contraste de significatividad*. El estadístico que se utiliza para contrastar $H_0 : \beta_j = 0$, contra cualquier otra alternativa, se denomina el *estadístico t*, o la *ratio t*, de $\hat{\beta}_j$ y se expresa como

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{ee(\hat{\beta}_j)}$$

Cuando vamos a contrastar $H_0 : \beta_j = 0$, es natural tener presente a $\hat{\beta}_j$ nuestro estimador insesgado de β_j. En una muestra dada $\hat{\beta}_j$ nunca será cero exactamente, pero un valor pequeño indicará una hipótesis nula verdadera, mientras que un valor grande indicará una hipótesis nula falsa. La pregunta es: ¿hasta qué punto $\hat{\beta}_j$ está alejada de cero?

Debemos recordar que hay un error de muestreo en la estimación de $\hat{\beta}_j$, por lo que el tamaño de $\hat{\beta}_j$ debe comparase con su error de muestreo. Esto es precisamente lo que hacemos cuando usamos $t_{\hat{\beta}_j}$, ya que este estadístico mide cuantos errores estándar, está $\hat{\beta}_j$ alejada de cero. A fin de determinar una regla para rechazar H_0, tenemos que decidir sobre la hipótesis alternativa relevante. Hay tres posibilidades: hipótesis alternativas unilaterales (cola derecha e izquierda) e hipótesis alternativa de dos colas.

Hipótesis alternativa de una cola: derecha

En primer lugar, vamos a considerar la hipótesis nula

$$H_0 : \beta_j = 0$$

contra la hipótesis alternativa

$$H_1 : \beta_j > 0$$

Este es un *contraste de significación positiva*. La regla de decisión es en este caso la siguiente:

Regla de decisión

Si	$t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$	se rechaza	H_0	(4-12)
Si	$t_{\hat{\beta}_j} < t^{\alpha}_{n-k}$	no se rechaza	H_0	

Por lo tanto, rechazamos $H_0 : \beta_j = 0$ en favor de cuando $t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$ como puede verse en la figura 4.3. Está muy claro que para rechazar H_0 contra $H_1 : \beta_j > 0$ el valor de $t_{\hat{\beta}_j}$ debe ser positivo. Un resultado negativo de $t_{\hat{\beta}_j}$, no importa lo grande que sea, no proporciona ninguna evidencia a favor de $H_1 : \beta_j > 0$. Por otra parte, para obtener t^{α}_{n-k}

en la tabla estadística de la *t*, sólo necesitamos conocer el nivel de significación α y los grados de libertad. En todo caso, es importante destacar que cuando α disminuye, t_{n-k}^{α} aumenta.

Hasta cierto punto, el enfoque clásico es, en algún sentido arbitrario, puesto que se tiene que elegir un α de antemano, y, dependiendo de esa elección, la H_0 se rechaza o no.

En la figura 4.4 se representa el enfoque alternativo. Como se desprende del examen de la figura, la determinación del *valor-p* es la operación inversa de encontrar el valor en las tablas estadísticas para un determinado nivel de significación. Una vez que el *valor-p* ha sido determinado, sabemos que se rechaza la H_0 para cualquier nivel de significación en que α>*valor-p*, por el contrario, la hipótesis nula no se rechaza cuando α<*valor-p*.

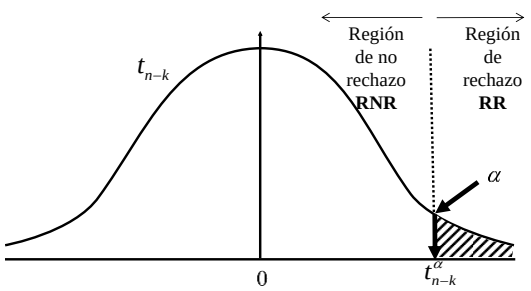


FIGURA 4.3. Región de rechazo utilizando la *t*: hipótesis alternativa de cola a la derecha.

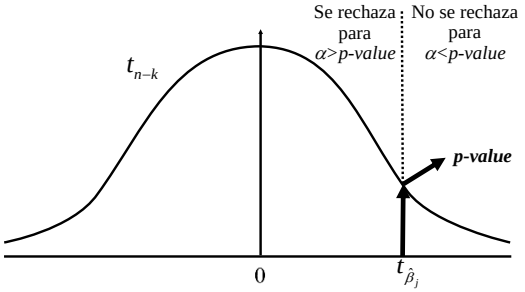


FIGURA 4.4. El *valor-p* utilizando la *t*: hipótesis alternativa de cola a la derecha.

EJEMPLO 4.1 ¿Es la propensión marginal a consumir menor que la propensión media al consumo?

Como se ve en el ejemplo 1.1, contrastar la proposición 3 de la función de consumo keynesiana, en un modelo lineal, es equivalente a contrastar si el término independiente es significativamente mayor que 0. Es decir, en el modelo

$$cons = \beta_1 + \beta_2 inc + u$$

Tenemos que contrastar si

$$\beta_1 > 0$$

Con una muestra aleatoria de 42 observaciones, se han obtenido los siguientes resultados

$$\underset{}{cons_i} = \underset{(0.350)}{0.41} + \underset{(0.062)}{0.843} inc_i$$

Los números entre paréntesis, debajo de los coeficientes, son los errores estándar (*ee*) de los estimadores.

La pregunta que nos planteamos es la siguiente: ¿es la tercera proposición de la teoría keynesiana admisible? A continuación, respondemos a esta pregunta.

1) En este caso, las hipótesis nula y alternativa son las siguientes:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

2) El contraste estadístico es:

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{ee(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{ee(\hat{\beta}_1)} = \frac{0.41}{0.35} = 1.171$$

3) Regla de decisión

Es conveniente utilizar varios niveles de significación. Comencemos con un nivel de significación de 0.10 porque el valor de *t* es relativamente pequeño (menor que 1.5). En este caso, los

grados de libertad son 40 (42 observaciones menos 2 parámetros estimados). Si nos fijamos en la tabla estadística de la *t* (fila 40 y columna de 0.10 o 0.20, en las tablas de una cola, o de dos colas, respectivamente), encontramos que $t_{40}^{0.10} = 1.303$.

Como $t<1.303$, no se rechaza la H_0. Si no rechazamos la H_0 para α=0.10, tampoco se rechazará para α=0.05 ($t_{40}^{0.05} = 1.684$) o α=0.01 ($t_{40}^{0.01} = 2.423$), como puede verse en la figura 4.5. En esta figura la región de rechazo corresponde a α=0.10. Por lo tanto, no se puede rechazar la H_0 en favor de la H_1. En otras palabras, los datos muestrales no son consistentes con la proposición 3 de Keynes.

En el enfoque alternativo, como puede verse en la figura 4.6, el *valor-p* correspondiente a $t_{\hat{\beta}_1}$ =1.171 para una *t* con 40 *gl* es igual a 0.124. Para α<0.124 - por ejemplo, 0.10, 0.05 y 0.01-, no se rechaza la H_0.

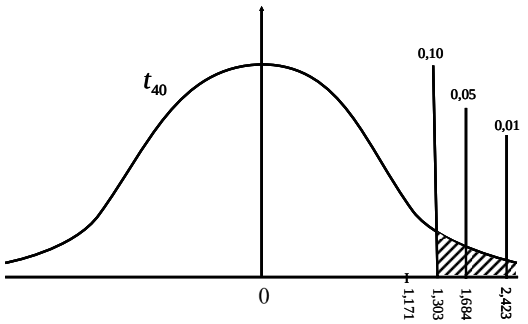


FIGURA 4.5. Ejemplo 4.1: Región de rechazo utilizando la *t* con hipótesis alternativa de cola a la derecha.

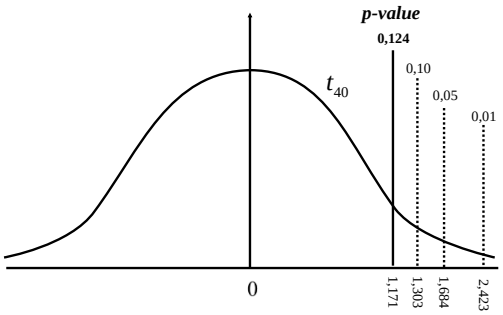


FIGURA 4.6. Ejemplo 4.1: El *valor-p* utilizando la *t* con hipótesis alternativa de cola a la derecha.

Hipótesis alternativa de una cola: izquierda

Consideremos ahora la hipótesis nula

$$H_0 : \beta_j = 0$$

contra la hipótesis alternativa

$$H_1 : \beta_j < 0$$

Este es un *contraste de significación negativa*.

La regla de decisión es en este caso es la siguiente:

Regla de decisión

Si	$t_{\hat{\beta}_j} \leq -t_{n-k}^{\alpha}$	se rechaza	H_0
Si	$t_{\hat{\beta}_j} > -t_{n-k}^{\alpha}$	no se rechaza	H_0

(4-13)

Por lo tanto, rechazamos $H_0 : \beta_j = 0$ en favor de $H_1 : \beta_j < 0$ para un α dado cuando $t_{\hat{\beta}_j} \leq -t_n^{\alpha}$, como puede verse en la figura 4.7. Está muy claro que para rechazar H_0 en contra de $H_1 : \beta_j < 0$, el valor de $t_{\hat{\beta}_j}$ debe ser negativo. Un valor positivo de $t_{\hat{\beta}_j}$, no importa lo grande que sea, no proporciona ninguna evidencia a favor de $H_1 : \beta_j < 0$.

En la figura 4.8 se representa el enfoque alternativo. Una vez que el *valor-p* ha sido determinado, se sabe que se rechaza H_0 para cualquier nivel de significación tal que α>*valor-p*; por el contrario, la hipótesis nula no se rechaza cuando α<*valor-p*.

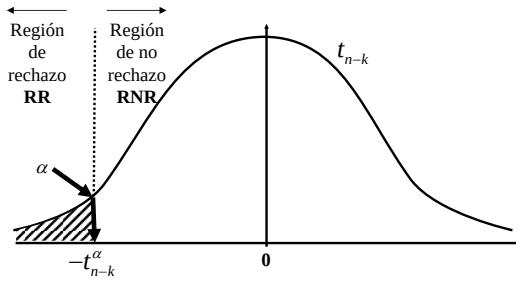


FIGURA 4.7. Región de rechazo utilizando la *t*: hipótesis alternativa de cola a la izquierda.

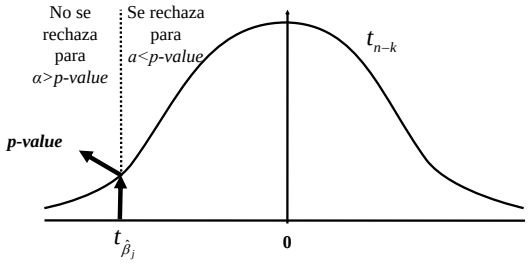


FIGURA 4.8. El *valor-p* utilizando la *t*: hipótesis alternativa de cola a la izquierda.

EJEMPLO 4.2 ¿Tiene la renta una influencia negativa sobre la mortalidad infantil?

El siguiente modelo ha sido planteado para explicar las muertes de niños menores de 5 años por 1000 nacidos vivos (fichero *deathun5*).

$$deathun5 = \beta_1 + \beta_2 gnipc + \beta_3 ilitrate + u$$

donde *gnipc* es la renta nacional bruta per cápita e *ilitrate* es la tasa de analfabetismo de adultos (15 años o más) en porcentaje.

Con una muestra de 130 países (fichero *hdr2010*) se ha realizado la siguiente estimación:

$$\underset{}{deathun5_i} = \underset{(5.93)}{27.91} - \underset{(0.00028)}{0.000826}\, gnipc_i + \underset{(0.183)}{2.043}\, ilitrate_i$$

Los números entre paréntesis, debajo de los coeficientes, son los errores estándar (*ee*) de los estimadores.

Una de las preguntas formuladas por los investigadores es si la renta tiene una influencia negativa sobre la mortalidad infantil. Para responder a esta pregunta se lleva a cabo un contraste en el que las hipótesis nula y alternativa y el estadístico de contraste son los siguientes:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 < 0$$
$$t = \frac{\hat{\beta}_2}{ee(\hat{\beta}_2)} = \frac{-0.000826}{0.00028} = -2.966$$

Dado que el valor de *t* es relativamente alto, vamos a empezar a contrastar con un nivel del 1%. Para α=0.01, $t^{0.01}_{130-2} \approx t^{0.01}_{60} = 2.390$. Teniendo en cuenta que *t*<-2.390, como se muestra en la figura 4.9, se rechaza la H_0 a favor de la H_1. Por lo tanto, la renta nacional bruta per cápita tiene una influencia que es significativamente negativa en la mortalidad de niños menores de 5 años; es decir, cuanto mayor sea la renta nacional bruta per cápita más bajo será el porcentaje de mortalidad de niños menores de 5 años. Como la H_0 se ha rechazado para α=0.01, también será rechazada por los niveles de 5% y 10%.

En el enfoque alternativo, como puede verse en la figura 4.10, el *valor-p* correspondiente a un $t_{\hat{\beta}_1}$=-2.966 para *t* con menos de 61 *gl* es igual a 0.0000. Para todos los α>0.0000 como 0.01, 0.05 y 0.10, se rechaza la H_0.

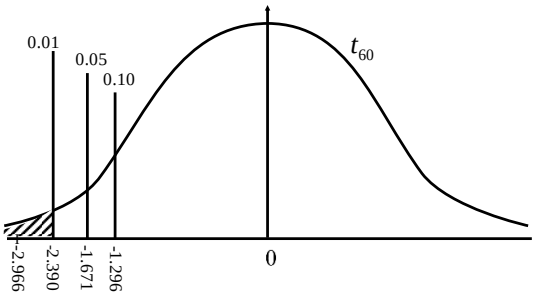


FIGURA 4.9. Ejemplo 4.2: Región de rechazo utilizando la *t* con una hipótesis alternativa de cola a la izquierda.

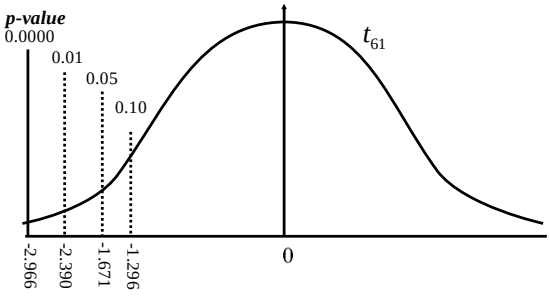
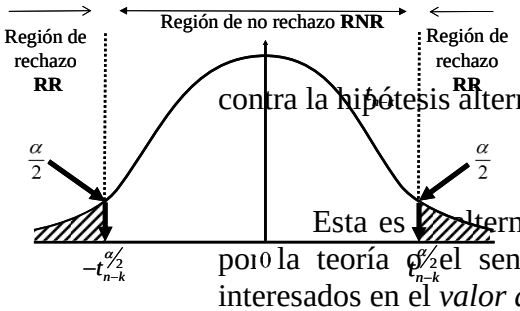


FIGURA 4.10. Ejemplo 4.2: El *valor-p* utilizando la *t* con una hipótesis alternativa de cola a la izquierda.

Hipótesis alternativa con dos colas

Consideremos ahora la hipótesis nula



$$H_0 : \beta_j = 0$$

contra la hipótesis alternativa

$$H_1 : \beta_j \neq 0$$

Esta es la alternativa relevante cuando el signo de β_j no está bien determinado por la teoría o el sentido común. Cuando la alternativa es de dos colas, estamos interesados en el *valor absoluto* del estadístico *t*. Este es un *contraste de significación*.

La regla de decisión en este caso es la siguiente:

Regla de decisión

$$\text{Si} \quad \left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2} \quad \text{se rechaza} \quad H_0$$

$$\text{Si} \quad \left| t_{\hat{\beta}_j} \right| < t_{n-k}^{\alpha/2} \quad \text{no se rechaza} \quad H_0 \qquad (4\text{-}14)$$

Por lo tanto, rechazamos $H_0 : \beta_j = 0$ en favor de $H_1 : \beta_j < 0$ para un α dado cuando $\left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2}$, como puede verse en la figura 4.11. En este caso, para rechazar la H_0 a favor de $H_1 : \beta_j \neq 0$, $t_{\hat{\beta}_j}$ debe ser lo suficientemente grande, bien sea positivo o negativo.

Es importante hacer notar que cuando α decrece, $t_{n-k}^{\alpha/2}$ aumenta en valor absoluto.

En el enfoque alternativo, una vez que el *valor-p* ha sido determinado, se sabe que se rechaza H_0 para cualquier nivel de significación si α>*valor-p*; por el contrario, la hipótesis nula no se rechaza cuando α<*valor-p*. En este caso *valor-p* se distribuye entre las dos colas de forma simétrica, como se muestra en la figura 4.12.

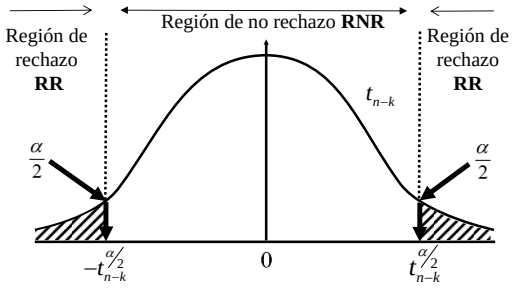


FIGURA 4.11. Región de rechazo usando *t*: hipótesis alternativa de dos colas.

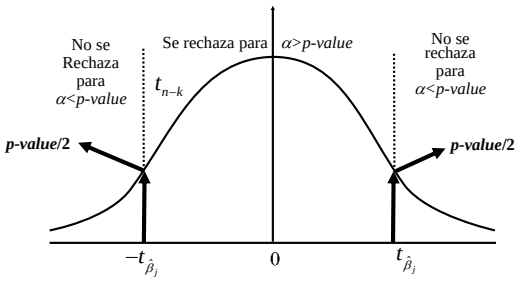


FIGURA 4.12. El *valor-p* usando *t*: hipótesis alternativa de dos colas.

Cuando no se especifica una hipótesis alternativa, por lo general, se considera que el contraste de hipótesis es de dos colas. Si se rechaza la H_0 a favor de la H_1 para un α dado, se suele decir que "x_j *es estadísticamente significativa* para el nivel α".

EJEMPLO 4.3 La tasa de delincuencia, ¿juega un papel en el precio de la vivienda de un área?

Para explicar el precio de la vivienda en una ciudad estadounidense se ha estimado el siguiente modelo:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

donde *rooms* es el número de habitaciones de la casa, *lowstat* es el porcentaje de personas de "clase marginal" en la zona y *crime* son los delitos cometidos per capita en la zona.

Los resultados del modelo ajustado realizado con E-views, utilizando el fichero *hprice2* (primeras 55 observaciones), aparece en el cuadro 4.2. El significado de las tres primeras columnas es claro: "Estadístico *t*" es el dato requerido para hacer un contraste de significación, es decir, es la relación entre el "Coeficiente" y el "Error estándar", y "Prob" es el *valor-p* para un contraste de dos colas.

En relación con este modelo la pregunta que se hacen los investigadores es si la tasa de criminalidad juega un papel importante en el precio de las casas de la zona. Para responder a esta pregunta, se ha llevado a cabo el siguiente procedimiento.

En este caso, la hipótesis nula y alternativa, y el estadístico de contraste, son los siguientes:

$$H_0 : \beta_4 = 0 \qquad t = \frac{\hat{\beta}_4}{ee(\hat{\beta}_4)} = \frac{-3854}{960} = -4.016$$
$$H_1 : \beta_4 \neq 0$$

CUADRO 4.2. Salida estándar en una regresión para explicar el precio de una casa. *n*=55.

Variable	*Coeficiente*	*Error estándar*	*Estadístico t*	*Prob.*
C	-15693.61	8021.989	-1.956324	0.0559
rooms	6788.401	1210.720	5.606910	0.0000
lowstat	-268.1636	80.70678	-3.322690	0.0017
crime	-3853.564	959.5618	-4.015962	0.0002

Dado que el valor de *t* es relativamente alto, vamos a empezar a contrastar para un nivel del 1%. Para α=0.01, $t_{51}^{0.01/2} \approx t_{50}^{0.01/2} = 2.69$. (En las tablas estadísticas habituales para la distribución *t*, no hay información para cada *gl*, uno a uno, más allá de 20). Teniendo en cuenta que $|t| > 2.69$, rechazamos la H_0 a favor de la H_1. Por lo tanto, la delincuencia tiene una influencia significativa en el precio de la vivienda con un nivel de significación del 1% y, por tanto, de un 5% y 10%.

En el enfoque alternativo podemos realizar el contraste con mayor precisión. En el cuadro 4.2, vemos que el valor de *valor-p* para el coeficiente *crime* es de 0.0002. Eso significa que la probabilidad de que el estadístico *t* sea mayor de 4.016 es 0.0001 y la probabilidad de que *t* sea menor de -4.016 es de 0.0001. Es decir, el valor de *valor-p*, como se muestra en la figura 4.13 se distribuye en los dos lados. Como puede verse en esta figura, se rechaza H_0 para todos los niveles de significación superiores a 0.0002, tales como 0.01, 0.05 y 0.10. Si se tratara de un contraste de una cola, en el enfoque alternativo el *valor-p* sería igual a 0.0002/1=0.0001.

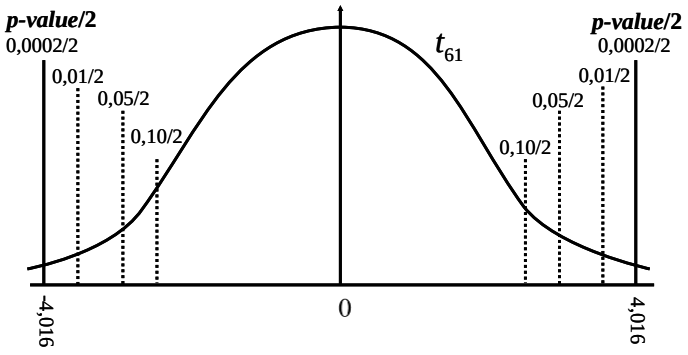


FIGURA 4.13. Ejemplo 4.3: El *valor-p* utilizando *t:* hipótesis alternativa de dos colas.

Hasta ahora hemos visto el contraste significativo de una cola y de dos colas en el que un parámetro toma el valor 0 en la H_0. Ahora vamos a ver un caso más general en que el parámetro en la H_0 toma un valor específico cualquiera:

$$H_0 : \beta_j = \beta_j^0$$

En este caso, el estadístico *t* adecuado es

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^0}{ee(\hat{\beta}_j)}$$

Al igual que antes, $t_{\hat{\beta}_j}$ mide la cantidad de desviaciones estándar está $\hat{\beta}_j$ distanciada de β_j^0, valor que toma el parámetro en la hipótesis nula.

EJEMPLO 4.4 ¿Es la elasticidad del gasto en frutas/renta igual a 1? ¿Es la fruta un bien de lujo?

Para responder a estas dos preguntas vamos a utilizar el siguiente modelo para explicar el gasto en fruta (*fruit*):

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 househsize + \beta_4 punders + u$$

donde *inc* es la renta disponible de los hogares, *househsize* es el número de miembros de la familia y *punder5* es la proporción de niños menores de cinco años en el hogar.

Dado que las variables *fruit* e *inc* aparecen expresadas en logaritmos naturales, entonces β_2 es la elasticidad del gasto/renta. Utilizando una muestra de 40 hogares (fichero *demand*), se han obtenido los resultados del cuadro 4.3.

CUADRO 4.3. Salida estándar de una regresión que explica los gastos en fruta. *n*=40.

Variable	*Coeficiente*	*Error estándar*	*Estadístico t*	*Prob.*
C	-9.767654	3.701469	-2.638859	0.0122
ln(inc)	2.004539	0.512370	3.912286	0.0004
househsize	-1.205348	0.178646	-6.747147	0.0000
punder5	-0.017946	0.013022	-1.378128	0.1767

¿Es la elasticidad del gasto en frutas/ renta igual a 1?

Para responder a esta pregunta, se ha llevado a cabo el siguiente procedimiento. En este caso las hipótesis nula y alternativa, y el estadístico de contraste, son los siguientes:

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 \neq 1$$

$$t = \frac{\hat{\beta}_2 - \beta_2^0}{ee(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 1}{ee(\hat{\beta}_2)} = \frac{2.005 - 1}{0.512} = 1.961$$

Para α=0.10, nos encontramos con que $t_{36}^{0.10/2} \approx t_{35}^{0.10/2} = 1.69$. Como $|t|$>1.69 se rechaza H_0. Para α=0.05, $t_{36}^{0.05/2} \approx t_{35}^{0.05/2} = 2.03$. Como $|t|$<2.03 no rechazamos H_0 para α=0.05, ni para α=0.01. Por lo tanto, rechazamos que la elasticidad del gasto en fruta/renta sea igual a 1 para α=0.10, pero no rechazamos para α=0.05, ni para α=0.01.

¿Es la fruta un bien de lujo?

Según la teoría económica, una mercancía es un bien de lujo cuando la elasticidad del gasto con respecto a la renta es mayor que 1. Por lo tanto, para responder a esta segunda cuestión, y teniendo en cuenta que el estadístico *t* es el mismo, se ha llevado a cabo el siguiente procedimiento:

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 > 1.$$

Para α=0.10, nos encontramos con que $t_{36}^{0.10} \approx t_{35}^{0.10} = 1.31$.Cuando *t*>1.31, rechazamos H_0 en favor de H_1. Para α=0.05, $t_{36}^{0.05} \approx t_{35}^{0.05} = 1.69$.Como *t*>1.69, rechazamos la H_0 en favor de la H_1. Para α=0.01, $t_{36}^{0.01} \approx t_{35}^{0.01} = 2.44$. Como *t*<2.44, no rechazamos H_0. Por lo tanto, la fruta es un bien de lujo para α=0.10 y α=0.05, pero no se puede rechazar la H_0 en favor de la H_1 para α=0.01.

EJEMPLO 4.5 ¿Es la Bolsa de Madrid un mercado eficiente?

Antes de responder a la cuestión planteada, vamos a examinar algunos conceptos previos. La *tasa de rendimiento de un activo* en un período de tiempo se define como la variación porcentual que experimenta el valor invertido en ese activo durante dicho período de tiempo. Vamos a considerar ahora como activo específico, a una acción de una compañía industrial adquirida en una Bolsa española al final de un año y que se mantiene hasta el final del año siguiente. A esos dos momentos de tiempo los designaremos *t*-1 y *t* respectivamente. La tasa de rendimiento de esa acción al cabo de ese año puede expresarse mediante la siguiente relación:

$$RA_t = \frac{\Delta P_t + D_t + A_t}{P_{t-1}} \tag{4-15}$$

donde

P_t: es la cotización de la acción al final del período t

D_t: son los dividendos percibidos por la acción durante el período t

A_t: es el valor de los derechos que eventualmente han podido corresponder a la acción durante el período t

Así pues, en el numerador de (4-15) se recogen los tres tipos de ganancias de capital que se han podido percibir por el mantenimiento de una acción durante el año t : incremento - o pérdida en su caso – en la cotización, dividendos y derechos de ampliación. Al dividir por P_{t-1} se obtiene la tasa de ganancia sobre el valor de la acción a finales del período anterior. De los tres componentes el más importante es el incremento en la cotización. Teniendo en cuenta solamente a esa componente, la tasa de rendimiento de la acción puede expresarse por

$$RA1_t = \frac{\Delta P_t}{P_{t-1}} \tag{4-16}$$

o, alternativamente si utilizamos una tasa de variación natural, por

$$RA2_t = \Delta \ln P_t \tag{4-17}$$

De la misma forma que RA_t, en cualquiera de las dos expresiones, representa la tasa de rendimiento de una acción concreta, se puede estimar también la tasa de rendimiento del conjunto de acciones cotizadas en Bolsa. A esta última tasa de rendimiento, a la que designaremos por RM_t, se le denomina tasa de rendimiento de mercado.

Hasta ahora hemos considerado la tasa de rendimiento en un año, pero igualmente se puede aplicar expresiones del tipo (4-16), o (4.17), para obtener tasas de rendimiento diario. Analizando el comportamiento de estas tasas cabe preguntarse si las tasas de rendimiento en el pasado son de utilidad para predecir las tasas de rendimiento en el futuro. Esta pregunta está relacionada con el concepto de eficiencia de un mercado. Un mercado es *eficiente* si los precios incorporan toda la información disponible, de modo que hay posibilidad de obtener ganancias extraordinarias al utilizar esta información.

Para contrastar la eficiencia de un mercado vamos a definir el siguiente modelo, utilizando tasas de rendimiento diarias definidas según (4-16):

$$rmad92_t = \beta_1 + \beta_2 rmad92_{t-1} + u_t \tag{4-18}$$

Si un mercado es eficiente, entonces el parámetro β_2 del anterior modelo debe ser 0. Vamos a contrastar ahora si la Bolsa de Madrid, en lo que respecta a la renta variable, es o no eficiente en su conjunto.

El modelo de (4-18) se ha estimado con datos diarios de la Bolsa de Madrid para el año 1992, utilizando el fichero *bolmadef*. Los resultados obtenidos han sido los siguientes:

$$rmad92_t = \underset{(0.0007)}{-0.0004} + \underset{(0.0629)}{0.1267}\, rmad92_{t-1}$$

$$R^2=0.0163 \qquad n=247$$

Los resultados obtenidos pueden resultar paradójicos. Por una parte, el valor del coeficiente de determinación es muy bajo (0.0163), lo que significa que solamente el 1.63% de la varianza total de la tasa de rendimiento se explica por la tasa de rendimiento del día anterior. Por otra parte, sin embargo, el coeficiente correspondiente a la tasa de significación del día anterior es estadísticamente significativo para un nivel del 5% pero no para un nivel de 1%, porque el estadístico t es igual a 0.1267/0.0629=2.02 que es ligeramente mayor en valor absoluto que $t_{245}^{0.01} \simeq t_{60}^{0.01}$=2.00. El motivo de esta aparente paradoja se debe a que el tamaño de la muestra es muy elevado. Así, aunque la incidencia de la variable explicativa sobre la variable endógena es relativamente reducida (como indica el coeficiente de determinación), sin embargo, esta incidencia es significativa (como lo confirma el estadístico t) debido a que se ha dispuesto de una muestra de datos suficientemente grande.

Contestando a la pregunta de si la Bolsa de Madrid es o no un mercado eficiente, en un primer análisis la respuesta es que no es totalmente eficiente. Sin embargo, esta respuesta debe matizarse. En economía financiera es conocida la existencia de una relación de dependencia entre la tasa de rendimiento de un día y la del día precedente. Esta relación no es muy fuerte, aunque sí es estadísticamente significativa en muchas bolsas mundiales, y se debe a las fricciones del mercado. En cualquier caso, este fenómeno no se puede explotar de forma lucrativa por los agentes del mercado, por lo que no se puede calificar a estos mercados de ineficientes, de acuerdo con la definición dada anteriormente sobre el concepto de eficiencia.

EJEMPLO 4.6 La rentabilidad de la Bolsa de Madrid, ¿se ve afectada por la rentabilidad de la Bolsa de Tokio?

El estudio de la relación entre distintos mercados de acciones (Bolsa de Nueva York, Bolsa de Tokio, Bolsa de Madrid, Bolsa de Londres, etc.) ha recibido una gran atención en los últimos años, debido a una mayor libertad en la circulación de capitales y a la conveniencia de utilizar mercados extranjeros para reducir el riesgo en la gestión de carteras, ya que la ausencia de una perfecta integración de los mercados permite la diversificación del riesgo. De todas formas, cada vez se camina hacia una mayor integración mundial de los mercados financieros, en general, y de los mercados de acciones, en particular.

Si los mercados son eficientes, y hemos visto en el ejemplo 4.5 que se puede admitir que lo sean, las noticias que se van produciendo, a las que se denominan *innovaciones*, durante un período de 24 horas se irán viendo reflejadas en los distintos mercados.

Conviene distinguir entre dos tipos de innovaciones: a) *innovaciones globales*, que son noticias que se generan alrededor del mundo y que se captan en los precios de las acciones en todos los mercados; b) *innovaciones específicas*, que es la información generada durante un período de 24 horas que solo afecta a los precios de un mercado particular. Así, la información sobre evolución de los precios del petróleo se puede considerar como una innovación global, mientras que una nueva regulación del sector financiero en un país sería considerada posiblemente como una innovación específica.

De acuerdo con la exposición anterior, en una sesión de Bolsa de un determinado mercado, los precios de las acciones que en él se cotizan vendrán afectados por las innovaciones globales recogidas en otro mercado que haya cerrado antes. Así, las innovaciones globales recogidas en el mercado de Tokio influirán en las cotizaciones del mercado de Madrid de ese mismo día. El siguiente modelo recoge la transmisión de efectos entre la Bolsa de Tokio y la Bolsa de Madrid:

$$rmad92_t = \beta_1 + \beta_2 rtok92_t + u_t \tag{4-19}$$

donde $rmad92_t$ es la tasa de rentabilidad de la Bolsa Madrid en el período *t*, y $rtok92_t$ es la tasa de rentabilidad de la Bolsa de Tokio en el período *t*. Las rentabilidades de los mercados se han calculado de acuerdo con (4-19).

En el fichero *madtok* se pueden encontrar los índices generales de la Bolsa de Madrid y la Bolsa de Valores de Tokio durante los días en que ambas bolsas estaban abiertas de forma simultánea. Es decir, se han eliminado las observaciones de los días en los que cualquiera de las bolsas estuviese cerrada. En total, el número de observaciones es de 234, en comparación con los 247 y 246 días en que las Bolsa de Madrid y de Tokio estuvieron abiertas respectivamente.

La estimación del modelo (4-19) es como sigue:

$$\widehat{rmad92}_t = \underset{(0.0007)}{-0.0005} + \underset{(0.0375)}{0.1244}\, rtok92_t$$

$$R^2 = 0.0452 \qquad n = 235$$

Obsérvese que el coeficiente de determinación es relativamente bajo. Sin embargo, para contrastar H_0: $\beta_2=0$, se obtiene que el estadístico t = (0.1244/0.0375) = 3.32 lo que implica que se rechaza la hipótesis nula de que la tasa de rendimiento de la Bolsa de Valores de Tokio no tenga ningún efecto sobre la tasa de rentabilidad de la Bolsa de Madrid, para un nivel de significación de 0.01.

Una vez más nos encontramos con la misma paradoja aparente que apareció cuando se analizó la eficiencia de la Bolsa de Madrid en el ejemplo 4.5 a excepción de una diferencia. En este último caso, la tasa de rendimiento del día anterior aparecía como significativa, debido a problemas derivados de la elaboración del índice general de la Bolsa de Madrid.

En consecuencia, el hecho de que la hipótesis nula sea rechazada implica que hay evidencia empírica que apoya la teoría de que las innovaciones globales de la Bolsa de Tokyo se transmiten a las cotizaciones de la Bolsa de Madrid ese mismo día.

4.2.2 Los intervalos de confianza

Bajo los supuestos del *MLC*, es fácil construir un *intervalo de confianza* (*IC*) para el parámetro de la población, β_j. A los *IC* también se les denomina estimaciones por intervalo, ya que proporcionan un rango de valores verosímiles para β_j, y no solamente una estimación puntual.

El *IC* está construido de tal manera que el parámetro desconocido está contenido dentro del recorrido del *IC* con una probabilidad previamente especificada.

Utilizando el hecho de que

$$\frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \sim t_{n-k}$$

$$\Pr\left[-t_{n-k}^{\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \leq t_{n-k}^{\alpha/2} \right] = 1-\alpha$$

Operando para situar al parámetro β_j solo en el centro del intervalo, obtenemos que

$$\Pr\left[\hat{\beta}_j - ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2} \leq \beta_j \leq \hat{\beta}_j + ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2} \right] = 1-\alpha$$

Por lo tanto, los límites inferior y superior, respectivamente, de un *IC* de probabilidad (1-α) están dados por

$$\underline{\beta}_j = \hat{\beta}_j - ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2}$$

$$\overline{\beta}_j = \hat{\beta}_j + ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2}$$

Si las muestras se obtuvieron al azar de forma repetida calculando $\underline{\beta}_j$ y $\overline{\beta}_j$ cada vez, el parámetro poblacional (desconocido) caería en el intervalo $(\underline{\beta}_j, \overline{\beta}_j)$ en un $(1-\alpha)$% de las muestras. Desafortunadamente, para la muestra individual que se utiliza en la construcción del *IC*, no sabemos si β_j está o no realmente contenida dentro del intervalo.

Una vez que el *IC* se ha construido, es fácil llevar a cabo contrastes de hipótesis de dos colas. Si la hipótesis nula es $H_0 : \beta_j = a_j$, entonces la H_0 se rechaza contra la $H_1 : \beta_j \neq a_j$ para el nivel de significación del 5%, si y sólo si, a_j *no está* en el *IC* del 95%.

Para ilustrar todo lo anterior, en la figura 4.14 se han construidos intervalos de confianza del 90%, 95% y 99%, para la propensión marginal al consumo -β_2- correspondiente al ejemplo 4.1.

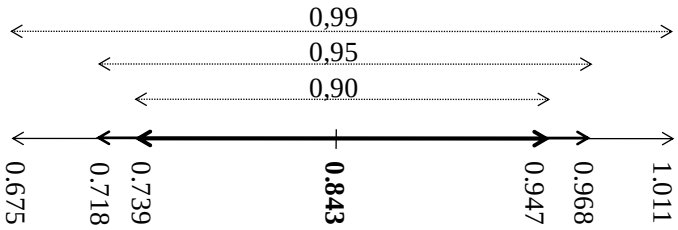


FIGURA 4.14. Intervalos de confianza para la propensión marginal al consumo en el ejemplo 4.1.

4.2.3 Contraste de hipótesis sobre una combinación lineal de parámetros

En muchas aplicaciones, estamos interesados en contrastar hipótesis en las que están implicados más de un parámetro poblacional. El estadístico *t* también se puede utilizar para contrastar una combinación de parámetros, en la que dos o más parámetros están implicados.

El contraste sobre una combinación lineal de parámetros se puede hacer por dos procedimientos diferentes. En el primer procedimiento, el error estándar de la combinación lineal de los parámetros correspondientes a la hipótesis nula se calcula utilizando información sobre la matriz de covarianza de los estimadores. En el segundo procedimiento, el modelo se reparametriza mediante la introducción de un nuevo parámetro deducido de la hipótesis nula, después se estima el modelo reparametrizado y el contraste del nuevo parámetro nos indica sí se rechaza, o no, la hipótesis nula. El siguiente ejemplo ilustra ambos procedimientos.

EJEMPLO 4.7 ¿Hay rendimientos constantes a escala en el sector de metales primario?

Para examinar si hay rendimientos constantes a escala en el sector de metales primario se va a utilizar la función de producción Cobb-Douglas, dada por

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u \qquad (4\text{-}20)$$

En el anterior modelo los parámetros β_2 y β_3 son elasticidades (producción/trabajo y producción/capital).

Antes de hacer inferencias hemos de recordar que los *rendimientos a escala* se refieren a una característica técnica de la función de producción que analiza los cambios en la producción debidas a un cambio en la misma proporción de todos los inputs, que en este caso son el trabajo y el capital. Si la producción cambia en la misma proporción que los inputs entonces se dice que hay *rendimientos constantes a escala*. Los rendimientos constantes a escala implican que si los factores *trabajo y capital* aumentan a una cierta tasa (digamos el 10%), la producción aumentará en la misma proporción (esto es en un 10%). Si la producción aumenta en una mayor proporción, entonces hay *rendimientos crecientes a escala*. Si aumenta la producción en una menor proporción, existen *rendimientos decrecientes a escala*. En el modelo anterior, sucede que:

- si $\beta_2+\beta_3=1$, hay *rendimientos constantes a escala*.

- si $\beta_2+\beta_3>1$, hay *rendimientos crecientes a escala*.

- si $\beta_2+\beta_3<1$, hay *rendimientos decrecientes a escala*

Los datos utilizados en este ejemplo son una muestra de 27 empresas del sector de metales primario (fichero *prodmet*), donde *output* es el valor añadido bruto, *labor* es una medida de la mano de obra y *capital* es el valor bruto de la planta y equipo. Detalles adicionales sobre la construcción de estos datos se ofrecen en Aigne *et al.* (1977) y en Hildebrand y Liu (1957). Los resultados obtenidos en la estimación del modelo (4-20) aparecen en el cuadro 4.4.

CUADRO 4.4. Salida estándar de la estimación de la función de producción: modelo(4-20)..

Variable	*Coeficiente*	*Error estándar*	*Estadístico t*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	0.602999	0.125954	4.787457	0.0001
ln(*capital*)	0.375710	0.085346	4.402204	0.0002

Para responder a la pregunta planteada en este ejemplo, tenemos que contrastar:

$$H_0 : \beta_2 + \beta_3 = 1$$

contra la hipótesis alternativa siguiente:

$$H_1 : \beta_2 + \beta_3 \neq 1$$

De acuerdo con la H_0, se deduce que $\beta_2 + \beta_3 - 1 = 0$. Por lo tanto, el estadístico *t* se basa ahora en si la suma estimada $\hat{\beta}_2 + \hat{\beta}_3 - 1$ es lo suficientemente diferente de 0 como para rechazar la H_0 a favor de la H_1.

Para contrastar esta hipótesis se van a utilizar dos procedimientos. En el primer procedimiento se utiliza la matriz de covarianzas de los estimadores. En el segundo, el modelo se reparametriza introduciendo un nuevo parámetro.

Procedimiento que utiliza la matriz de covarianzas de los estimadores

De acuerdo con H_0, se estima que $\beta_2 + \beta_3 - 1 = 0$. Por lo tanto, el estadístico *t* está basado ahora en si la suma estimada $\hat{\beta}_2 + \hat{\beta}_3 - 1$es lo suficientemente diferente de 0 para rechazar la H_0 a favor de la H_1. Para tener en cuenta el error de muestreo en nuestros estimadores, se estandariza esta suma dividiendo por su error estándar:

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{ee(\hat{\beta}_2 + \hat{\beta}_3)}$$

Así que, si $t_{\hat{\beta}_2+\hat{\beta}_3}$ es lo suficientemente grande, vamos a concluir, en un contraste de dos colas, que *no hay rendimientos constantes a escala*. Por otro lado, si $t_{\hat{\beta}_2+\hat{\beta}_3}$ es positivo y lo suficientemente grande, vamos a rechazar, en un contraste alternativo de una cola (la derecha), a H_0 en favor de $H_1 : \beta_2 + \beta_3 > 1$, concluyendo que sí hay rendimientos crecientes a escala.

Por otro parte, tenemos que

$$ee(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\text{var}(\hat{\beta}_2 + \hat{\beta}_3)}$$

donde

$$\text{var}(\hat{\beta}_2 + \hat{\beta}_3) = \text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) + 2 \times \text{covar}(\hat{\beta}_2, \hat{\beta}_3)$$

Por lo tanto, para estimar $ee(\hat{\beta}_2 + \hat{\beta}_3)$ se necesita información sobre la covarianza estimada de los estimadores. Muchos paquetes de software econométrico, como el E-Views, tiene una opción para mostrar las estimaciones de la matriz de covarianzas del vector de los estimadores. En este caso, la matriz de covarianzas obtenida aparece en el cuadro 4.5. Con esta información se tiene que

$$ee(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{0.015864 + 0.007284 - 2 \times 0.009616} = 0.0626$$

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{ee(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{-0.02129}{0.0626} = -0.3402$$

CUADRO 4.5. Matriz de covarianzas de la función de producción.

	constante	ln(*labor*)	ln(*capital*)
constant	0.106786	-0.019835	0.001189
ln(*labor*))	-0.019835	0.015864	-0.009616
ln(*capital*)	0.001189	-0.009616	0.007284

Teniendo en cuenta que t=0.3402, es evidente que no podemos rechazar la existencia de rendimientos constantes de escala, para los niveles de significación habituales. Dado que el estadístico t toma un valor negativo, no tiene sentido contrastar si existen rendimientos crecientes a escala.

Procedimiento en el que se reparametriza el modelo mediante la introducción de un nuevo parámetro

La aplicación de este segundo procedimiento es una forma más fácil de realizar este contraste. En este procedimiento se estima un modelo diferente que proporciona directamente el error estándar en que estamos interesados. Así, en el ejemplo anterior vamos a definir:

$$\theta = \beta_2 + \beta_3 - 1$$

por lo que la hipótesis nula de que *hay rendimientos constantes a escala* es equivalente a postular que $H_0 : \theta = 0$.

De la definición de θ se tiene que $\beta_2 = \theta - \beta_3 + 1$. Sustituyendo β_2 en la ecuación original:

$$\ln(output) = \beta_1 + (\theta - \beta_3 + 1)\ln(labor) + \beta_3 \ln(capital) + u$$

Por lo tanto,

$$\ln(output / labor) = \beta_1 + \theta \ln(labor) + \beta_3 \ln(capital / labor) + u$$

Así pues, contrastar si existen rendimientos constantes a escala es equivalente a realizar un contraste de significación del coeficiente ln(*labor*) en el modelo anterior. La estrategia de reescribir el modelo, de modo que contenga el parámetro de interés, funciona en todos los casos y normalmente es fácil de implementar. Si aplicamos esta transformación en este ejemplo, se obtienen los resultados del cuadro 4.6.

Como puede verse se obtiene el mismo resultado:

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{ee(\hat{\theta})} = -0.3402$$

CUADRO 4.6. Salida de la estimación de la función de producción: modelo reparametrizado.

Variable	*Coeficiente*	*Error estándar*	*Estadístico t*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	-0.021290	0.062577	-0.340227	0.7366
ln(*capital/labor*)	0.375710	0.085346	4.402204	0.0002

EJEMPLO 4.8 ¿Publicidad o incentivos?

La Compañía Bush se dedica a la venta y distribución de regalos importados de Oriente Próximo. El artículo más popular en el catálogo es la pulsera de Guantánamo. Tiene algunas propiedades relajantes. Los agentes de ventas reciben una comisión del 30% del importe total de las ventas. Con el fin de aumentar las ventas sin necesidad de expandir la red comercial, la compañía estableció incentivos especiales para aquellos agentes que rebasen el objetivo de ventas durante el último año.

Por otra parte, se emiten anuncios publicitarios en radio en diferentes regiones para fortalecer la promoción de las ventas. En esos lugares se hizo especial hincapié en destacar el bienestar de llevar un brazalete de Guantánamo.

El gerente de la Empresa Bush se pregunta si un dólar gastado en incentivos especiales tiene, o no, una mayor incidencia en las ventas que un dólar gastado en publicidad. Para responder a esa pregunta el económetra de la compañía sugiere el siguiente modelo para explicar las ventas (*sales*):

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u$$

donde *incent* son los incentivos a los vendedores y *advert* son los gastos en publicidad. Las variables *sales*, *incent* y *advert* están expresadas en miles de dólares.

Utilizando una muestra de 18 áreas de venta (fichero *advincen*), se han obtenido los resultados de la regresión y la matriz de covarianzas de los coeficientes que aparecen en los cuadros 4.7 y 4.8, respectivamente.

CUADRO 4.7. Salida estándar de la regresión para el ejemplo 4.8.

Variable	*Coeficiente*	*Error estándar*	*Estadístico t*	*Prob.*
constant	396.5945	3548.111	0.111776	0.9125
advert	18.63673	8.924339	2.088304	0.0542
incent	30.69686	3.604420	8.516448	0.0000

CUADRO 4.8. Matriz de covarianzas para el ejemplo 4.8.

	C	*advert*	*incent*
constant	12589095	-26674	-7101
advert	-26674	79.644	2.941
incent	-7101	2.941	12.992

En este modelo, el coeficiente β_2 indica el aumento de las ventas producidas por un dólar de incremento en el gasto en publicidad, mientras β_3 indica el aumento que se produce en las ventas por un dólar de incremento en los incentivos especiales, manteniendo fijo en ambos casos el otro regresor.

Para responder a la pregunta planteada en este ejemplo, la hipótesis nula y la hipótesis alternativa son las siguientes:

$$H_0 : \beta_3 - \beta_2 = 0$$
$$H_1 : \beta_3 - \beta_2 > 0$$

El estadístico *t* se construye utilizando la información sobre la matriz de covarianzas de los estimadores:

$$t_{\hat{\beta}_3-\hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{ee(\hat{\beta}_3 - \hat{\beta}_2)}$$

$$ee(\hat{\beta}_3 - \hat{\beta}_2) = \sqrt{79.644 + 12.992 - 2 \times 2.941} = 9.3142$$

$$t_{\hat{\beta}_3-\hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{ee(\hat{\beta}_3 - \hat{\beta}_2)} = \frac{30.697 - 18.637}{9.3142} = 1.295$$

Para α=0.10, nos encontramos con que $t_{15}^{0.10} = 1.341$. Como t<1.341, no rechazamos H_0 para α=0.10, ni para α=0.05 o α=0.01. Por lo tanto, no hay evidencia empírica de que un dólar gastado en incentivos especiales tenga una mayor incidencia en las ventas que un dólar gastado en publicidad.

EJEMPLO 4.9 Contraste de la hipótesis de homogeneidad en la demanda de pescado

En el caso de estudio del capítulo 2 fueron estimados diversos modelos, utilizando datos de corte transversal, para explicar la demanda de productos lácteos en los que la renta disponible era la única variable explicativa. Sin embargo, el precio del producto estudiado y, en mayor o menor medida, los precios de otros productos son determinantes en la demanda. El análisis de la demanda sobre la base de datos de corte transversal, tiene precisamente la limitación de que no es posible examinar el efecto de los precios en la demanda porque los precios se mantienen constantes, ya que los datos se refieren todos al mismo punto en el tiempo. Para analizar el efecto de los precios es necesario el uso de datos de series temporales o, alternativamente, de datos de panel. A continuación, vamos a examinar, brevemente, algunos aspectos de la teoría de la demanda de un bien para luego pasar a la estimación de una función de demanda con datos de series temporales. Como colofón a este caso, vamos a contrastar una de las hipótesis que, en determinadas circunstancias, debe satisfacer un modelo teórico.

La demanda de un bien -como el bien *j*- se puede expresar, de acuerdo con un proceso de optimización llevado a cabo por el consumidor, en términos de renta disponible, del precio de la mercancía y de los precios del resto de bienes. Analíticamente:

$$q_j = f_j(p_1, p_2, \cdots, p_j, \cdots, p_m, R) \tag{4-21}$$

donde

- *R* es la renta disponible de los consumidores.
- $p_1, p_2, \cdots, p_j, \cdots p_m$ son los precios de los bienes que se tienen en cuenta por los consumidores cuando adquieren el bien *j*.

En los estudios de demanda, los modelos logarítmicos son atractivos, ya que los coeficientes son directamente elasticidades. El modelo logarítmico se expresa de la siguiente forma:

$$\ln(q_j) = \beta_1 + \beta_2 \ln(p_1) + \beta_3 \ln(p_2) + \cdots + \beta_j \ln(p_j) + \cdots + \beta_{m+1} \ln(p_m) + \beta_{m+2} \ln(R) + u \qquad (4\text{-}22)$$

Como puede verse de forma inmediata, todos los coeficientes β, excluyendo el término constante, son elasticidades de diferentes tipos y, por lo tanto, son independientes de las unidades de medida de las variables. Cuando no hay ilusión monetaria, si todos los precios y la renta crecen a la misma tasa, la demanda de un bien no se ve afectada por estos cambios. Por lo tanto, suponiendo que los precios y la renta se multiplican por λ, si el consumidor no tiene ilusión monetaria, se debe satisfacer que

$$f_j(\lambda p_1, \lambda p_2, \cdots, \lambda p_j, \cdots, \lambda p_m, \lambda R) = f_j(p_1, p_2, \cdots, p_j, \cdots p_m, R) \qquad (4\text{-}23)$$

Desde un punto de vista matemático, la condición anterior implica que la función de demanda debe ser homogénea de grado 0. Esta condición se llama la *restricción de homogeneidad*. Aplicando el teorema de Euler, la restricción de homogeneidad, a su vez implica que la suma de la elasticidad de demanda/renta y de todas las elasticidades de demanda/precio es cero, es decir:

$$\sum_{h=1}^{m} \varepsilon_{q_j/p_h} + \varepsilon_{q_j/R} = 0 \qquad (4\text{-}24)¡$$

Esta restricción aplicada al modelo logarítmico (4-22) implica que

$$\beta_2 + \beta_3 + \cdots + \beta_j + \cdots + \beta_{m+1} + \beta_{m+2} = 0 \qquad (4\text{-}25)$$

En la práctica, cuando se estima una función de demanda, los precios de muchos bienes no están incluidos, sino sólo aquellos que están estrechamente relacionados, ya sea por ser complementarios o por ser sustitutivos del bien estudiado. También es bien sabido que la asignación presupuestaria del gasto se suele realizar en varias etapas.

A continuación, estudiaremos la demanda de pescado en España, utilizando un modelo similar a (4-22). Tengamos en cuenta que en una primera asignación, el consumidor distribuye su renta entre el consumo total y el ahorro. En una segunda etapa, el gasto de consumo por función se lleva a cabo teniendo en cuenta el consumo total y los precios relevantes en cada función. En concreto, en la demanda de pescado se ha supuesto que sólo es relevante el precio del pescado y el precio de la carne que es el sustitutivo más importante.

Teniendo en cuenta las consideraciones anteriores, se ha formulado el siguiente modelo:

$$\ln(fish) = \beta_1 + \beta_2 \ln(fishpr) + \beta_3 \ln(meatpr) + \beta_4 \ln(cons) + u \qquad (4\text{-}26)$$

donde *fish* es el gasto de pescado a precios constantes, *fishpr* es el precio del pescado, *meatpr* es el precio de la carne y *cons* es el consumo total a precios constantes.

El fichero *fishdem* contiene información acerca de esta serie para el período 1964-1991. Los precios son números índices con base 1986, y *fish* y *cons* son magnitudes a precios constantes también con base en 1986. Los resultados de la estimación del modelo (4-26) son los siguientes:

$$\ln(fish_i) = \underset{(2.30)}{7.788} - \underset{(0.133)}{0.460} \ln(fishpr_i) + \underset{(0.112)}{0.554} \ln(meatpr_i) + \underset{(0.137)}{0.322} \ln(cons_i)$$

Como se puede observar, los signos de las elasticidades son correctos: la elasticidad de la demanda es negativo con respecto al precio del propio bien, mientras que las elasticidades con respecto al precio del bien sustitutivo y con respecto al consumo total son positivos.

En el modelo de (4-26) la restricción de homogeneidad implica la siguiente hipótesis nula:

$$\beta_2 + \beta_3 + \beta_4 = 0 \qquad (4\text{-}27)$$

Para realizar este contraste vamos a utilizar un procedimiento similar al del ejemplo 4.6. Ahora, el parámetro θ se define de la siguiente forma

$$\theta = \beta_2 + \beta_3 + \beta_4 \qquad (4\text{-}28)$$

Haciendo $\beta_2 = \theta - \beta_3 - \beta_4$, se ha estimado el siguiente modelo:

$$\ln(fish) = \beta_1 + \theta \ln(fishpr) + \beta_3 \ln(meatpr / fishpr) + \beta_4 \ln(cons / fishpr) + u \qquad (4\text{-}29)$$

Los resultados obtenidos han sido los siguientes:

$$\ln(\mathit{fish}_i) = \underset{(2.30)}{7.788} - \underset{(0.1334)}{0.4596}\ln(\mathit{fishpr}_i) + \underset{(0.112)}{0.554}\ln(\mathit{meatpr}_i) + \underset{(0.137)}{0.322}\ln(\mathit{cons}_i)$$

Usando (4-28), contrastar la hipótesis nula (4-27) es equivalente a contrastar que el coeficiente de ln(*fishpr*) en (2-29) es igual a 0. Dado que el estadístico *t* para dicho coeficiente es igual a -3.44 y $t_{24}^{0.01/2}$ =2.8, rechazamos la hipótesis de homogeneidad de la demanda de pescado.

4.2.4 Importancia económica versus significación estadística

Hasta ahora hemos hecho hincapié en la significación estadística. Sin embargo, es importante recordar que debemos prestar atención a la magnitud y el signo de los coeficientes estimados, además de a los estadísticos *t.*

La significación estadística de la variable x_j se determina completamente por el tamaño de $t_{\hat{\beta}_j}$, mientras que la importancia económica de una variable se relaciona con el tamaño (y signos) de $\hat{\beta}_j$. Si se pone demasiado énfasis en la significación estadística puede conducirnos a la falsa conclusión de que una variable es "importante" para explicar *y*, aunque su efecto estimado sea modesto.

Así que, incluso si una variable es estadísticamente significativa, es necesario analizar la magnitud del coeficiente estimado para tener una idea de su importancia práctica o económica.

4.3 Contrast de restriccions lineals múltiples utilitzant l'estadístic *F*.

Fins ara, només hem considerat hipòtesi que impliquen una sola restricció. Però amb freqüència, desitgem contrastar hipòtesis múltiples sobre els paràmetres. $\beta_1, \beta_2, \beta_3, ..., \beta_k$.

En las restricciones lineales múltiples distinguiremos tres tipos: las *restricciones de exclusión*, la *significatividad del modelo* y *otras restricciones lineales*.

4.3.1 Restricciones de exclusión

Hipótesis nula y alternativa; modelos no restringido y restringido

Comenzamos contrastando si un conjunto de variables independientes tiene o no un efecto parcial sobre la variable dependiente, *y*. A este contrate se le denomina de *restricciones de exclusión*. Así, considerando el modelo

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u \qquad (4\text{-}30)$$

la hipótesis nula en un típico ejemplo de restricciones de exclusión podría ser la siguiente:

$$H_0 : \beta_4 = \beta_5 = 0$$

Este es un ejemplo de restricciones múltiples, ya que se ha impuesto más de una restricción en los parámetros del modelo. Un contraste de restricciones múltiples es denominado también contraste *conjunto* de hipótesis.

La hipótesis alternativa se puede expresar de la siguiente manera

H_1: H_0 no es cierta

Es importante destacar que se estima H_0 conjuntamente y no individualmente. Ahora, vamos a distinguir entre el modelo *no restringido* (*NR*) y el *restringido* (*R*). El modelo no restringido es el modelo de referencia o modelo inicial. En este ejemplo el modelo no restringido es el modelo que figura en (4-30). El modelo restringido se obtiene mediante la imposición de H_0 en el modelo original. En el anterior ejemplo el modelo restringido es

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Por definición, el modelo restringido siempre tiene un menor número de parámetros que el modelo no restringido. Además, siempre se verifica que

$$SCR_R \geq SCR_{NR}$$

donde SCR_R es la *SCR* del modelo restringido, y SCR_{NR} es la *SCR* del modelo no restringido. Recuerde que, debido a que las estimaciones de los coeficientes por *MCO* se eligen de forma que minimicen la suma de los cuadrados de los residuos, la *SCR* no disminuye (y en general aumenta) cuando algunas restricciones (como la eliminación de variables) se introducen en el modelo.

El aumento en la *SCR* cuando se imponen restricciones puede indicarnos algo sobre si es verosímil la H_0. Si el incremento es grande, esto es una evidencia en contra de H_0, y esta hipótesis será rechazada. Si el incremento es pequeño, esto no será una evidencia en contra de H_0, y esta hipótesis no será rechazada. La pregunta es entonces si el aumento observado en la *SCR* cuando se imponen las restricciones, es suficientemente grande, en relación con la *SCR* en el modelo no restringido, para justificar el rechazo de H_0.

La respuesta depende claro está de α, pero no podemos llevar a cabo el contraste sobre el α seleccionado hasta que tengamos un estadístico cuya distribución sea conocida y esté tabulado bajo la H_0. Por lo tanto, necesitamos una manera de combinar la información de la SCR_R y de la SCR_{NR} para obtener un estadístico de contraste con una distribución conocida bajo H_0.

Ahora, veamos el caso general, donde el *modelo no restringido* es el siguiente:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u \tag{4-31}$$

Supongamos que hay q restricciones de exclusión a contrastar. Entonces, H_0 postula que q variables tienen coeficientes cero. Si se asume que son las últimas q variables, la H_0 se expresa como

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = ... = \beta_k = 0 \tag{4-32}$$

El modelo restringido se obtiene mediante la imposición de q restricciones de la H_0 en el modelo no restringido:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-q} x_{k-q} + u \tag{4-33}$$

La H_1 se expresa como

$$H_1: H_0 \text{ no es cierta} \tag{4-34}$$

Estadístico de contraste: la ratio F

El estadístico *F*, o la *ratio F*, está definido por

$$F = \frac{(SCR_R - SCR_{SR})/q}{SCR_{SR}/(n-k)} \tag{4-35}$$

donde SCR_R es el *SCR* del modelo restringido, SCR_{NR} es el *SCR* del modelo no restringido y *q* es el número de restricciones, es decir, el número de igualdades en la hipótesis nula.

Para poder utilizar el estadístico *F* para el contraste de hipótesis, debemos conocer su distribución muestral bajo H_0 con el fin de elegir el valor de *c* para un α dado, y determinar la regla de rechazo. Se puede demostrar que, bajo la H_0, y asumiendo que los supuestos del *MLC* se mantienen, el estadístico *F* se distribuye como una variable aleatoria *F* de Snedecor con *q* y *n-k grados de libertad*. Escribimos este resultado de la siguiente manera

$$F \mid H_0 \sim F_{q,n-k} \tag{4-36}$$

Una *F* de Snedecor con *q grados de libertad* en el numerador y *n-k grados de libertad* en el denominador es igual a

$$F_{q,n-k} = \frac{\chi_q^2/q}{\chi_{n-k}^2/n-k} \tag{4-37}$$

donde x_q^2 y x_{n-k}^2 son distribuciones *Chi-cuadrado* independientes la una de la otra.

En (4-35) se observa que los *grados de libertad* que corresponden a la SRC_{NR} (gl_{NR}) son *n-k.* Recuerde que

$$\hat{\sigma}_{NR}^2 = \frac{SCR_{NR}}{n-k} \tag{4-38}$$

Por otro lado, los *grados de libertad* que corresponden a la SRC_R (gl_R) son *n-k+q*, porque en el modelo restringido se estiman *k-q* parámetros. Los *grados de libertad* que corresponde a SCR_R-SCR_{SR} son

$$(n\text{-}k\text{+}q)\text{-}(n\text{-}k)=q = \textit{número de grados de libertad}=gl_R\text{-}gl_{NR}$$

Así, en el numerador de la *F*, la diferencia entre las *SCR* se divide por *q*, que es el número de restricciones impuestas al pasar del modelo no restringido al restringido. En el denominador de la *F*, SCR_{NR} es divideix per gl_{NR}. De hecho, el denominador de la *F* es justamente el estimador de σ^2 en el modelo no restringido.

La *ratio F* debe ser mayor que o igual a 0, puesto que $SCR_R - SCR_{NR} \geq 0$.

A menudo es conveniente tener una forma del estadístico *F* tal que pueda ser calculada a partir del R^2 de los modelos restringido y no restringido.

Utilizando el hecho de que $SCR_R = SCT(1-R_R^2)$ y $SCR_{NR} = SCT(1-R_{NR}^2)$, podemos expresar (4-35) del siguiente modo

$$F = \frac{(R_{SR}^2 - R_R^2)/q}{(1-R_{SR}^2)/(n-k)} \tag{4-39}$$

ya que el término *SCT* se cancela.

Esto expresión se denomina la forma *R-cuadrado* del estadístico *F*.

Considerando que, aunque la forma *R-cuadrado* del estadístico *F* es muy conveniente para contrastar restricciones de exclusión, no puede aplicarse para contrastar todo tipo de restricciones lineales. Por ejemplo, la ratio *F* (4-39) no puede utilizarse cuando el modelo no tiene término independiente ni cuando la forma funcional de la variable endógena en el modelo restringido no es la misma que en el modelo no restringido.

Regla de decisión

La distribución $F_{q,n-k}$ está tabulada y disponible en tablas estadísticas, donde se busca el valor crítico ($F^{\alpha}_{q,n-k}$), que depende de α (nivel de significación), q (*gl* del numerador), y *n-k*, (*gl* del denominador). Teniendo en cuenta lo anterior, la regla de decisión es muy simple.

Regla de decisión

Si	$F \geq F^{\alpha}_{q,n-k}$	se rechaza	H_0
Si	$F < F^{\alpha}_{q,n-k}$	no se rechaza	H_0

(4-40)

Por lo tanto, rechazamos la H_0 en favor de la H_1 en α cuando $F \geq F^{\alpha}_{q,n-k}$, como puede verse en la figura 4.15. Es importante destacar que cuando α disminuye, aumenta $F^{\alpha}_{q,n-k}$. Si se rechaza la H_0, entonces decimos que $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ son *estadísticamente significativos conjuntamente*, o, más breve, *significativos conjuntamente*, para el nivel de significación seleccionado.

Este contraste por sí solo no nos permite decir cuáles de las variables tienen un efecto parcial sobre *y*, ya que todas ellas pueden afectar a *y*, o tal vez sólo una afecta a *y*. Si no se rechaza H_0, entonces decimos que no son estadísticamente significativas conjuntamente, o simplemente que no son significativas conjuntamente, lo que a menudo justifica su eliminación del modelo. El estadístico *F* es a menudo útil para contrastar la exclusión de un grupo de variables cuando las variables del grupo están altamente correlacionadas entre sí.

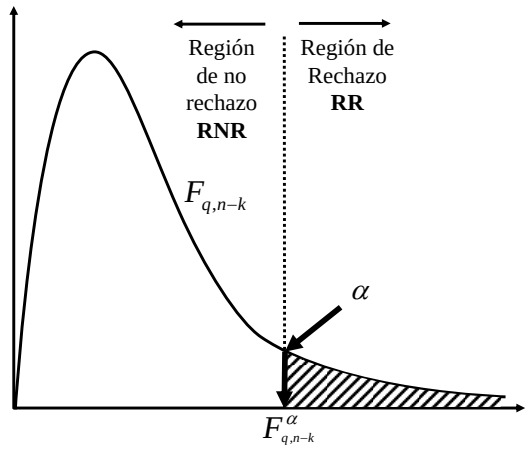


FIGURA 4.15. Región de rechazo y región de no rechazo utilizando la distribución *F*.

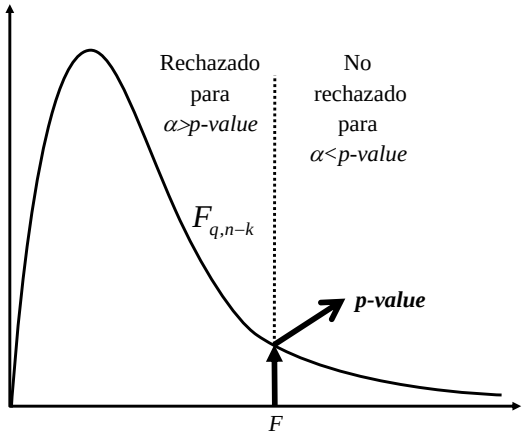


FIGURA 4.16. *Valor-p* utilizando la distribución *F*.

En el contexto del estadístico *F*, el *valor-p* se define como

$$valor\text{-}p = \Pr(F > F' \mid H_0)$$

donde *F* es el valor real del estadístico de contraste y F' designa una variable aleatoria *F* de Snedecor con *q* y *n-k grados de libertad*.

El *valor-p* tiene la misma interpretación que en el estadístico *t*. Un *valor-p* pequeño es una evidencia en contra de la H_0. Por el contrario, un *valor-p* elevado no constituye una evidencia en contra de la H_0. Una vez que *valor-p* ha sido calculado, el contraste *F* puede llevarse a cabo para cualquier nivel de significación. En la figura 4.16 se representa este enfoque alternativo. Como se desprende de la observación de la figura, la determinación del *valor-p* es la operación inversa a la de encontrar el valor en las tablas estadísticas para un determinado nivel de significación. Una vez que el *valor-p* ha sido determinado, se sabe que se rechaza H_0 para cualquier nivel de significación tal que α>*valor-p*; por el contrario, la hipótesis nula no se rechaza cuando α<*valor-p*.

EJEMPLO 4.10 Salarios, experiencia, antigüedad y edad

El siguiente modelo ha sido formulado para analizar los factores determinantes de los salarios:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 tenure + \beta_5 age + u$$

donde *wage* son los salarios mensuales, *educ* son los años de educación, *exper* son los años de experiencia laboral, *tenure* son los años trabajando con la empresa actual, y *age es* la edad en años.

El investigador tiene la intención de excluir *tenure* del modelo, ya que en muchos casos es igual a la experiencia, y también la edad, ya que está altamente correlacionada con la experiencia. ¿Es aceptable la exclusión de ambas variables?

Las hipótesis nula y alternativa son las siguientes:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El modelo restringido correspondiente a esta H_0 es

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + u$$

Utilizando una muestra de 53 observaciones del fichero *wage2*, se han obtenido las siguientes estimaciones para los modelos no restringido y restringido:

$$\ln(wage_i) = 6.476 + 0.0658educ_i + 0.0267exper_i - 0.0094tenure_i - 0.0209age_i \quad SCR = 5.954$$

$$\ln(wage_i) = 6.157 + 0.0457educ_i + 0.0121exper_i \quad SCR = 6.250$$

La ratio *F* obtenida es la siguiente:

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(n-k)} = \frac{(6.250 - 5.954)/2}{5.954/48} = 1.193$$

Teniendo en cuenta que el estadístico *F* es bajo, vamos a ver qué sucede con un nivel de significación del 0.10. En este caso, los grados de libertad para el denominador son 48 (53 observaciones menos 5 parámetros estimados). Si buscamos en las tablas el estadístico *F* para 2 grados de libertad en el numerador y 45 grados de libertad en el denominador, encontramos $F_{2,48}^{0.10} \simeq F_{2,45}^{0.10} = 2.42$. Como $F<2.42$ no se rechaza H_0. Si no se rechaza la H_0 para 0.10, no se rechazará tampoco para 0.05 o 0.01, como se puede en la figura 4.17. Por lo tanto, no podemos rechazar H_0 en favor de H_1. En otras palabras, la antigüedad en la empresa y la edad no son conjuntamente significativas.

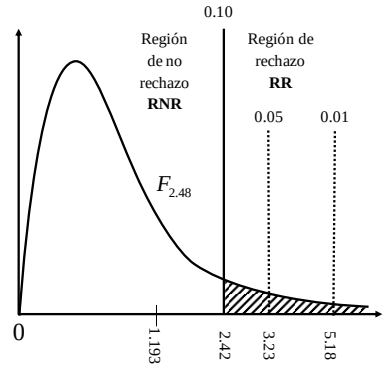


FIGURA 4.17. Ejemplo 4.10: Región de rechazo en la distribución *F* (los valores α son para $F_{2.40}$).

4.3.2 Significación global del modelo

Contrastar la significación del modelo, o significación global del modelo, es un caso particular de los contrastes de restricciones de exclusión. Se podría pensar que este contraste la H_0 debería ser la siguiente:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0 \tag{4-41}$$

Sin embargo, esto no es la H_0 adecuada para contrastar la significatividad global del modelo. Si $\beta_2 = \beta_3 = ... = \beta_k = 0$, entonces el modelo restringido sería el siguiente:

$$y = \beta_1 + u \tag{4-42}$$

Si tomamos esperanzas en (4-42), tenemos que

$$E(y) = \beta_1 \tag{4-43}$$

Así, la H_0 en (4-41) implica que no sólo que las variables explicativas no tienen ninguna influencia sobre la variable endógena, sino también que la media de la variable endógena -por ejemplo, el consumo medio- es igual a 0.

Por lo tanto, si queremos conocer si el modelo es globalmente significativo, la H_0 debe ser la siguiente:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \tag{4-44}$$

El correspondiente modelo restringido dado en (4-42) no explica nada y, por lo tanto, R_R^2 es igual a 0. Entonces, contrastar la H_0 dada en (4-44) es muy fácil utilizando la forma *R-cuadrado* del estadístico *F*:

$$F = \frac{R^2 / k}{(1-R^2)/(n-k)} \tag{4-45}$$

on $R^2 = R_{NR}^2$, ya que sólo es necesario estimar el modelo no restringido, debido a que el R^2 del modelo (4-42) -modelo restringido- es igual a 0.

EJEMPLO 4.11 Salarios de directores ejecutivos

Considere la siguiente ecuación para explicar los salarios (*salary)* de los directores ejecutivos en función de las ventas anuales de la firma (*sales)*, la rentabilidad sobre recursos propios (*roe)* (en forma de porcentaje), y el rendimiento de las acciones de la empresa (*ros*) (en forma de porcentaje):

$$\ln(salary) = \beta_1 + \beta_2 \ln(sales) + \beta_3 roe + \beta_4 ros + u.$$

La pregunta que se plantea es si el rendimiento de la empresa medido por las variables *sales*, *roe* y *ros* es crucial para establecer los salarios de los directores ejecutivos. Para responder a esta pregunta vamos a hacer un contraste de significación global. Las hipótesis nula y alternativa son las siguientes:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

H_1: H_0 no es cierta

El cuadro 4.9 muestra una salida de E-views completa de *mínimos cuadrados* (*ls*), utilizando el fichero *ceosal1*. En la parte inferior puede verse el "estadístico *F*" para el contraste de significación global, así como "Prob.", que es el *valor-p* correspondiente a este estadístico. En este caso *valor-p* es igual a 0, es decir, se rechaza H_0 para todos los niveles de significación (Véase la figura 4.18). Por lo tanto, podemos rechazar que el rendimiento de la empresa no tenga ninguna influencia sobre los salarios de los directores ejecutivos.

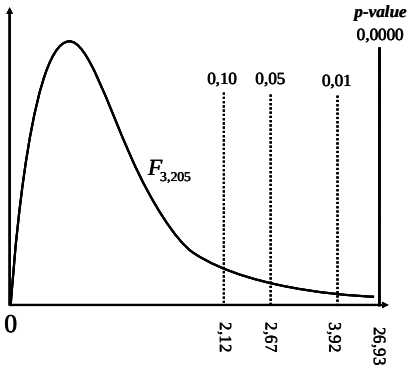


FIGURA 4.18. Ejemplo 4.11: *Valor-p* utilizando la distribución *F* (los valores α para una $F_{3,140}$).

CUADRO 4.9. Salida completa de E-views en el ejemplo 4.11.

Dependent Variable: LOG(SALARY)
Method: Least Squares
Date: 04/12/12 Time: 19:39
Sample: 1 209
Included observations: 209

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.311712	0.315433	13.66919	0.0000
LOG(SALES)	0.280315	0.03532	7.936426	0.0000
ROE	0.017417	0.004092	4.255977	0.0000
ROS	0.000242	0.000542	0.446022	0.6561

R-squared	0.282685	Mean dependent var	6.950386
Adjusted R-squared	0.272188	S.D. dependent var	0.566374
S.E. of regression	0.483185	Akaike info criterion	1.402118
Sum squared resid	47.86082	Schwarz criterion	1.466086
Log likelihood	-142.5213	**F-statistic**	**26.9293**
Durbin-Watson stat	2.033496	**Prob(F-statistic)**	**0.0000**

4.3.3 Estimando otras restricciones lineales

Hasta el momento, hemos contrastado hipótesis con restricciones de exclusión mediante la utilización estadístico *F*. Pero también podemos contrastar hipótesis con restricciones lineales de cualquier tipo. Por lo tanto, podemos combinar varios tipos de restricciones: restricciones de exclusión, restricciones que imponen determinados valores a los parámetros y restricciones relativas a una combinación lineal de los parámetros.

Así, consideremos el siguiente modelo

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

con la hipótesis nula:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_4 = 3 \\ \beta_5 = 0 \end{cases}$$

El modelo restringido correspondiente a esta hipótesis nula es

$$(y - x_2 - 3x_4) = \beta_1 + \beta_3 (x_3 - x_2) + u$$

En el ejemplo 4.12, que se verá a continuación, la hipótesis nula consiste en dos restricciones: una combinación lineal de parámetros y una restricción de exclusión.

EJEMPLO 4.12 Una restricción adicional en la función de producción. (Continuación del ejemplo 4.7)

En la función de producción de Cobb-Douglas, vamos a contrastar la siguiente H_0, que tiene 2 restricciones:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_1 = 0 \end{cases}$$

$$H_1 : H_0 \text{ no es cierta}$$

En la primera restricción se impone que hay rendimientos constantes a escala. En la segunda restricción β_1, el parámetro relacionado con la productividad total de los factores, es igual a 0.

Sustituyendo la restricción de H_0 en el modelo original (modelo sin restricciones), tenemos

$$\ln(output) = (1-\beta_3)\ln(labor) + \beta_3 \ln(capital) + u$$

Operando, se obtiene el modelo restringido:

$$\ln(output / labor) = \beta_3 \ln(capital / labor) + u$$

En la estimación de los modelos no restringido y restringido, obtenemos que SCR_R=3.1101 y SCR_{NR}=0.8516. Por lo tanto, la *ratio F* es igual a

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(n-k)} = \frac{(3.1101 - 0.8516)/2}{0.8516/(27-3)} = 13.551$$

Hay dos razones para no utilizar R^2 en este caso. En primer lugar, el modelo restringido no tiene término independiente. En segundo lugar, el regresando del modelo restringido es diferente del regresando del modelo no restringido.

Dado que el valor de F es relativamente alto, vamos a empezar el contrastar con un nivel de significación del 1%. Para α=0.01, $F_{2,24}^{0.01} = 5.61$. Teniendo en cuenta que F>5.61, rechazamos H_0 en favor de H_1. Por lo tanto, se rechaza la hipótesis conjunta de que hay rendimientos constantes a escala y de que el parámetro β_1 es igual a 0. Si se rechaza H_0 para α=0.01, también será rechazada para los niveles de 5% y 10%.

4.3.4 Relación entre los estadísticos *F* y *t*

Hasta ahora, hemos visto cómo se utiliza el estadístico F para contrastar varias restricciones en el modelo, pero también puede utilizarse para contrastar una sola restricción. En este caso, podemos elegir entre el estadístico F o el estadístico t para hacer un contraste de dos colas. En cualquier caso, las conclusiones serán exactamente las mismas.

Ahora bien, ¿cuál es la relación entre una F con un grado de libertad en el numerador (para contrastar una sola restricción) y una t? Se puede demostrar que

$$t_{n-k}^2 = F_{1,n-k} \tag{4-46}$$

Este hecho se ilustra en la figura 4,19. Se observa que la cola de la F ha desdoblado en las dos colas de la t. Por lo tanto, los dos enfoques conducen exactamente al mismo resultado, siempre que la hipótesis alternativa sea de dos colas. Sin embargo, el estadístico t es más flexible para contrastar una hipótesis con una sola restricción, ya que puede utilizarse para el contrastar la H_0 contra una alternativa de sola cola.

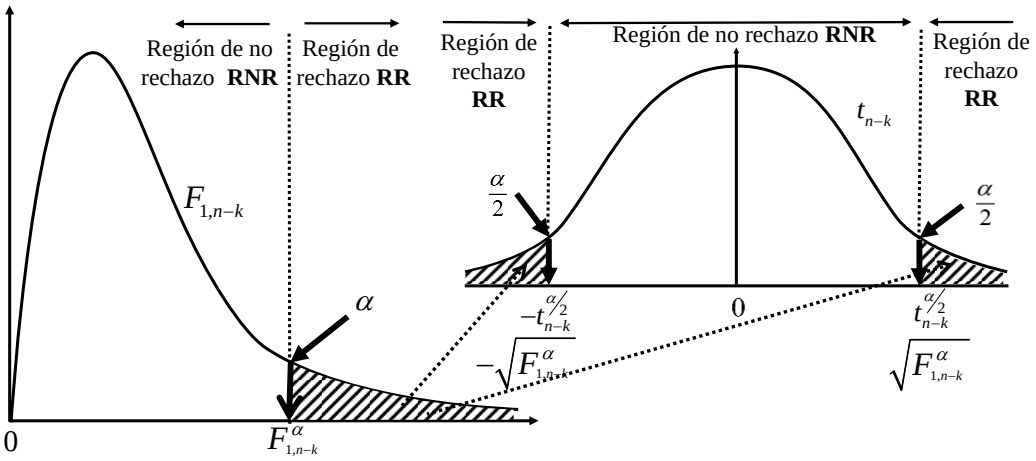


FIGURA 4.19. Relación entre $F_{1,n-k}$ y t_{n-k}.

Además, dado que los estadísticos *t* son también más fáciles de obtener que los estadísticos *F*, en realidad no hay una buena razón para usar un estadístico *F* para contrastar una hipótesis con una sola restricción.

4.4 Contrastes sin normalidad

La normalidad de los estimadores de *MCO* depende crucialmente del supuesto de normalidad de las perturbaciones. ¿Qué sucede si las perturbaciones no tienen una distribución normal? Hemos visto que las perturbaciones bajo los supuestos de Gauss-Markov, y en consecuencia, los estimadores de *MCO* tienen una distribución asintótica normal, es decir, tienen aproximadamente una distribución normal.

Si las perturbaciones no son normales, el estadístico *t* no tiene una distribución *t exacta*, sino sólo *aproximada*. Como puede verse en la tabla de la *t* de Student, para un tamaño muestral de 60 observaciones los puntos críticos son prácticamente igual a los de la distribución normal estándar.

Del mismo modo, si las perturbaciones no son normales, el estadístico *F* no tiene una distribución *F exacta*, sino sólo *aproximada*, cuando el tamaño muestral es lo suficientemente grande y los supuestos de Gauss-Markov se cumplen. Entonces, podemos utilizar el estadístico *F* para contrastar las restricciones lineales en modelos lineales como un contraste aproximado.

Hay otros contrastes asintóticos (la razón de verosimilitud, el multiplicador de Lagrange y el contraste de Wald) basados en las funciones de verosimilitud que pueden utilizarse en contrastar restricciones lineales cuando las perturbaciones no se distribuyen normalmente. Estos tres estadísticos también pueden aplicarse cuando a) las restricciones son no lineales, y b) el modelo es no lineal en los parámetros. Para las restricciones no lineales, en los modelos lineales y no lineales, el contraste más utilizado es el contraste de Wald.

Para contrastar los supuestos del modelo (por ejemplo, los de homoscedasticidad y no autocorrelación) se aplica generalmente el multiplicador de Lagrange (*ML*). En la aplicación del contraste de *ML* se estima a menudo una *regresión auxiliar*. El nombre de regresión auxiliar significa que los coeficientes no son de interés directo: de esta regresión auxiliar sólo se conserva el R^2. En una regresión auxiliar el regresando es, por lo general, los residuos o funciones de los residuos, obtenidos en la estimación por

MCO del modelo original, mientras que las variables independientes son, a menudo, los regresores (y/o funciones de los mismos) del modelo original.

4.5 Predicción

En este epígrafe se examinarán dos tipos de predicción: predicción puntual y predicción por intervalos.

4.5.1 Predicción puntual

La obtención de una predicción puntual no plantea ningún problema especial, puesto que es una operación simple de extrapolación en el contexto de métodos descriptivos.

Designemos por $x_2^0, x_3^0, \cdots, x_k^0$ los valores específicos de cada uno de los *k* regresores en la predicción; estos valores pueden o no corresponder a un punto real de datos en nuestra muestra. Si sustituimos estos valores en el modelo de regresión múltiple, tenemos

$$y^0 = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 + u^0 = \theta^0 + u^0 \tag{4-47}$$

Por tanto, la esperanza, o media, del valor de *y* viene dada por

$$E(y^0) = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \tag{4-48}$$

La predicción puntual se obtiene de forma inmediata mediante la sustitución de los parámetros de (4-48) por los estimadores de *MCO* correspondientes:

$$\hat{\theta}^0 = \hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0 \tag{4-49}$$

Para obtener (4-49) no se necesita postular ningún supuesto estadístico. Pero, si adoptamos los supuestos 1 a 6 del *MLC*, es fácil concluir que $\hat{\theta}^0$ es un predictor insesgado de θ^0

$$E\left[\hat{\theta}^0\right] = E\left[\hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0\right] = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \tag{4-50}$$

Por otra parte, adoptando los supuestos de Gauss Markov (1 a 8), se puede demostrar que este predictor puntual es el estimador lineal insesgado óptimo (ELIO).

Tenemos ahora una predicción puntual para θ^0, pero, ¿cuál es la predicción puntual para y^0? Para responder a esta pregunta tenemos que predecir u_0. Como el error no es observable, el mejor predictor de u_0 es su valor esperado, que es 0. Por lo tanto,

$$\hat{y}^0 = \hat{\theta}^0 \tag{4-51}$$

4.5.2 Predicción por intervalos

Las predicciones puntuales hechas con un modelo econométrico no coincidirán, en general, con los valores observados, debido a la incertidumbre que rodea a los fenómenos económicos.

La primera fuente de incertidumbre es que no podemos utilizar la función de regresión poblacional, ya que no conocemos los parámetros β. En su lugar tenemos que utilizar la función de regresión muestral. El *intervalo de confianza para el valor*

esperado -es decir, para θ^0-, que examinaremos a continuación, incluye sólo este tipo de incertidumbre.

La segunda fuente de incertidumbre es que en un modelo econométrico, además de la parte sistemática, hay una perturbación que no es observable. La *predicción por intervalo para un valor individual* -es decir para y^0-, que se discutirá más adelante incluye tanto la incertidumbre derivada de la estimación, como del término de perturbación.

Una tercera fuente de incertidumbre puede provenir del hecho de no saber exactamente los valores que las variables explicativas tomarán en el momento de predicción. Esta tercera fuente de incertidumbre, que no se aborda aquí, complica los cálculos para la construcción de intervalos.

Intervalo de confianza para el valor esperado

Si queremos predecir el valor esperado de *y*, esto es θ^0, entonces, el error de predicción $\hat{e}_1^0$ será $\hat{e}_1^0 = \theta^0 - \hat{\theta}^0$. De acuerdo con (4-50), el error de predicción esperado es cero. Bajo los supuestos del *MLC*,

$$\frac{\hat{e}_1^0}{ee(\hat{\theta}^0)} = \frac{\theta^0 - \hat{\theta}^0}{ee(\hat{\theta}^0)} \sim t_{n-k}$$

Por lo tanto, podemos escribir que

$$\Pr\left[-t_{n-k}^{\alpha/2} \leq \frac{\theta^0 - \hat{\theta}^0}{ee(\hat{\theta}^0)} \leq t_{n-k}^{\alpha/2} \right] = 1 - \alpha$$

Operando, podemos construir un intervalo de confianza (*IC*) del (1-α)% para θ^0 con la siguiente estructura:

$$\Pr\left[\hat{\theta}^0 - ee(\hat{\theta}^0) \times t_{n-k}^{\alpha/2} \leq \theta^0 \leq \hat{\theta}^0 + ee(\hat{\theta}^0) \times t_{n-k}^{\alpha/2} \right] = 1 - \alpha \qquad (4\text{-}52)$$

Para obtener un *IC* para θ^0, necesitamos conocer el error estándar ($ee(\hat{\theta}_0)$) para $\hat{\theta}^0$. En cualquier caso, hay una manera fácil de estimarlo. Así, resolviendo (4-48) para β_1 vemos que $\beta_1 = \theta^0 - \beta_2 x_2^0 - \beta_3 x_3^0 - ... - \beta_k x_k^0$. Poniendo este resultado en la ecuación (4-47), obtenemos

$$y = \theta^0 + \beta_2 (x_2 - x_2^0) + \beta_3 (x_3 - x_3^0) + \ldots + \beta_k (x_k - x_k^0) + u \qquad (4\text{-}53)$$

Aplicando *MCO* a (4-53), además de la predicción puntual, se obtiene $ee(\hat{\theta}^0)$, que es el error estándar correspondiente al término independiente de esa regresión. El método anterior nos permite obtener un *IC* de *E*(*y*), de la estimación por *MCO*, para distintos valores de los regresores.

Predicción por intervalos para un valor individual

Ahora vamos a construir un intervalo para y^0, al que se denomina *predicción por intervalos para un valor individual*, o de un modo más breve, *predicción por intervalos*. De acuerdo con (4-47), y^0 tiene dos componentes:

$$y^0 = \theta^0 + u^0 \tag{4-54}$$

El *intervalo para el valor esperado* construido previamente es un intervalo de confianza alrededor de θ^0, que es una combinación de los parámetros. En contraste, el intervalo para y^0 es aleatorio, porque uno de sus componentes, u^0, es aleatorio. Por lo tanto, el intervalo para y^0 es un intervalo probabilístico y no un intervalo de confianza. El mecanismo para su obtención es el mismo, pero teniendo en cuenta que ahora vamos a considerar que el conjunto $x_2^0, x_3^0, \cdots, x_k^0$ se encuentra fuera de la muestra utilizada para estimar la regresión.

El error de predicción de $\hat{y}^0$ al predecir y^0 es

$$\hat{e}_2^0 = y^0 - \hat{y}^0 = \theta^0 + u^0 - \hat{y}^0 \tag{4-55}$$

Teniendo en cuenta (4-51) y (4-50), y que $E(u^0)=0$, entonces el error de predicción esperado es cero. En la búsqueda de la varianza de $\hat{e}_2^0$, debe tenerse en cuenta que u^0 no está correlacionado con $\hat{y}^0$, porque $x_2^0, x_3^0, \cdots, x_k^0$ no pertenecen a la muestra.

Por lo tanto, la *varianza del error de predicción* (condicionada a x) es la suma de las varianzas:

$$Var(\hat{e}_2^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2 \tag{4-56}$$

Así pues, hay dos fuentes de variación en $\hat{e}_2^0$:

1. El error de muestreo de $\hat{y}^0$ surge porque se han estimado las β_j.
2. La ignorancia de los factores no observados que afectan a y, los cuales se reflejan en σ^2.

Bajo los supuestos del *MLC*, $\hat{e}_2^0$ también se distribuye normalmente. Utilizando el estimador insesgado de σ^2 y teniendo en cuenta que $var(\hat{y}^0) = var(\hat{\theta}^0)$, podemos definir el error estándar (*ee*) de $\hat{e}_2^0$ como

$$ee(\hat{e}_2^0) = \left\{ \left[ee(\hat{\theta}^0) \right]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}} \tag{4-57}$$

Por lo general, $\hat{\sigma}^2$ es mayor que $\left[ee(\hat{\theta}^0) \right]^2$. Bajo los supuestos del *MLC*,

$$\frac{\hat{e}_2^0}{ee(\hat{e}_2^0)} \sim t_{n-k} \tag{4-58}$$

Por lo tanto, podemos escribir que

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{e}_2^0}{ee(\hat{e}_2^0)} \le t_{n-k}^{\alpha/2} \right] = 1 - \alpha \tag{4-59}$$

Al sustituir $\hat{e}_2^0 = y^0 - \hat{y}^0$ en (4-59) y reordenando se obtiene un intervalo *de predicción* del (1-α)% para y^0:

$$\Pr\left[\hat{y}^0 - ee(\hat{e}_2^0)\times t_{n-k}^{\alpha/2} \le y^0 \le \hat{y}^0 + ee(\hat{e}_2^0)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha \qquad (4\text{-}60)$$

EJEMPLO 4. 13 ¿Cuál es la puntuación esperada en el examen final si se han obtenido 7 puntos en la primera evaluación?

Se ha estimado el siguiente modelo para comparar las puntuaciones en el examen final (*finalmrk*) y en la primera evaluación (*primeval*) de Econometría:

$$finalmrk_i = \underset{(0.715)}{4.155} + \underset{(0.123)}{0.491}\, primeval_i$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

Para estimar la nota final esperada para un estudiante con $primeval^0 = 7$ en la primera evaluación, se estimó el siguiente modelo, de acuerdo con (4-53):

$$finalmrk_i = \underset{(0.497)}{7.593} + \underset{(0.123)}{0.491}\ primeval_i - 7$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

La predicción puntual de partida para $primeval^0 = 7$ es $\hat{\theta}_0 = 7.593$ y los límites inferior y superior de un *IC* del 95%, respectivamente, vienen dados por

$$\underline{\theta}^0 = \hat{\theta}^0 - ee(\hat{\theta}^0)\times t_{14}^{0.05/2} = 7.593 - 0.497\times 2.14 = 6.5$$

$$\overline{\theta}^0 = \hat{\theta}^0 + ee(\hat{\theta}^0)\times t_{14}^{0.05/2} = 7.593 + 0.497\times 2.14 = 8.7$$

Por lo tanto, el estudiante tendrá una confianza del 95% de obtener, en promedio, una calificación final situada entre 6.5 y 8.7.

La predicción puntual puede obtenerse también a partir de la primera ecuación estimada:

$$finalmrk = 4.155 + 0.491\times 7 = 7.593$$

Ahora, vamos a estimar un intervalo de probabilidad del 95% para el valor individual. El error estándar de $\hat{e}_2^0$ es igual a

$$ee(\hat{e}_2^0) = \left\{\left[ee(\hat{y}^0)\right]^2 + \hat{\sigma}^2\right\}^{\frac{1}{2}} = \sqrt{0.497^2 + 1.649^2} = 1.722$$

donde 1.649 es el error estándar de la regresión (E.E.), se ha obtenido de la salida de E-views directamente.

Los límites inferior y superior de un *intervalo de probabilidad* del 95%, respectivamente, vienen dados por

$$\underline{y}^0 = \hat{y}^0 - ee(\hat{e}_2^0)\times t_{14}^{0.025} = 7.593 - 1.722\times 2.14 = 3.7$$

$$\overline{y}^0 = \hat{y}^0 + ee(\hat{e}_2^0)\times t_{14}^{0.025} = 7.593 + 1.722\times 2.14 = 11.3$$

Hay que tener en cuenta que este intervalo de probabilidad es bastante grande debido a que el tamaño de la muestra es muy pequeño.

EJEMPLO 4.14 Prediciendo el salario de los directores ejecutivos

Utilizando los datos de las empresas más importantes de EE.UU. tomadas de Forbes (fichero *ceoforbes*), se ha estimado la siguiente ecuación para explicar los salarios (incluyendo bonificaciones) anuales obtenidos (en miles de dólares) en 1999 por los directores ejecutivos de esas empresas:

$$salary_i = \underset{(104)}{1381} + \underset{(0.0013)}{0.008377}\, assets_i + \underset{(8.671)}{32.508}\, tenure_i + \underset{(0.0538)}{0.2352}\, profits_i$$

$$\hat{\sigma} = 1506 \quad R^2 = 0.2404 \quad n = 447$$

donde *assets* son los activos totales de la firma en millones de dólares, *tenure* es el número de años como director ejecutivo de la compañía, y *profits* son los beneficios en millones de dólares.

En el cuadro 4.10 aparecen las medidas descriptivas de las variables explicativas del modelo de los salarios de los directores generales.

CUADRO 4.10. Medidas descriptivas de las variables del modelo sobre el salario de los ejecutivos.

	assets	*tenure*	*profits*
Media	27054	7.8	700
Mediana	7811	5.0	333
Máximo	668641	60.0	22071
Mínimo	718	0.0	-2669
N. observaciones	447	447	447

Los salarios predichos y los correspondiente $ee(\hat{\theta}_0)$ para valores seleccionados (máximo, media, mediana y mínimo), utilizando un modelo como el (4-53), se presentan en el cuadro 4.11.

CUADRO 4.11. Predicciones para los valores seleccionados.

	Predicción $\hat{\theta}_0$	*E. estándar* $ee(\hat{\theta}_0)$
Valor medio	2026	71
Valor de la mediana	1688	78
Valor del máximo	14124	1110
Valor del mínimo	760	195

4.5.3 Predicción de *y* en un modelo logarítmico

Considere el modelo en logaritmos:

$$\ln(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \tag{4-61}$$

Una vez obtenidas las estimaciones por *MCO*, podemos predecir ln(*y*) de la siguiente forma

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{4-62}$$

Tomando antilogaritmos en (4-62), obtenemos el valor de predicción para *y*:

$$\tilde{y} = \exp(\ln(y)) = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \tag{4-63}$$

Sin embargo, esta predicción es sesgada e inconsistente, ya que *sistemáticamente* subestima el valor esperado de *y*. Veamos el por qué. Si aplicamos antilogaritmos a (4-61), tenemos que

$$y = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(u) \tag{4-64}$$

Antes de tomar las esperanzas en (4-64), hay que tener en cuenta que si $u \sim N(0,\sigma^2)$, entonces $E(\exp(u)) = \exp\left(\frac{\sigma^2}{2}\right)$. Por lo tanto, bajo los supuestos 1 al 9 del *MLC* tenemos que

$$E(y) = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(\sigma^2 / 2) \tag{4-65}$$

Tomando como referencia (4-65) el predictor adecuado de *y* es

$$\hat{y} = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \times \exp(\hat{\sigma}^2 / 2) = \tilde{y} \times \exp(\hat{\sigma}^2 / 2) \tag{4-66}$$

donde $\hat{\sigma}^2$ es el estimador insesgado de σ^2.

Es importante destacar que $\hat{y}$ es un predictor sesgado pero consistente, mientras que $\tilde{y}$ es sesgado e inconsistente.

EJEMPLO 4.15 Prediciendo el salario de los directores ejecutivos con un modelo logarítmico (continuación de 4.14)

Utilizando los mismos datos que en el ejemplo 4.14, se estimó el siguiente modelo:

$$\ln(salary_i) = \underset{(0.210)}{5.5168} + \underset{(0.0232)}{0.1885}\ln(assets_i) + \underset{(0.0032)}{0.0125}tenure_i + \underset{(0.0000195)}{0.00007}\, profits_i$$

$$\hat{\sigma} = 0.5499 \quad R^2 = 0.2608 \quad n = 447$$

En el salario, *salary*, y en los activos, *assets*, se han tomado logaritmos naturales, pero las ganancias, *profits*, están en niveles – es decir, sin ninguna transformación, debido a que algunas observaciones son negativas, por lo que no es posible tomar logaritmos.

En primer lugar, vamos a estimar la predicción de acuerdo con (4-63) para un director ejecutivo que trabaja en una empresa con *assets* =10000, *tenure*=10 años y *profits*=1000:

$$salary_i = \exp(\ln(salary_i))$$
$$= \exp(5.5168 + 0.1885\ln(10000) + 0.0125\times 10 + 0.00007\times 1000) = 1716$$

Ahora, utilizando (4-66), se obtiene la siguiente predicción que es consistente:

$$salary = \exp(0.5499^2/2)\times 1716 = 1996$$

4.5.4 Evaluación de las predicciones y predicción dinámica

En esta sección vamos a comparar las predicciones realizadas con un modelo econométrico y los valores realmente observados a fin de evaluar la capacidad predictiva del modelo. También examinaremos la predicción dinámica en modelos en los que hay variables endógenas retardadas incluidas como regresores.

Estadísticos para evaluación de predicciones

Supongamos que se realizan predicciones para $i=n+1, n+2,\ldots, n+h$, y que se designa el valor real y el previsto en el período i como y_i e $\hat{y}_i$, respectivamente. Ahora, vamos a presentar algunos de los estadísticos más usuales utilizados para la evaluación de las predicciones.

Error absoluto medio (EAM)

El *EAM* se define como la media de los valores absolutos de los errores:

$$EAM = \frac{\sum_{i=n+1}^{n+h} |\hat{y}_i - y_i|}{h} \tag{4-67}$$

Al tomar valores absolutos de los errores se evita que los errores positivos se compensen con los negativos.

Error absoluto medio en porcentaje (EAMP),

$$EAMP = \frac{\sum_{i=n+1}^{n+h} \frac{|\hat{y}_i - y_i|}{y_i}}{h} \times 100 \tag{4-68}$$

Raíz del error cuadrático medio (RECM)

Este estadístico se define como la raíz cuadrada del error cuadrático medio:

$$RECM = \sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i - y_i^{\ 2}}{h}} \tag{4-69}$$

Como se toman los cuadrados de los errores, se evita la compensación entre errores positivos y negativos. Es importante remarcar que la *RECM* impone una penalización mayor a los errores de predicción que el *EAM*.

Coeficiente de desigualdad de Theil (U)

Este coeficiente se define como sigue:

$$U = \frac{\sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i - y_i^{\ 2}}{h}}}{\sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i^2}{h}} + \sqrt{\frac{\sum_{i=n+1}^{n+h} y_i^2}{h}}} \tag{4-70}$$

Cuanto más pequeño es *U* más precisas son las predicciones. La escala de *U* es tal que siempre está entre 0 y 1. Si U=0, $y_i = \hat{y}_i$ para todas las predicciones; si U=1, la predicción es tan mala como pueda ser. El estadístico *U* de Theil puede ser reescalado y descomponerse en 3 proporciones: el sesgo, la varianza y la covarianza. Por supuesto, la suma de estas tres proporciones es 1. La interpretación de estas tres proporciones es la siguiente:

1) El *sesgo* refleja errores sistemáticos. Cualquiera que sea el valor de *U*, esperaríamos que el sesgo esté cercano a 0. Un sesgo grande sugiere que las predicciones están sistemáticamente por encima, o por debajo, de los valores reales.
2) La *varianza* refleja también errores sistemáticos. El tamaño de esta proporción es una indicación de la incapacidad de las predicciones para replicar la variabilidad de la variable a predecir.
3) La *covarianza* mide errores de carácter no sistemático. Idealmente, ésta debería ser la mayor proporción de la desigualdad de Theil.

Además del coeficiente definido en (4-70), Theil propuso otros coeficientes para la evaluación de predicciones.

Predicción dinámica

Sea el siguiente modelo:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t \tag{4-71}$$

Supongamos que se desean realizar predicciones para los periodos $i=n+1,\ldots,i=n+h$, y que se designa el valor real y previsto en el periodo *i* como y_i y $\hat{y}_i$ respectivamente. La predicción para el periodo $n+1$ es

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + \hat{\beta}_3 y_n \tag{4.72}$$

Como se puede observar, para la predicción se utiliza el valor observado de *y* (y_n) porque está dentro de la muestra utilizada en la estimación. Para la predicción del resto de los períodos utilizamos de forma recursiva la predicción del valor retardado de la variable dependiente (predicción dinámica), es decir,

$$\hat{y}_{n+i} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+i} + \hat{\beta}_3 \hat{y}_{n-1+i} \qquad i = 2,3,\cdots,h \tag{4-73}$$

Así, desde el periodo n+2 al n+h la predicción realizada para un periodo se utiliza para pronosticar la variable endógena en el periodo siguiente.

Ejercicios

Ejercicio 4.1 Para explicar el precio de la vivienda en una ciudad americana se formula el siguiente modelo:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

donde *rooms* es el número de habitaciones de la casa, *lowstat* es el porcentaje de personas de "clase marginal" en la zona y *crime* es el número de delitos per cápita en la zona. Los precios de las casas están medidos en dólares.

Utilizando los datos del fichero *hprice2*, se ha estimado el modelo anterior:

$$\underset{}{price} = \underset{(8022)}{-15694} + \underset{(1211)}{6788}\, rooms - \underset{(81)}{268}\, lowstat - \underset{(960)}{3854}\, crime$$

$$R^2 = 0.771 \qquad n = 55$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Interprete el significado de los coeficientes $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\beta}_4$.

b) ¿Tiene el porcentaje de personas de "clase marginal" una influencia negativa sobre el precio de las casas en esa área?

c) ¿Tiene el número de habitaciones una influencia positiva sobre el precio de la vivienda?

Ejercicio 4.2 Considere el siguiente modelo:

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + \beta_4 punder5 + u$$

donde *fruit* es el gasto en frutas, *inc* es la renta disponible del hogar, *hhsize* es el número de miembros del hogar y *punder5* es la proporción de niños menores de cinco años en el hogar.

Usando los datos del fichero *demand*, se ha estimado el modelo anterior:

$$\ln(fruit) = \underset{(3.701)}{-9.768} + \underset{(0.512)}{2.005}\ln(inc) - \underset{(0.179)}{1.205}\, hhsize - \underset{(1.302)}{1.795}\, punder5$$

$$R^2 = 0.728 \qquad n = 40$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Interprete el significado de los coeficientes $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\beta}_4$.

b) ¿Tiene el número de miembros del hogar un efecto estadísticamente significativo sobre el gasto en fruta?

c) ¿Es la proporción de niños menores de cinco años en el hogar un factor que tiene influencia negativa en el gasto fruta?

d) ¿Es la fruta un bien de lujo?

Ejercicio 4.3 (Continuación del ejercicio 2.5). Teniendo en cuenta el modelo

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad \text{i} = 1,2,\ldots,n$$

se han obtenido los siguientes resultados con un tamaño muestral de 11 observaciones:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

(Recuerde que $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$)

a) Construya un estadístico para contrastar $H_0 : \beta_2 = 0$ contra $H_1 : \beta_2 \neq 0$

b) Contraste las hipótesis de la cuestión *a)* cuando $EB = 2F^2$.

c) Contraste las hipótesis de la cuestión *a)* cuando $EB = F^2$.

Ejercicio 4.4 Se ha formulado el siguiente modelo para explicar el gasto de alimentos (*alim*):

$$alim = \beta_1 + \beta_2 renta + \beta_3 pralim + u$$

donde *renta* es la renta disponible y *pralim* es el índice de precios relativos de los alimentos con respecto a los demás productos de consumo.

Tomando una muestra de observaciones correspondientes a 20 años sucesivos se obtienen los siguientes resultados:

$$\underset{}{alim_i} = \underset{(4.92)}{1.40} + \underset{(0.01)}{0.126}\, renta_i - \underset{(0.07)}{0.036}\, pralim_i$$

$$R^2 = 0.996; \qquad \sum \hat{u}_t^2 = 0.196$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la hipótesis nula de que el coeficiente de *pralim* es menor que 0.

b) Obtenga un intervalo de confianza del 95% para la propensión marginal al consumo de alimentos respecto a la renta.

c) Contraste la significatividad conjunta del modelo.

Ejercicio 4.5 Se han utilizado mínimos cuadrados ordinarios para estimar la siguiente función de demanda de alquiler de viviendas:

$$\ln(galq_i) = \beta_1 + \beta_2 \ln(palq_i) + \beta_3 \ln(renta_i) + \varepsilon_i$$

donde *galq* es el gasto en alquiler de viviendas, *palq* es el precio de alquiler, y *renta* es la renta disponible.

Utilizando una muestra de 403 observaciones, se obtienen los siguientes resultados:

$$\ln(galq_i) = 10 - 0.7 \ln(palq_i) + 0.9 \ln(renta_i)$$

con $R^2 = 0.39$ y la matriz estimada de covarianzas

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 0.09 & 0.085 \\ 0 & 0.085 & 0.09 \end{bmatrix}$$

a) Interprete los coeficientes de ln(*galq*) y ln(*palq*).
b) Utilizando un nivel de significación del 0.01, contraste la hipótesis nula de que $\beta_2=\beta_3=0$.
c) Contraste la hipótesis nula de que $\beta_2=0$, frente a la alternativa de que $\beta_2<0$.
d) Contraste la hipótesis nula de que $\beta_3=1$ frente a la alternativa de que $\beta_3 \neq 1$.
e) Contraste la hipótesis nula de que un aumento en el precio de la vivienda y un aumento en la renta, en la misma proporción, no tienen ningún efecto sobre la demanda de viviendas.

Ejercicio 4.6 Utilizando una muestra de 30 empresas se han estimado los siguientes modelos correspondientes a las funciones del coste medio (*cm*):

$$\begin{gathered} cm_i = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, cant_i \\ R^2 = 0.838 \qquad SCR = 8090 \end{gathered} \tag{1}$$

$$\begin{gathered} cm_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, cant_i + \underset{(11.61)}{26.73}\, cant_i^2 - \underset{(1.22)}{1.40}\, cant_i^3 \\ R^2 = 0.978 \quad SCR = 1097 \end{gathered} \tag{2}$$

donde *cm* es el coste medio y *cant* es la cantidad producida.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste si los términos cuadrático y cúbico de la cantidad producida son significativos en la determinación el coste medio.
b) Contraste la significación global del modelo 2.

Ejercicio 4.7 Utilizando una muestra de 35 observaciones, se han estimado los siguientes modelos para explicar el gasto en café

$$\begin{gathered} \ln(coffee) = 21.32 + \underset{(0.01)}{0.11}\ln(inc) - \underset{(0.23)}{1.33}\ln(cprice) + 1.35\ln(tprice) \\ R^2 = 0.905 \qquad SCR = 254 \end{gathered} \tag{1}$$

$$\begin{gathered} \ln(coffee) = 19.9 + \underset{(0.02)}{0.14}\ln(inc) - \underset{(0.21)}{1.42}\ln(cprice) \\ SCR = 529 \end{gathered} \tag{2}$$

donde *inc* es la renta disponible, *cprice* es el precio del café y *tprice* es el precio del té.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la significatividad global del modelo (1)
b) El error estándar de ln(*tprice*) no aparece en el modelo (1), ¿lo puede calcular?
c) Contraste si el precio del té es estadísticamente significativo.
d) ¿Cómo se contrastaría la hipótesis de que la elasticidad del precio del café es igual pero con signo opuesto al de la elasticidad del precio del té? Detalle el procedimiento.

Ejercicio 4.8 Ha sido formulado el siguiente modelo para analizar los determinantes de la calidad del aire (*airqual*) en 30 áreas SMSA (Standard Metropolitan Statistical Areas) de California:

$$airqual = \beta_1 + \beta_2\, popln + \beta_3 medincm + \beta_4\, poverty + \beta_5\, fueoil + \beta_6 valadd + u$$

donde *airqual* es el peso en $\mu g/m^3$ de partículas en suspensión, *popln* es la población en miles, *medincm* es la renta media per cápita en dólares, *poverty* es el porcentaje de familias con una renta inferior al nivel de pobreza, *fueloil* son miles de barriles de petróleo consumido en industrias manufactureras, y *valadd* es el valor añadido de las industrias manufactureras en el año 1972 en miles de dólares.

Utilizando los datos del fichero *airqualy*, se ha estimado el modelo anterior:

$$\underset{}{airqual_i} = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$
$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i$$
$$R^2 = 0.415 \qquad n = 30$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Interprete los coeficientes de *medincm*, *poverty* y *valadd*

b) ¿Son los coeficientes de pendiente significativos al 10% de forma individual?

c) Contraste de significación conjunta de *fueloil* y *valadd*, sabiendo que

$$\underset{}{airqual_i} = \underset{(10.41)}{97.67} + \underset{(0.020)}{0.0566}\, popln_i - \underset{(0.0039)}{0.0102}\, medincm_i - \underset{(0.0078)}{0.0174}\, poverty_i$$
$$R^2 = 0.339 \qquad n = 30$$

d) Si se omite en el primer modelo la variable de la pobreza se obtienen los siguientes resultados:

$$\underset{}{airqual_i} = \underset{(10.02)}{82.98_i} + \underset{(0.031)}{0.0523}\, popln_i - \underset{(0.0055)}{0.0097}\, medincm_i$$
$$- \underset{(0.0017)}{0.00063}\, fueoil_i - \underset{(0.0028)}{0.00037} valadd_i$$
$$R^2 = 0.218 \quad n = 30$$

¿Son los coeficientes de pendiente significativos al 10% de forma individual en el nuevo modelo? ¿Considera que estos resultados son razonables en comparación con los obtenidos en el apartado *b*)?

Comparando R^2 en los dos modelos estimados, ¿cuál es el papel desempeñado por la pobreza en la determinación de la calidad del aire?

e) Si se hace una regresión para explicar *airqual* usando como regresores únicamente el término independiente y *poverty*, se obtiene que R^2=0.037. ¿Considera razonable este valor teniendo en cuenta los resultados obtenidos en el apartado *d*)?

Ejercicio 4.9 Se han estimado por el método de *MCO* con una muestra de 39 observaciones las siguientes funciones de producción:

$$output_t = \hat{\alpha} labor_t^{1.30} capital_t^{0.32} \exp(0.0055 trend_t) \qquad R^2 = 0.9945$$

$$output_t = \hat{\beta} labor_t^{1.41} capital_t^{0.47} \qquad R^2 = 0.9937$$

$$output_i = \hat{\gamma} \exp(0.0055 trend_t) \qquad R^2 = 0.9549$$

donde *trend* es una variable de tendencia.

a) Contraste la significatividad conjunta de *labor* y *capital*.
b) Contraste la significatividad del coeficiente de la variable *trend*.
c) Indique los supuestos estadísticos bajo las cuales los contrastes realizados en los dos apartados anteriores son correctos. Una cuestión adicional: Escriba el modelo poblacional correspondiente a la primera de las tres especificaciones anteriores.

Ejercicio 4.10 En una investigación se ha formulado el siguiente modelo:

$$y = \beta_1 + \beta_2\ x_2 + \beta_3 x_3 + u$$

Con una muestra de 43 observaciones se han obtenido los siguientes resultados:

$$\hat{y}_i = \text{-}0.06 + 1.44\ x_{2i} - 0.48\ x_{3i}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1011 & -0.0007 & -0.0005 \\ & 0.0231 & -0.0162 \\ & & 0.0122 \end{bmatrix} \qquad \sum y_i^2 = 444 \qquad \sum \hat{y}_i^2 = 424.92$$

a) Contraste que el término independiente es menor que 0.
b) Contraste que β_2=2.
c) Contraste de la hipótesis nula $\beta_2+3\beta_3=0$.

Ejercicio 4.11 Dada la función de producción

$$q\ = ak^{\alpha} l^{\beta} \exp(u)$$

se ha procedido a su estimación con datos de la economía española de los últimos 20 años, obteniéndose los siguientes resultados:

$$\ln(q_i) = 0.15 + 0.73\ln(k_i) + 0.47\ln(l_i)$$

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4129 & -95 & -266 \\ -95 & 3 & 5 \\ -266 & 5 & 19 \end{bmatrix} \qquad SCR = 0.017$$

a) Contraste la significatividad individual de los coeficientes de *k* y *l.*
b) Contraste si el parámetro α es significativamente distinto de 1.
c) Contraste si existen rendimientos crecientes a escala.

Ejercicio 4.12 Sea el siguiente modelo de regresión múltiple:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$$

Con una muestra de 33 observaciones se ha estimado este modelo por *MCO*, obteniéndose los siguientes resultados:

$$\hat{y}_i = 12.7 + 14.2x_{1i} + 2.1x_{2i}$$

$$\hat{\sigma}^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4.1 & -0.95 & -0.266 \\ -0.95 & 3.8 & 0.5 \\ -0.266 & 0.5 & 1.9 \end{bmatrix}$$

a) Contraste la hipótesis nula $\alpha_0 = \alpha_1$.
b) Contraste si $\alpha_1 / \alpha_2 = 7$.

c) ¿Son significativos individualmente los coeficientes α_0, α_1, y α_2?

Ejercicio 4.13 Con una muestra de 30 empresas se han estimado las siguientes funciones de coste:

$$a)\ cost_i = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, x_i \qquad R^2 = 0.838 \quad \bar{R}^2 = 0.829 \quad SCR = 8090$$

$$b)\ cost_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, x_i + \underset{(11.61)}{26.73}\, x_i^2 - \underset{(1.22)}{1.40}\, x_i^3 \quad R^2 = 0.978 \quad \bar{R}^2 = 0.974 \quad SCR = 1097$$

donde *cost* es el coste medio y *x* es la cantidad producida.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Cuál de los dos modelos elegiría? ¿En base a qué criterio?

b) Contraste si los términos cuadrático y cúbico de la cantidad producida influyen significativamente en la determinación del coste medio.

c) Contraste la significatividad del modelo *b*).

Ejercicio 4.14 Un investigador formula el siguiente modelo:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Utilizando una muestra de 13 observaciones obtiene los siguientes resultados:

$$\begin{gathered} \hat{y}_i = 1.00 - 1.82 x_{2i} + 0.36 x_{3i} \\ R^2 = 0.50 \quad n = 13 \\ \operatorname{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.25 & -0.01 & 0.04 \\ -0.01 & 0.16 & -0.15 \\ 0.04 & -0.15 & 0.81 \end{bmatrix} \end{gathered} \qquad (1)$$

Utilizando la información disponible:

a) Contraste la hipótesis nula de que $\beta_2 = 0$ frente a la hipótesis alternativa de que $\beta_2 < 0$.

b) Contraste la hipótesis nula de que $\beta_2 + \beta_3 = -1$ frente a la hipótesis alternativa de que $\beta_2 + \beta_3 \neq -1$, con un nivel de significación del 5%.

c) ¿El modelo en su conjunto es significativo?

d) Suponiendo que las variables en (1) están medidas en logaritmos naturales, ¿cuál es la interpretación del coeficiente correspondiente a x_3?

Ejercicio 4.15 Con una muestra de 50 empresas del sector del automóvil se han estimado las siguientes funciones de producción, utilizando como variable endógena el valor añadido bruto de la producción de automóviles (*vab*) y como variables explicativas el factor trabajo (*labor*) y el factor capital (*capital*).

1) $$\begin{gathered} \ln(vab_i) = 3.87 + \underset{(0.11)}{0.80} \ln(labor_i) + \underset{(0.24)}{1.24} \ln(capital_i) \\ SCR = 254 \quad R^2 = 0.75 \quad \bar{R}^2 = 0.72 \end{gathered}$$

2) $$\begin{gathered} \ln(vab_i) = 19.9 + 1.04 \ln(capital_i) \\ SCR = 529 \quad R^2 = 0.84, \bar{R}^2 = 0.81 \end{gathered}$$

3) $$\begin{gathered} \ln(vab / labor_i) = 15.2 + 0.87 \ln(capital_i / labor_i) \\ SCR = 380 \end{gathered}$$

(Entre paréntesis aparecen los errores estándar de los estimadores).

a) Contraste la significatividad conjunta de los dos factores de la función de producción.
b) Contraste si el factor trabajo tiene una influencia significativamente positiva sobre el valor añadido de la producción de automóviles.
c) Contraste la hipótesis de rendimientos constantes a escala. Razone la respuesta.

Ejercicio 4.16 Con una muestra de 35 observaciones anuales se han estimado dos funciones de demanda de vino de Rioja, utilizando como variable endógena el gasto en vino de reserva (*vino*) y como variables explicativas la renta disponible (*renta*), el precio medio de una botella de vino de Rioja de reserva (*pvinrioj*) y el precio medio de una botella de vino de Ribera de Duero de reserva (*pvinduer)*. Los resultados son los siguientes:

$$\ln(vino_i) = 21.32 + \underset{(0.01)}{0.11}\ln(renta_i) - \underset{(0.23)}{1.33}\ln(pvinrioj_i) + \underset{(0.233)}{1.35}\ln(pvinduer_i)$$

$$R^2 = 0.905 \quad SCR = 254$$

$$\ln(vino_i) = 19.9 + \underset{(0.02)}{0.14}\ln(renta_i) - \underset{(0.21)}{1.42}\ln(pvinrioj_i)$$

$$SCR = 529$$

(Los números entre paréntesis son los errores estándar de los estimadores.)
a) Contraste la significatividad global del primer modelo.
b) Contraste si el precio del vino de Ribera de Duero tiene una influencia significativa, aplicando dos estadísticos que no utilicen la misma información. Muestre que ambos procedimientos son equivalentes
c) ¿Cómo contrastaría la hipótesis de que la elasticidad precio del vino de Rioja es igual pero con signo contrario a la elasticidad precio del vino de Ribera de Duero? Detalle el procedimiento que seguiría.

Ejercicio 4.17 Para analizar la demanda de té de Ceylán (*teceil*) se ha formulado el siguiente modelo econométrico:

$$\ln(teceil) = \beta_1 + \beta_2\ln(renta) + \beta_3\ln(pteceil) + \beta_4\ln(pteind) + \beta_5\ln(pcobras) + u$$

donde *renta* es la renta disponible, *pteceil* el precio del té de Ceilán, *pteind* es el precio del té de la India y *pcobras* es el precio del café de Brasil.

Con una muestra de 22 observaciones se han realizado las siguientes estimaciones:

$$\ln(teceil_i) = 2.83 + \underset{(0.17)}{0.25}\ln(renta_i) - \underset{(0.98)}{1.48}\ln(pteceil_i) + \underset{(0.69)}{1.18}\ln(pteind_i) + \underset{(0.15)}{0.20}\ln(pcofbras_i)$$

$$SCR=0.4277$$

$$\ln(teceil_i \times pteceil) = 0.74 + \underset{(0.16)}{0.26}\ln(renta_i) + \underset{(0.15)}{0.20}\ln(pcofbras_i)$$

$$SCR=0.6788$$

(Los números entre paréntesis son los errores estándar de los estimadores.)
a) Contraste la significatividad de la renta disponible.
b) Contraste la hipótesis de que $\beta_3 = -1$ y $\beta_4 = 0$, explicando el procedimiento aplicado.

c) Si en lugar de disponer de información sobre la *SCR*, solo conociera el R^2 de cada modelo, ¿cómo procedería para realizar el contraste de la parte *b)*?

Ejercicio 4.18 Con una muestra de 64 países se han obtenido las siguientes estimaciones para explicar las defunciones de menores de 5 años por 1000 habitantes nacidos vivos (*defmen5*)

1) $defmen5_i = 263.64 - \underset{(0.0019)}{0.0056}\, renpc_i + \underset{(0.21)}{2.23}\, tanalmuj_i ; \qquad R^2 = 0.7077$

2) $defmen5_i = 168.31 - \underset{(0.0018)}{0.0055}\, renpc_i + \underset{(0.25)}{1.76}\, tanalmuj_i + 12.87 tfec_i , \quad R^2 = 0.7474$

donde *renpc* es la renta per cápita, *tanalmuj* es la tasa de analfabetismo de las mujeres y *tfec* es la tasa de fecundidad (*tf*$_i$).

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la significatividad conjunta de renta, tasa de analfabetismo y tasa de fecundidad.
b) Contraste la significatividad de la tasa de fecundidad.
c) ¿Cuál de los dos modelos elegiría? Razone la respuesta.

Ejercicio 4.19 Se ha estimado la siguiente función de ventas de automóviles de una determinada marca utilizando una muestra de 32 observaciones anuales:

$$\hat{v}_i = \underset{(6.48)}{104.8} - \underset{(3.19)}{6.64}\, p_i + \underset{(0.16)}{2.98}\, g_i$$

$$\sum \hat{u}_i^2 = 1805.2; \qquad \sum (v_i - \bar{v})^2 = 13581.4$$

donde *v* son ventas, *p* es el precio de los automóviles y *g* son gastos en publicidad.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Son significativos conjuntamente el precio y los gastos en publicidad? Razone la respuesta.
b) ¿Es admisible la hipótesis de que los precios tengan una influencia negativa sobre las ventas? Razone la respuesta.
c) Describa detalladamente como contrastaría la hipótesis de que el impacto de los gastos en publicidad sobre las ventas es mayor que menos 0.4 veces el impacto del precio.

Ejercicio 4.20 En un estudio sobre los costes de producción (*cp*$_i$) de 62 minas de carbón se elabora el siguiente modelo:

$$cp_i = \underset{(3.4)}{2.20} - \underset{(0.005)}{0.104}\, gm_i + \underset{(2.2)}{3.48}\, dg_i + \underset{(0.15)}{0.104}\, pa_i$$

donde *gm*$_i$ es el grado de mecanización, *dg*$_i$ es una medida de dificultades geológicas y *pa*$_i$ es el porcentaje de absentismo.

(Los números entre paréntesis son los errores estándar de los estimadores.)

Se dispone además de la siguiente información:

$$\sum \left[cp_i - \overline{cp} \right]^2 = 109.6 \qquad \sum \hat{u}_i^2 = 18.48$$

a) Contraste la significatividad de cada uno de los coeficientes del modelo.
b) Contraste la significatividad global del modelo.

Ejercicio 4.21 Con quince observaciones se ha obtenido la siguiente estimación:

$$\hat{y}_i = 8.04 - \underset{(1.00)}{2.46}\, x_{i2} + \underset{(0.60)}{0.23}\, x_{i3}$$

$$\bar{R}^2 = 0.30$$

donde los valores entre paréntesis son los errores estándar de los estimadores y el coeficiente de determinación es el corregido.

a) ¿Es significativo el coeficiente de la variable x_3?
b) ¿Se rechaza la hipótesis de que un aumento en x_2 de dos unidades ocasiona una disminución en y de ocho unidades?
c) Analice la significatividad conjunta del modelo.

Ejercicio 4.22 Considérese la siguiente especificación econométrica:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

Con una muestra de 26 observaciones se han obtenido las dos siguientes estimaciones:

1) $\hat{y}_i = 2 + \underset{(1.9)}{3.5}\, x_{1i} - \underset{(2.2)}{0.7}\, x_{2i} - \underset{(1.5)}{2}\, x_{3i} + u_i$ $\qquad R^2$=0.982

2) $\hat{y}_i = 1.5 + \underset{(2.7)}{3}\,(x_{1i} + x_{2i}) - \underset{(2.4)}{0.6}\, x_{3i} + u_i$ $\qquad R^2$= 0.876

(Entre paréntesis figuran los estadísticos t)

a) Demostrar que las siguientes expresiones del estadístico F son equivalentes:

$$F = \frac{(SCR_R - SCR_{NR})/r}{SCR_{NR}/(n-k)} \qquad\qquad F = \frac{\left(R^2_{NR} - R^2_R\right)/q}{(1-R^2_{NR})/(n-k)}$$

b) Contraste la hipótesis nula $\beta_2 = \beta_3$.

Ejercicio 4.23 En la estimación del modelo de Brown, realizada en el ejercicio 3.19 utilizando el fichero *consumsp*, se obtuvo el siguiente resultado:

$$conspc_t = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

R^2=0.997; $\qquad SCR$=1891320; $\qquad n$=56

También se realiza la estimación de los siguientes modelos:

$$conspc_t - conspc_{t-1} = \underset{(84.43)}{-98.13} + \underset{(0.0803)}{0.2757}(incpc_t - conspc_{t-1})$$

R^2=0.1792; $\qquad SCR$=2199474; $\qquad n$=56

$$conspc_t - incpc_{t-1} = \underset{(84.88)}{-7.156} - \underset{(0.0090)}{0.0264}\, incpc + \underset{(0.0903)}{0.5771}(conspc_{t-1} - incpc_t)$$

R^2=0.6570; $\qquad SCR$=1891320; $\qquad n$=56

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la significatividad de cada uno de los coeficientes del primer modelo.
b) Contraste que el coeficiente de *incpc* en el primer modelo es menor que 0.5.
c) Contraste la significatividad global del primer modelo.
d) ¿Es admisible la hipótesis de que $\beta_2 + \beta_3 = 1$?

e) Demuestre que operando en el tercer modelo puede llegar a los mismos coeficientes que en el primer modelo.

Ejercicio 4.24 El siguiente modelo fue formulado para analizar los determinantes del salario medio en dólares que se obtiene al graduarse en la clase del 2010 en las mejores escuelas de negocios de EE.UU. (*salMBAgr*):

$$salMBAgr = \beta_1 + \beta_2 tuition + \beta_3 salMBApr + u$$

donde *tuition* son los derechos de matrícula, incluyendo además todos los demás honorarios para el programa completo, pero con exclusión de los gastos de manutención, y *salMBApr* es el salario medio anual en dólares obtenido previamente por la clase de 2010.

Utilizando los datos del fichero *MBAtui10*, se ha estimado el modelo anterior:

$$\underset{}{salMBAgr_i} = \underset{(5415)}{42489} + \underset{(0.0628)}{0.1881} tuition_i + \underset{(0.1015)}{0.5992} salMBApr_i$$

$$R^2=0.703 \quad n=39$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Cuáles de los regresores incluidos en el modelo anterior son individualmente significativos al 1% y el 5%?

b) Contraste la significación global del modelo.

c) ¿Cuál es el valor predicho de *salMBAgr* para un estudiante de posgrado que pagó 100000 dólares de derechos de matrícula en un máster MBA de dos años y que anteriormente tuvo un salario de 70000 dólares anuales? ¿Cuántos años de trabajo necesita el estudiante para compensar los gastos de matrícula? Para responder a esta pregunta suponga que la tasa de descuento es igual a la tasa esperada de aumento salarial y que el estudiante no recibió ningún ingreso salarial durante los 2 cursos de duración del máster.

d) Si añadimos el regresor *rank2010* (el rango de cada escuela de negocios en 2010), se obtienen los siguientes resultados:

$$\underset{}{salMBAgr_i} = \underset{(8520)}{61320} + \underset{(0.0626)}{0.1229} tuition_i + \underset{(0.1055)}{0.4662} salMBApr_i$$
$$\underset{(85.13)}{-232.06} rank2010_i$$

$$R^2=0.755 \quad n=39$$

¿Cuál de los regresores incluidos en este modelo son individualmente significativos al 5%?

¿Cuál es la interpretación del coeficiente de *rank2010*?

e) La variable *rank2010* se ha construido a partir de tres componentes: *gradpoll* es una clasificación basada en encuestas a los graduados en MBA y contribuye con un 45 por ciento a la clasificación final; *corppoll* es una clasificación basada en encuestas realizadas a reclutadores de los MBA y contribuye con el 45 por ciento a la clasificación final; e *intellec* es una clasificación basada en la revisión de la investigación universitaria publicada durante en un periodo de cinco años en las 20 principales revistas académicas, y en los libros universitarios revisados por *The New York Times, The Wall Street Journal* y *Bloomberg Businessweek* en el mismo período; esta última clasificación aporta el 10 por ciento a

clasificación final. En el siguiente modelo estimado *rank2010* ha sido sustituido por sus tres componentes:

$$\underset{}{salMBAgr_i} = \underset{(10700)}{79904} + \underset{(0.0696)}{0.0305}\,tuition_i + \underset{(0.107)}{0.3751}\,salMBApr_i$$
$$\underset{(94.54)}{-303.82}\,gradpoll_i - \underset{(61.26)}{33.829}\,corppoll_i - \underset{(64.09)}{113.36}\,intellec_i$$
$$R^2=0.797 \quad n=39$$

¿Cuál es el peso en porcentaje de cada uno de estas tres componentes en la determinación de *salMBAgr*? Compare los resultados con la contribución de cada una en la definición de *rank2010*.

f) ¿Son *gradpoll*, *corppoll* e *intellec* conjuntamente significativos al 5%? ¿Son individualmente significativos al 5%?

Ejercicio 4.25 (Continuación del ejercicio 3.12). El modelo poblacional que corresponde a este ejercicio es el siguiente:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 tenure + \beta_4 age + u$$

Utilizando el fichero *wage06s*, se ha estimado el modelo anterior:

$$\ln(wage)_i = \underset{(0.073)}{1.565} + \underset{(0.0035)}{0.0448}\,educ_i + \underset{(0.0019)}{0.0177}\,tenure_i + \underset{(0.0016)}{0.0065}\,age_i$$
$$R^2=0.337 \quad n=800$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la significatividad global del modelo.

b) ¿Es *tenure* estadísticamente significativa al 10%? ¿Es *age* positivamente significativa al 10%?

c) ¿Es admisible que el coeficiente de *educ* sea igual al de *tenure*? ¿Es admisible que el coeficiente de *educ* sea el triple del de *tenure*? Para responder a estas preguntas dispone de la siguiente información adicional:

$$\ln(wage)_i = \underset{(0.073)}{1.565} + \underset{(0.0042)}{0.0271}\,educ_i + \underset{(0.0019)}{0.0177}(educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

$$\ln(wage)_i = \underset{(0.073)}{1.565} - \underset{(0.0071)}{0.0082}\,educ_i + \underset{(0.0019)}{0.0177}(3\times educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

¿Se puede calcular el R^2 en las dos ecuaciones del apartado *c*)? Por favor, si la respuesta es positiva, hágalo.

Ejercicio 4.26 (Continuación del ejercicio 3.13). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. En el modelo estimado, con el fichero *housecan*, los errores estándar de los coeficientes aparecen entre paréntesis:

$$price_i = \underset{(3379)}{-2418} + \underset{(1207)}{5827}\,bedrooms_i + \underset{(1785)}{19750}\,bathrms_i + \underset{(0.388)}{5.411}\,lotsize_i$$
$$R^2=0.486 \quad n=546$$

a) Contraste la significatividad global del modelo.

b) Contraste la hipótesis nula de que un baño adicional tiene la misma influencia sobre el precio de la vivienda que 4 dormitorios adicionales. Por otra parte, contraste si un cuarto de baño adicional tiene más influencia sobre el precio de la vivienda que 4 dormitorios adicionales. (Información adicional: $\text{var}(\hat{\beta}_2)$ =1455813, $\text{var}(\hat{\beta}_3)$ =3186523 y $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =-764846).

c) Si añadimos al modelo el regresor *stories* (número de plantas, excluido el sótano), se obtienen los siguientes resultados:

$$price_i = \underset{(3603)}{-4010} + \underset{(1215)}{2825}\, bedrooms_i + \underset{(1734)}{17105}\, bathrms_i + \underset{(0.369)}{5.429}\, lotsize_i + \underset{(1008)}{7635}\, stories_i$$

$$R^2=0.536 \quad n=546$$

¿Cuál es su opinión acerca del signo y magnitud del coeficiente de *stories*? ¿Es un resultado sorprendente? ¿Cuál es la interpretación de este coeficiente? Estime si el número de *stories* tiene una influencia significativa sobre el precio de la vivienda.

d) Repita los contrastes del apartado *b*) con el modelo estimado en el apartado *c*). (Información adicional: $\text{var}(\hat{\beta}_2) = 1475758$, $\text{var}(\hat{\beta}_3) = 3008262$ y $\text{var}(\hat{\beta}_2, \hat{\beta}_3) = -554381$).

Ejercicio 4.27 (Continuación del ejercicio 3.14). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. Usando el fichero *ceoforbes*, el modelo estimado es el siguiente:

$$\ln(salary)_i = \underset{(0.377)}{4.641} + \underset{(0.0033)}{0.0054}\, roa_i + \underset{(0.0425)}{0.2893}\ln(sales_i) + \underset{(0.0000220)}{0.0000564}\, profits_i + \underset{(0.0032)}{0.0122}\, tenure_i$$

$$R^2=0.232 \quad n=447$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Tiene *roa* un efecto significativo sobre el salario? ¿Tiene *roa* un efecto positivamente significativo sobre el salario? Realice ambos contrastes para el 10% y el 5% de nivel de significación.

b) En el caso de que *roa* se incrementara en 20 puntos, ¿en qué porcentaje se incrementaría el salario?

c) Contraste la hipótesis nula de que la elasticidad salario/ventas es igual a 0.4.

d) Si a este modelo añadimos el regresor *age*, se obtienen los siguientes resultados:

$$\ln(salary)_i = \underset{(0.442)}{4.159} + \underset{(0.0033)}{0.0055}\, roa_i + \underset{(0.0423)}{0.2903}\ln(sales_i) + \underset{(0.0000220)}{0.0000539}\, profits + \underset{(0.0035)}{0.00924}\, tenure_i + \underset{(0.0043)}{0.00880}\, age_i$$

$$R^2=0.240 \quad n=447$$

¿Son los coeficientes estimados muy diferentes de los obtenidos en la estimación del modelo de referencia? ¿Qué ocurre con el coeficiente de *tenure*? Explíquelo.

e) ¿Tiene la edad (*age*) del director ejecutivo un efecto significativo sobre el salario?

f) ¿Es admisible que el coeficiente de *age* sea igual al coeficiente de *tenure*? (Información adicional: $\text{var}(\hat{\beta}_5) = 1.24\text{E-}05$; $\text{var}(\hat{\beta}_6) = 1.82\text{E-}05$ y $\text{var}(\hat{\beta}_5, \hat{\beta}_6) = -6.09\text{E-}06$).

Ejercicio 4.28 (Continuación del ejercicio 3.15). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. Utilizando el fichero *rdspain*, el modelo estimado fue el siguiente:

$$rdintens_i = \underset{(0.428)}{-1.8168} + \underset{(0.0278)}{0.1482}\ln(sales_i) + \underset{(0.0021)}{0.0110}\,exponsal_i$$

$$R^2=0.048 \quad n=1983$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Es la variable de ventas (*sales*) individualmente significativa al 1%?

b) Contraste la hipótesis nula de que el coeficiente de las ventas es igual a 0.2?

c) Contraste la significatividad global del modelo de referencia.

d) Si añadimos el regresor ln(*workers*), - donde *workers* es el número de trabajadores-, se obtienen los siguientes resultados:

$$rdintens = \underset{(0.750)}{0.480} - \underset{(0.0687)}{0.08585}\ln(sales) + \underset{(0.0021)}{0.01049}\,exponsal + \underset{(0.09198)}{0.3422}\ln(workers)$$

$$R^2=0.055 \quad n=1983$$

¿Es la variable ventas individualmente significativa al 1% en el nuevo modelo estimado?

e) Contraste la hipótesis nula de que el coeficiente de ln(*workers*) es mayor que 0.5

Ejercicio 4.29 (Continuación del ejercicio 3.16). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. Utilizando el fichero *hedcarsp*, el modelo estimado fue el siguiente:

$$\ln(price)_i = \underset{(0.154)}{14.42} + \underset{(0.0000438)}{0.000581}cid_i + \underset{(0.0079)}{0.003823}\,hpweight_i - \underset{(0.0122)}{0.07854}\,fueleff_i$$

$$R^2=0.830 \quad n=214$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Cuál de las variables explicativas incluidas en el modelo de referencia son individualmente significativas al 1%?

b) Añada la variable *volume* (volumen) al modelo de referencia. ¿Tiene *volume* un efecto estadísticamente significativo sobre ln(*price*)? ¿Tiene el volumen tiene un efecto positivo estadísticamente significativo sobre el ln(*price*)?

c) ¿Es admisible que el coeficiente estimado de *volume* en el apartado *b*) sea igual pero con signo opuesto que el coeficiente de *fueloff*?

d) Añada las variables *length*, *width* y *height* (longitud, anchura y altura) al modelo estimado en el apartado *b*). Teniendo en cuenta que *volume=length*×*width*×*height*, ¿hay multicolinealidad perfecta en el nuevo modelo? ¿Por qué? ¿Por qué no? Estime el nuevo modelo si es posible.

e) Añada la variable ln(*volume*) al modelo de referencia. Contraste la hipótesis nula de que la elasticidad de precio/volumen es igual a 1?

f) ¿Qué sucedería si añade al modelo estimado en el apartado *e*) los regresores ln(*length*), ln(*width*) y ln(*height*)?

Ejercicio 4.30 (Continuación del ejercicio 3.17). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. Utilizando el fichero *timuse03*, el correspondiente modelo ajustado fue el siguiente:

$$housework_i = \underset{(23.27)}{141.9} + \underset{(1.621)}{3.850}\,educ_i - \underset{(0.00539)}{0.00917}\,hhinc_i + \underset{(0.311)}{1.767}\,age_i - \underset{(0.0229)}{0.2289}\,paidwork_i$$

$$R^2=0.1440 \quad n=1000$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Cuál de las variables explicativas incluidas en el modelo de referencia son individualmente significativas al 5% y al 1%?

b) Estime un modelo en el que se pueda contrastar directamente si un año más de educación tiene el mismo efecto sobre el tiempo dedicado al trabajo doméstico que 2 años de edad adicionales. ¿Cuál es su conclusión?

c) Contraste la significación conjunta de *educ* y *hhnc*.

d) Estime una regresión en la que se añade la variable *childup3* (número de niños de hasta 3 años) al modelo de referencia. En el nuevo modelo, ¿cuales de los regresores son individualmente significativos al 5% y al 1%?

e) En el modelo formulado en el apartado *d*), ¿cuál es la variable más influyente? ¿Por qué?

Ejercicio 4.31 (Continuación del ejercicio 3.18). Tomemos el modelo poblacional de este ejercicio como modelo de referencia. Utilizando el fichero *hdr2010*, el correspondiente modelo ajustado fue el siguiente:

$$\underset{}{stsfglo_i} = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i$$

$$R^2=0.642 \qquad n=144$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) ¿Cuál de las variables explicativas incluidas en el modelo de referencia son individualmente significativas al 1%?

b) Estime un modelo en el que se han añadido las variables *popnosan* (porcentaje de población sin acceso a servicios de saneamiento mejorado) y *gnirank* (rango de la renta nacional bruta) al modelo de referencia. ¿Cuál de los regresores incluidos en el nuevo modelo son individualmente significativos al 1%? Interprete los coeficientes de *popnosan* y *gnirank*.

c) ¿Son *popnosan* y *gnirank* conjuntamente significativos?

d) Contraste la significatividad global del modelo formulado en el apartado b).

Ejercicio 4.32 Con una muestra de 42 observaciones se ha estimado el siguiente modelo

$$\hat{y}_t = -670.591 + 1.008 x_t$$

Para la observación 43 se sabe que el valor de x es 1571.9.

a) Calcule el predictor puntual para la observación 43.

b) Sabiendo que la varianza de del error de predicción $\hat{e}_2^{43} = y^{43} - \hat{y}^{43}$ es igual a $(24.9048)^2$, calcule un intervalo de probabilidad del 90% del valor individual.

Ejercicio 4.33 Sobre la función de consumo Brown, además de la estimación presentada en el ejercicio 4.23, se dispone de la siguiente estimación:

$$\underset{}{conspc_t} = \underset{(64.35)}{12729} + \underset{(0.0857)}{0.3965}(incpc_t - 13500) + \underset{(0.0903)}{0.5771}(conspc_{t-1} - 12793.6)$$

$$R^2=0.997; \qquad SCR=1891320; \qquad n=56$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Obtenga el predictor puntual del consumo per cápita en 2011, sabiendo que $conspc_{2010}$=12793.6 y $incpc_{2011}$=13500.

b) Obtenga un intervalo de confianza del 95% para el valor esperado del consumo per cápita en 2011.

c) Obtenga un intervalo de predicción del 95% para el valor individual del consumo per cápita en 2011.

Ejercicio 4.34 (Continuación del ejercicio 4.30) Conteste a las siguientes preguntas:

a) Utilizando la primera estimación del ejercicio 4.30, obtener una predicción para *houswork* (minutos dedicados a las tareas domésticas por día), cuando en la ecuación se hace *educ*=10 (años), *hhinc* =1200 (euros al mes), *age*=50 (años) y *paidwork*=400 (minutos por día).

b) Realice una regresión, utilizando el fichero *timuse03*, que le permita calcular un *IC* del 95% con las características tenidas en cuenta en el apartado *a*).

c) Obtenga un intervalo de predicción del 95% para el valor individual de *houswork* con las características tenidas en cuenta en el apartado *a*).

Ejercicio 4.35 (Continuación del ejercicio 4.29) Conteste a las siguientes preguntas:

a) En la primera ecuación del ejercicio 4.29 considere que *cid*=2000 (pulgadas cúbicas de desplazamiento), *hpweight*=10 (relación potencia/peso en kg, expresados en porcentaje) y *fueleff*=6 (minutos por día) Obtenga el predictor puntual del consumo per cápita en 2011, sabiendo que *incpc2011*=12793.6 y *conspc2010*=13500.

b) Obtenga una estimación consistente de *price* para las características utilizadas en el apartado *a*).

c) Realice una regresión que le permita calcular un *IC* del 95% para las características utilizadas en el apartado *a*).

d) Obtenga un intervalo de predicción del 95% para el valor individual del precio.

5 ANÁLISIS DE REGRESIÓN MÚLTIPLE CON INFORMACIÓN CUALITATIVA

5.1 Introducción de información cualitativa en los modelos econométricos

Hasta ahora las variables que hemos utilizado para explicar la variable endógena tenían un carácter cuantitativo. Sin embargo, hay otras variables de carácter cualitativo que pueden ser importantes para explicar el comportamiento de la variable endógena, como el sexo, raza, religión, nacionalidad, región geográfica, etc. Por ejemplo, manteniendo todos los demás factores constantes, se ha constatado que las mujeres trabajadoras tienen unos salarios inferiores que sus homólogos masculinos. Este resultado puede ser consecuencia de la discriminación por género, pero cualquiera que sea la razón, las variables cualitativas como el género parece que influyen en la variable endógena y deberían incluirse en muchos casos entre las variables explicativas. Los factores cualitativos, a menudo, pero no siempre, se presentan en forma de información binaria, es decir, una persona es hombre o mujer, está casada o no, etc. Cuando los factores cualitativos se presentan en forma dicotómica la información relevante puede mostrarse como una variable binaria o una variable de cero-uno. En econometría, las variables binarias que se utilizan como regresores son comúnmente llamadas variables ficticias. En la definición de una variable dicotómica, debemos decidir a qué caso se le asigna el valor 1 y a cual se le asigna el valor 0.

En el caso del género podemos definir

$$mujer = \begin{cases} 1 & \text{si la persona es una mujer} \\ 0 & \text{si la persona es un hombre} \end{cases}$$

Pero, por supuesto, también podemos definir

$$hombre = \begin{cases} 1 & \text{si la persona es un hombre} \\ 0 & \text{si la persona es una mujer} \end{cases}$$

Es importante señalar que ambas variables, mujer y hombre, contienen la misma información. Utilizar las variables cero-uno para captar información cualitativa es una decisión arbitraria, pero con esta elección los parámetros tienen una interpretación natural.

5.2 Una sola variable ficticia independiente.

Vamos a analizar cómo se puede incorporar la información dicotómica en los modelos de regresión. Considere el siguiente modelo para la determinación del *salario* por hora, en función de los años de educación (*educ*):

$$salario = \beta_1 + \beta_2 educ + u \tag{5-1}$$

Para medir la discriminación salarial debida al género se introduce una variable ficticia (*mujer*) como variable independiente en el modelo definido anteriormente,

$$salario = \beta_1 + \delta_1 mujer + \beta_2 educ + u \tag{5-2}$$

El *atributo* género tiene dos *categorías: mujer* y *hombre*. La categoría *mujer* ha sido incluida en el modelo; mientras que la categoría *hombre*, que ha sido omitida, es la *categoría de referencia*. El modelo (5-2) se muestra en la figura 5.1, tomando $\delta_1<0$. La interpretación de δ_1 es la siguiente: δ_1 es la diferencia en el salario por hora entre mujeres y hombres, dado el mismo nivel de educación (y el mismo término de perturbación, *u*). Así, el coeficiente δ_1 determina si existe una discriminación contra las mujeres o no. Si $\delta_1<0$, entonces, para el mismo nivel de otros factores (educación, en este caso), las mujeres ganan menos que los hombres en promedio. Suponiendo que la esperanza de la perturbación es cero, si se toman esperanzas en ambas categorías se obtiene:

$$\begin{aligned} \mu_{salario|mujer} &= E(salario \mid mujer = 1, educ) = \beta_1 + \delta_1 + \beta_2 educ \\ \mu_{salario|hombre} &= E(salario \mid mujer = 0, educ) = \beta_1 + \beta_2 educ \end{aligned} \tag{5-3}$$

Como puede verse en (5-3), el término independiente para los hombres es β_1, y $\beta_1+\delta_1$ para las mujeres. Gráficamente, como puede verse en la figura 5.1, hay un desplazamiento del término independiente, pero las líneas para hombres y mujeres son paralelas.

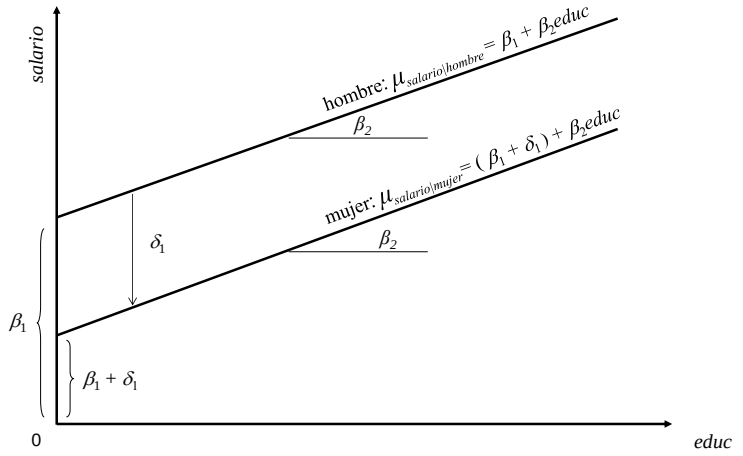


FIGURA 5.1. Misma pendiente, término independiente diferente.

En (5-2) hemos incluido una variable ficticia para las mujeres, pero no para los hombres porque incluir las dos variables ficticias habría sido redundante. De hecho, todo lo que necesitamos es dos términos independientes, uno para mujeres y otro para los hombres. Como hemos visto, con la introducción de la variable ficticia *mujer*, nos permite obtener un término independiente para cada género. La introducción de dos variables ficticias causaría multicolinealidad perfecta, ya que *mujer+hombre*=1, lo que significa que *hombre* es una función lineal exacta de *mujer* y del término independiente. La inclusión de variables ficticias para ambos sexos, además el término independiente, es el ejemplo más sencillo de la llamada trampa de las variables ficticias, como veremos más adelante.

Si usamos *hombre* en lugar de *mujer*, la ecuación de salarios sería la siguiente:

$$salario = \alpha_1 + \gamma_1 hombre + \beta_2 educ + u \tag{5-4}$$

Nada ha cambiado con la nueva ecuación, con excepción de la interpretación de α_1 y γ_1: α_1 es el término independiente para las mujeres, que ahora es la *categoría de referencia*, y $\alpha_1+\gamma_1$ es el término independiente para los hombres. Esto implica la siguiente relación entre los coeficientes:

$$\alpha_1=\beta_1+\delta_1 \text{ y } \alpha_1+\gamma_1=\beta_1 \Rightarrow \gamma_1=-\delta_1$$

En cualquier aplicación, no importa cómo elijamos la categoría de referencia, ya que sólo afecta a la interpretación de los coeficientes asociados a las variables ficticias, pero es importante tener presente qué categoría es la categoría de referencia. La elección de una categoría de referencia es, generalmente, una cuestión de conveniencia. También es posible eliminar el término independiente e incluir una variable ficticia para cada categoría. La ecuación será entonces

$$salario = \mu_1 hombre + \nu_1 mujer + \beta_2 educ + u \tag{5-5}$$

donde el término independiente es μ_1 para los hombres y ν_1 para las mujeres.

El contraste de hipótesis se realiza como de costumbre. En el modelo (5-2) la hipótesis nula de no discriminación entre hombres y mujeres es $H_0 : \delta_1 = 0$, mientras que la hipótesis alternativa de que existe discriminación contra la mujer es $H_1 : \delta_1 < 0$. Por lo tanto, en este caso, debemos aplicar un contraste *t* de una sola cola (la izquierda).

En las especificaciones formuladas en el trabajo aplicado es usual transformar la variable dependiente tomando logaritmos, ln(*y*), en modelos de este tipo. Por ejemplo:

$$\ln(salario) = \beta_1 + \delta_1 mujer + \beta_2 educ + u \tag{5-6}$$

Veamos la interpretación del coeficiente de la variable ficticia en un modelo con logaritmos. En el modelo (5-6), tomando u=0, el salario para una mujer y para un hombre son los siguientes:

$$\ln(salario_M) = \beta_1 + \delta_1 + \beta_2 educ \tag{5-7}$$

$$\ln(salario_H) = \beta_1 + \beta_2 educ \tag{5-8}$$

Dado el mismo nivel de educación, si restamos (5-8) de (5-7), tenemos

$$\ln(salario_M) - \ln(salario_H) = \delta_1 \tag{5-9}$$

Tomando antilogaritmos en (5-9) y restando 1 de ambos miembros de (5-9), obtenemos

$$\frac{salario_M}{salario_H} - 1 = e^{\delta_1} - 1 \tag{5-10}$$

es decir

$$\frac{salario_M - salario_H}{salario_H} = e^{\delta_1} - 1 \tag{5-11}$$

De acuerdo con (5.11), la tasa de variación entre el salario femenino y el salario de los hombres, para un mismo nivel de educación, es igual a $e^{\delta_1} - 1$. Por lo tanto, la

tasa exacta de variación porcentual entre el salario por hora de hombres y mujeres es de $100\times(e^{\delta_1}-1)$. Como una aproximación a este cambio se puede utilizar $100\times\delta_1$, pero si la magnitud del porcentaje es alta esta aproximación no es tan buena.

EJEMPLO 5.1 ¿Existe discriminación salarial para la mujer en España?

Utilizando datos de la *Encuesta de Estructura Salarial de Españ*a para 2002 (fichero *wage02sp*), se ha estimado el modelo (5-6) y se han obtenido los siguientes resultados:

$$\ln(wage) = \underset{(0.026)}{1.731} - \underset{(0.022)}{0.307}\ female + \underset{(0.0025)}{0.055}\,educ$$

$$SCR=393 \quad R^2=0.243 \quad n=2000$$

donde *wage* es el salario hora en euros, *female* es una variable ficticia que toma el valor 1 si es mujer, y *educ* son los años de educación. (Los números entre paréntesis son los errores estándar de los estimadores.)

Para responder a la pregunta planteada más arriba, tenemos que contrastar $H_0:\delta_1=0$ contra de $H_1:\delta_1<0$. Dado que el estadístico *t* es igual a -14.27 se rechaza la hipótesis nula para α=0.01. Es decir, hay evidencias de una discriminación en España contra la mujer en el año 2002. De hecho, la diferencia porcentual en el salario por hora entre hombres y mujeres es $100\times(e^{0.307}-1)=35.9\%$, dados unos mismos años en educación.

EJEMPLO 5.2 Análisis de la relación entre la capitalización de mercado y el valor contable: el papel del IBEX35

Un investigador desea estudiar la relación entre la capitalización de mercado y el valor contable de las acciones cotizadas en el mercado continuo de la Bolsa de Madrid. En este mercado algunas empresas están incluidas en el Ibex35, que es un índice selectivo. El investigador también quiere saber si acciones incluidas en el Ibex 35 tienen una mayor capitalización en promedio. Con este propósito en mente, el investigador formula el siguiente modelo:

$$\ln(marktval) = \beta_1 + \delta_1 ibex35 + \beta_2 \ln(bookval) + u \qquad (5\text{-}12)$$

- *marktval* es el valor de mercado de una compañía. Se calcula multiplicando el precio de la acción por el número de acciones emitidas.
- *bookval* es el valor contable de una compañía. También se conoce como valor neto de la compañía. El valor contable se calcula como la diferencia entre los activos de una compañía y sus pasivos.
- *ibex35* es una variable ficticia que toma el valor 1 si la compañía está incluida en el selectivo Ibex 35.

Utilizando las 92 compañías que cotizaron el 15 de noviembre 2011, y que suministraron información sobre el valor contable (fichero *bolmad11*), se obtuvieron los siguientes resultados:

$$\ln(marktval) = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.690}\,ibex35 + \underset{(0.037)}{0.675}\ln(bookval)$$

$$SCR=35.672 \quad R^2=0.893 \quad n=92$$

La elasticidad *marktval*/*bookval* es igual a 0.690, es decir, si el valor contable se incrementa en 1%, la capitalización bursátil de las acciones que cotizan aumentará en un 0.675%.

Contrastar si las acciones incluidas en el Ibex35 tienen en promedio una mayor capitalización implica contrastar $H_0:\delta_1=0$ contra $H_1:\delta_1>0$. Dado que el estadístico *t* es (0.690/0.179) =3.85, entonces rechazamos la hipótesis nula para los niveles habituales de significación. Por otro lado, vemos que las acciones incluidas en el Ibex35 cotizan un 99.4% más elevado que las acciones no incluidas. El porcentaje se obtiene como sigue: $100\times(e^{0.690}-1)=99.4\%$

EJEMPLO 5.3 ¿Gastan más en pescado las personas que viven en zonas urbanas que las que viven en zonas rurales?

Para ver si las personas que viven en zonas urbanas gastan más en pescado que las personas que viven en zonas rurales, se ha propuesto el modelo siguiente:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + u \qquad (5\text{-}13)$$

donde *fish* es gasto en pescado, *urban* es una variable ficticia que toma el valor 1 si la persona vive en una zona urbana e *inc* es la renta disponible.

Utilizando una muestra de tamaño 40 (fichero *demand*), se estimó el modelo (5-13):

$$\ln(fish) = \underset{(0.511)}{-6.375} + \underset{(0.055)}{0.140}\,urban + \underset{(0.070)}{1.313}\ln(inc)$$

$$SCR=1.131 \quad R^2=0.904 \quad n=40$$

De acuerdo con estos resultados, las personas que viven en zonas urbanas gastan en pescado aproximadamente un 14% más que las personas que viven en zonas rurales. Si se contrasta $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 > 0$, constatamos que el estadístico *t* es (0.140/0.055)=2.55. Teniendo en cuenta que $t_{37}^{0.01} \approx t_{35}^{0.01}$=2.44, se rechaza la hipótesis nula en favor de la alternativa para los niveles habituales de significación. Es decir, hay evidencia empírica de que las personas que viven en las zonas urbanas gastan más en pescado que las personas que viven en las zonas rurales.

5.3 Categorías múltiples para un atributo

En el epígrafe anterior consideramos un atributo (género) que tiene dos categorías (mujer y hombre). Ahora vamos a considerar atributos con más de dos categorías. En concreto, vamos a examinar un atributo con 3 categorías

Para medir el impacto del tamaño de la empresa sobre el salario, podemos utilizar variables dicotómicas. Supongamos que las empresas se clasifican en tres grupos según su tamaño: pequeñas (hasta 49 trabajadores), medianas (de 50 a 199 trabajadores) y grandes (más de 199 trabajadores). Con esta información podemos construir 3 variables ficticias:

$$pequeña = \begin{cases} 1 & \text{hasta 49 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$

$$mediana = \begin{cases} 1 & \text{de 50 a 199 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$

$$grande = \begin{cases} 1 & \text{mas de 199 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$

Si queremos explicar el salario por hora introduciendo en el modelo el tamaño de la empresa, es necesario omitir una de las categorías. En el siguiente modelo la categoría omitida son las empresas pequeñas:

$$salario = \beta_1 + \theta_1 mediana + \theta_2 grande + \beta_2 educ + u \qquad (5\text{-}14)$$

La interpretación de los coeficientes θ_j es la siguiente: θ_1 (θ_2) es la diferencia en el salario por hora entre las empresas medianas (grandes) y las pequeñas, dado un mismo nivel de educación (y un mismo término de perturbación, *u*).

Vamos a ver qué pasa si también incluimos en (5-14) la categoría *pequeña*. En ese caso, tendríamos el siguiente modelo:

$$salario = \beta_1 + \theta_0 pequeña + \theta_1 mediana + \theta_2 grande + \beta_2 educ + u \qquad (5\text{-}15)$$

Ahora, consideremos que tenemos una muestra de seis observaciones: las observaciones 1 y 2 corresponden a empresas pequeñas, la 3 y la 4 a medianas, y la 5 y 6 a grandes. En este caso, la matriz **X** de regresores tendría la siguiente configuración:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & educ_1 \\ 1 & 1 & 0 & 0 & educ_2 \\ 1 & 0 & 1 & 0 & educ_3 \\ 1 & 0 & 1 & 0 & educ_4 \\ 1 & 0 & 0 & 1 & educ_5 \\ 1 & 0 & 0 & 1 & educ_6 \end{bmatrix}$$

Como puede verse en la matriz **X**, la columna 1 de esta matriz es igual a la suma de las columnas 2, 3 y 4. Por lo tanto, existe multicolinealidad perfecta, debido a la llamada *trampa de las variables ficticias*. Generalizando, si un atributo tiene *g* categorías, en el modelo únicamente tenemos que incluir *g*-1 variables ficticias, junto con el término independiente. El término independiente para la categoría de referencia es el término independiente general del modelo, y el coeficiente de la variable ficticia de un grupo particular representa la diferencia estimada entre los términos independientes entre esa categoría y la categoría de referencia. Si incluimos *g* variables ficticias, junto con un término independiente se caerá en la trampa de las variables ficticias. Una alternativa es incluir *g* variables ficticias, y excluir el término independiente general. En el caso que nos ocupa, el modelo sería el siguiente:

$$salario = \theta_0 pequeña + \theta_1 mediana + \theta_2 grande + \beta_1 educ + u \qquad (5\text{-}16)$$

Esta solución no es aconsejable por dos razones. Con esta configuración del modelo es más difícil contrastar las diferencias con respecto a una categoría de referencia. En segundo lugar, esta solución sólo funciona en el caso de un modelo con sólo un atributo.

EJEMPLO 5.4 ¿Influye el tamaño de la empresa en la determinación de los salarios?

Utilizando la muestra del ejemplo 5.1 (fichero *wage02sp*), se estimó el modelo (5-14), tomando log en *wage*:

$$\ln(wage) = \underset{(0.027)}{1.566} + \underset{(0.025)}{0.281} medium + \underset{(0.024)}{0.162} large + \underset{(0.003)}{0.048} educ$$

$$SCR=406 \quad R^2=0.218 \quad n=2000$$

donde *medium* y *large* son dos variables dicotómicas para designar a empresas de tamaño mediano y grande respectivamente

Para responder a la pregunta inicial no haremos un contraste *individual* de θ_1 o θ_2. En vez de ello, contrastaremos conjuntamente si el tamaño de las empresas tiene una influencia significativa sobre el salario. Es decir, debemos contrastar si las medianas y grandes empresas tomadas conjuntamente tienen una influencia significativa en la determinación del salario. En este caso, las hipótesis nula y alternativa, tomando a (5-14) como el modelo no restringido, serán las siguientes:

$$H_0 : \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El modelo restringido en este caso es el siguiente:

$$\ln(wage) = \beta_1 + \beta_2 educ + u \qquad (5\text{-}17)$$

La estimación de este modelo es la siguiente:

$$\ln(wage) = \underset{(0.026)}{1.657} + \underset{(0.003)}{0.053} educ$$

$$SCR=433 \quad R^2=0.166 \quad n=2000$$

Por lo tanto, el estadístico *F* es

$$F = \frac{\left[SCR_R - SCR_{NR}\right]/q}{SCR_{NR}/(n-k)} = \frac{\left[433-406\right]/2}{406/(2000-4)} = 66.4$$

Así, de acuerdo con el valor del estadístico *F*, se puede concluir que el tamaño de la firma tiene una influencia significativa en la determinación de los salarios para los niveles usuales de significación.

Ejemplo 5.5 En el caso de Lydia E. Pinkham, ¿son significativas las variables temporales ficticias de forma individual y conjunta?

En el ejemplo 3.4 vimos el caso de Lydia E. Pinkham en el que las ventas, *sales*, de un extracto de hierbas de esa empresa (en miles de dólares) se explicaba en términos del gasto en publicidad en miles de dólares (*advexp*) así como las ventas del año anterior ($sales_{t-1}$). Sin embargo, su autor, además de estas dos variables, incluyó tres variables temporales ficticias: *d*1, *d*2 y *d*3. Estas variables ficticias abarcan las distintas situaciones por las que pasó la compañía. Así, *d*1 toma el valor 1 en el periodo de 1907-1914 y 0 en los periodos restantes, *d*2 toma el valor 1 en el periodo de 1915-1925 y 0 en otros periodos, y finalmente, *d*3 toma el valor 1 en el periodo de 1926 a 1940 y 0 en los restantes periodos. Por lo tanto, la categoría de referencia es el periodo 1941-1960. En consecuencia, la formulación final del modelo fue la siguiente:

$$sales_t = \beta_1 + \beta_2 advexp_t + \beta_3\, sales_{t-1} + \beta_4 d1_t + \beta_5 d2_t + \beta_6 d3_t + u_t \qquad (5\text{-}18)$$

Los resultados obtenidos en la regresión, utilizando el fichero *pinkham*, fueron los siguientes:

$$\underset{}{sales_t} = \underset{(96.3)}{254.6} + \underset{(0.136)}{0.5345}\, advexp_t + \underset{(0.0814)}{0.6073}\, sales_{t-1} - \underset{(89)}{133.35}\, d1_t + \underset{(67)}{216.84}\, d2_t - \underset{(67)}{202.50}\, d3_t$$

$$R^2 = 0.929 \qquad n = 53$$

Para contrastar si las variables ficticias de forma individual tienen un efecto significativo en las ventas, las hipótesis nula y alternativa son:

$$\begin{cases} H_0 : \theta_i = 0 \\ H_1 : \theta_i \neq 0 \end{cases} \qquad i = 1,2,3$$

Los correspondientes estadísticos *t* son los siguientes:

$$t_{\hat{\theta}_1} = \frac{-133.35}{89} = -1.50 \qquad t_{\hat{\theta}_2} = \frac{216.84}{67} = 3.22 \qquad t_{\hat{\theta}_3} = \frac{-202.50}{67} = -3.02$$

Como puede verse, el regresor *d*1 no es significativo para los niveles habituales de significación, mientras que por el contrario los regresores *d*2 y *d*3 son significativos para cualquiera de los niveles habituales.

La interpretación del coeficiente del regresor *d*2, por ejemplo, es la siguiente: manteniendo fijo el gasto en publicidad y dadas las ventas del año anterior, las ventas para un año del periodo 1915-1920 son 21.684 dólares mayores, en promedio, que las de un año cualquiera del periodo 1941-1960.

Para estimar el efecto conjunto de las variables temporales ficticias, las hipótesis nula y alternativa son

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \theta_3 = 0 \\ H_1 : H_0 \text{ no es cierto} \end{cases}$$

y el contraste estadístico correspondiente es

$$F = \frac{(R^2_{NR} - R^2_R)/q}{(1-R^2_{NR})/(n-k)} = \frac{(0.9290-0.8770)/3}{(1-0.9290)/(53-6)} = 11.47$$

Para cualquiera de los niveles habituales de significación la hipótesis nula es rechazada. Por lo tanto, las variables temporales ficticias tienen un efecto significativo sobre las ventas

5.4 Varios atributos

Ahora vamos a considerar la posibilidad de tener en cuenta dos atributos para explicar la determinación del salario: el género y duración de la jornada de trabajo (a tiempo parcial y a tiempo completo). La variable ficticia *tiempar*, va ser una variable binaria que toma el valor 1 cuando el tipo de contrato es a tiempo parcial y 0 si es a

tiempo completo. En el siguiente modelo se introducen las dos variables ficticias: *mujer* y *tiempar*:

$$salario = \beta_1 + \delta_1 mujer + \phi_1 tiempar + \beta_2 educ + u \tag{5-19}$$

En este modelo, ϕ_1 es la diferencia en el salario por hora entre las personas que trabajan a tiempo parcial, para un género dado y para el mismo nivel de educación (y también el mismo término de perturbación, *u*).

Cada uno de estos dos atributos tiene una categoría de referencia, que es la categoría omitida. En este caso, *hombre* es la categoría de referencia para el género y tiempo completo para el tipo de contrato. Si tomamos las esperanzas para las cuatro categorías implicadas, se obtiene:

$$\begin{aligned}
\mu_{wage|mujer,tiempar} &= E\left[wage \mid mujer, tiempar, educ\right] = \beta_1 + \delta_1 + \phi_1 + \beta_2 educ \\
\mu_{wage|mujer,tiemcom} &= E\left[wage \mid mujer, tiemcom, educ\right] = \beta_1 + \delta_1 + \beta_2 educ \\
\mu_{wage|hombre,tiempar} &= E\left[wage \mid hombre, tiempar, educ\right] = \beta_1 + \phi_1 + \beta_2 educ \\
\mu_{wage|hombre,tiemcom} &= E\left[wage \mid hombre, tiemcom, educ\right] = \beta_1 + \beta_2 educ
\end{aligned} \tag{5-20}$$

El término independiente general en la ecuación refleja el efecto de ambas categorías de referencia, hombre y tiempo completo; es decir, la categoría de referencia es hombre con jornada a tiempo completo. En (5.20) puede verse el término independiente para cada combinación de categorías.

EJEMPLO 5.6 La influencia de género y duración de la jornada de trabajo en la determinación de los salarios

El modelo (5-19), tomando log en *wage*, se estimó utilizando datos de la Encuesta de Estructura Salarial de España para el año 2006 (fichero *wage06sp*):

$$\ln(wage) = \underset{(0.026)}{2.005} - \underset{(0.021)}{0.233}\, female - \underset{(0.027)}{0.087}\, partime + \underset{(0.002)}{0.053} educ$$

$$SCR=365 \quad R^2=0.235 \quad n=2000$$

donde *partime* es un contrato a tiempo parcial.

De acuerdo con los valores de los coeficientes y los correspondientes errores estándar, es evidente que cada una de las dos variables ficticias, *female* y *partime*, son estadísticamente significativas para los niveles habituales de significación.

EJEMPLO 5.7 Análisis del absentismo laboral en la empresa Buenosaires

Buenosaires es una empresa dedicada a la fabricación de ventiladores, habiendo tenido resultados relativamente aceptables en los últimos años. Los directivos consideran que éstos habrían sido mejores si el absentismo en la empresa no fuera tan alto. Con el propósito de analiza los factores que determinan el absentismo, se propone el siguiente modelo:

$$absent = \beta_1 + \delta_1 bluecoll + \phi_1 male + \beta_2 age + \beta_3 tenure + \beta_4 wage + u \tag{5-21}$$

donde *bluecoll* es una variable ficticia que indica que la persona es un trabajador manual (la categoría de referencia es cuello blanco), male es una variable dicotómica que toma el valor 1 si el trabajador es hombre. Las variables *tenure* y *age* son continuas que reflejan los años trabajando en la empresa y la edad respectivamente.

Utilizando una muestra de tamaño 48 (fichero *absent*), se ha estimado la siguiente ecuación

$$absent = \underset{(1.640)}{12.444} + \underset{(0.669)}{0.968}\, bluecoll + \underset{(0.712)}{2.049}\, male - \underset{(0.047)}{0.037}\, age - \underset{(0.065)}{0.151}\, tenure - \underset{(0.007)}{0.044}\, wage$$

$$SCR=161.95 \quad R^2=0.760 \quad n=48$$

Ahora vamos a ver si *bluecoll* es significativa. Contrastando $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 \neq 0$, el estadístico *t* es (0.968/0.669)=1.45. Como $t_{40}^{0.10/2}$=1.68, fracasamos en rechazar la hipótesis nula para α=0.10. Entonces no hay evidencia empírica para afirmar que el absentismo de los trabajadores

manuales (cuello azul) es diferente del de los trabajadores de oficina (cuello blanco). Pero si se contrasta $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 > 0$, como $t_{40}^{0.10}$=1.30 para α=0.10, no se puede rechazar que el absentismo de los trabajadores de cuello azul sea mayor que el de los trabajadores de cuello blanco.

Por el contrario, en el caso de la variable ficticia *male*, contrastando $H_0 : \varphi_1 = 0$ contra $H_1 : \varphi_1 \neq 0$, dado que el estadístico *t* es (2.049/0.712)=2.88 y $t_{40}^{0.01/2}$ =2.70, rechazamos que el absentismo sea igual en hombres y mujeres para los niveles habituales de significación.

EJEMPLO 5.8 Tamaño de la empresa y género en la determinación del salario

Para conocer si el tamaño de la empresa y el género, de forma conjunta, son dos factores relevantes en la determinación del salario, se formula el siguiente modelo:

$$\ln(wage) = \beta_1 + \delta_1 female + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (5\text{-}22)$$

En este caso tenemos que hacer un contraste conjunto, donde las hipótesis nula y alternativa son,

$$H_0 : \delta_1 = \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El modelo restringido en este caso es el modelo de (5-17), que se estimó en el ejemplo 5.4 (fichero *wage02sp*). La estimación del modelo no restringido es la siguiente:

$$\ln(wage) = \underset{(0.026)}{1.639} - \underset{(0.021)}{0.327}\, female + \underset{(0.023)}{0.308}\, medium + \underset{(0.023)}{0.168}\, large + \underset{(0.0024)}{0.050}\, educ$$

$$SCR=361 \quad R^2=0.305 \quad n=2000$$

El estadístico *F* es

$$F = \frac{[SCR_R - SCR_{NR}]/q}{SCR_{NR}/(n-k)} = \frac{[433-361]/3}{361/(2000-5)} = 133$$

Por lo tanto, de acuerdo con el valor de *F*, se puede concluir que el tamaño de la firma y el género tienen conjuntamente una influencia significativa en la determinación del salario.

5.5 Las interacciones que implican variables ficticias

5.5.1 Interacciones entre dos variables ficticias

Para permitir la posibilidad de que exista una interacción entre el género y duración de la jornada de trabajo en la determinación salarial podemos añadir al modelo (5-19) un término de interacción entre *mujer* y *tiempar*, con lo que el modelo a estimar será el siguiente:

$$salario = \beta_1 + \delta_1 mujer + \phi_1 tiempar + \varphi_1 mujer \times tiempar + \beta_2 educ + u \qquad (5.23)$$

Esto permite determinar si el efecto de la duración de la jornada de trabajo en el salario depende, o no, del género. Análogamente, también permite si la influencia del género en el salario depende, o no, de la duración de la jornada de trabajo.

EJEMPLO 5.9 ¿Es la interacción entre las mujeres y el trabajo a tiempo parcial significativa?

El modelo (5-23) se estimó utilizando los datos de la Encuesta de Estructura Salarial de España para 2006 (fichero *wage06sp*):

$$\ln(wage) = \underset{(0.026)}{2.007} - \underset{(0.022)}{0.259}\, female - \underset{(0.047)}{0.198}\, partime + \underset{(0.058)}{0.167}\, female \times partime + \underset{(0.002)}{0.054}\, educ$$

$$SCR=363 \quad R^2=0.238 \quad n=2000$$

Para responder a la pregunta planteada, tenemos que contrastar $H_0 : \varphi_1 = 0$ contra $H_0 : \varphi_1 \neq 0$. Dado que el estadístico *t* es (0.167/0.058)=2.89, y teniendo en cuenta que $t_{60}^{0.01/2}$ =2.66 se rechaza la hipótesis nula en favor de la hipótesis alternativa. Por lo tanto, existe evidencia empírica de que la interacción entre *female* y *partime* es estadísticamente significativa.

EJEMPLO 5.10 ¿Discriminan las empresas pequeñas a las mujeres más, o menos, que las empresas grandes?

Para responder a esta pregunta se formula el siguiente modelo:

$$\ln(wage) = \beta_1 + \delta_1 female + \theta_1 medium + \theta_2 large + \varphi_1 female \times medium + \varphi_2 female \times large + \beta_2 educ + u \tag{5-24}$$

Utilizando la muestra del ejemplo 5.1 (fichero *wage02sp*), fue estimado el modelo (5-24):

$$\ln(wage) = \underset{(0.027)}{1.624} - \underset{(0.034)}{0.262}\, female + \underset{(0.028)}{0.361}\, medium + \underset{(0.027)}{0.179}\, large - \underset{(0.050)}{0.159}\, female \times medium - \underset{(0.051)}{0.043}\, female \times large + \underset{(0.0024)}{0.050}\, educ$$

$$SCR=359 \quad R^2=0.308 \quad n=2000$$

Si en (5-24) los parámetros φ_1 y φ_2 son igual a 0, esto implica que, en la ecuación para la determinación del salario, no hay interacción entre género y tamaño de la empresa. Así para responder a la pregunta planteada tomamos (5-24) como el modelo no restringido. Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \varphi_1 = \varphi_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

Por lo tanto, el modelo restringido es, en este caso, el modelo (5-22), que se estimó en el ejemplo 5.7. El estadístico F toma el valor

$$F = \frac{[SCR_R - SCR_{NR}]/q}{SCR_{NR}/(n-k)} = \frac{[361-359]/2}{359/(2000-7)} = 5.55$$

Para α=0.01, resulta que $F^{0.01}_{2,1993} \simeq F^{0.01}_{2,60} = 4.98$. Como F>5.61, rechazamos H_0 en favor de H_1. Si se rechaza H_0 para α=0.01, también será rechazada para los niveles de 5% y 10%. Por tanto, para los niveles usuales de significación, la interacción entre género y tamaño de empresa es relevante en la determinación del salario.

5.5.2 Interacciones entre una variable ficticia y una variable cuantitativa

Hasta ahora, en los ejemplos sobre determinación del salario se ha utilizado una variable ficticia para desplazar el término independiente o para estudiar su interacción con otra variable ficticia, pero manteniendo la pendiente de *educ* constante. Ahora bien, también se pueden utilizar las variables ficticias para desplazar pendientes si interactúan con cualquier variable explicativa continua. Por ejemplo, en el siguiente modelo la variable ficticia *mujer* interactúa con la variable continua *educ*:

$$salario = \beta_1 + \beta_2 educ + \delta_1 mujer \times educ + u \tag{5-25}$$

En este modelo, como puede verse en la figura 5.2, el término independiente es el mismo para hombres y para mujeres, pero la pendiente es mayor en hombres que en mujeres, porque δ_1 es negativa.

En el modelo de (5-25), los rendimientos de un año adicional en educación dependen del género del individuo. De hecho,

$$\frac{\partial salario}{\partial educ} = \begin{cases} \beta_2 + \delta_1 & \text{para mujeres} \\ \beta_2 & \text{para hombres} \end{cases} \tag{5-26}$$

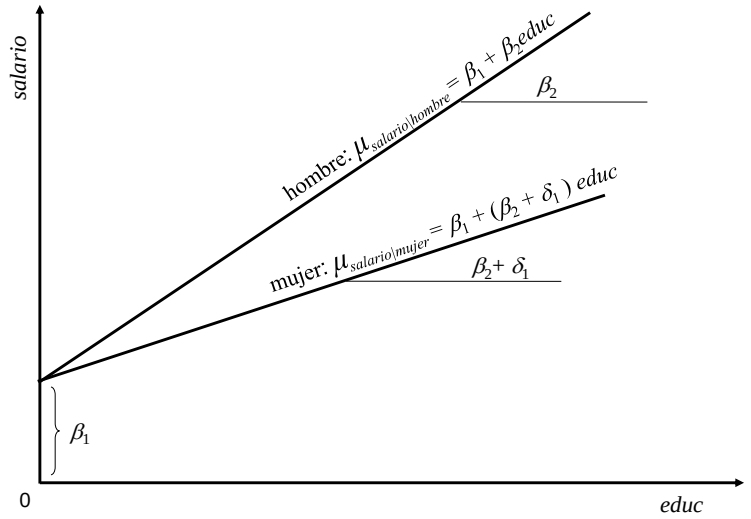


FIGURA 5.2. Diferente pendiente, mismo término independiente.

EJEMPLO 5.11 ¿Es el rendimiento de la educación para los hombres mayor que para las mujeres?

Utilizando la muestra del ejemplo 5.1 (fichero *wage02sp*), tomando logaritmos en *wage*, se ha estimado el modelo(5-25)::

$$\ln(wage) = \underset{(0.025)}{1.640} + \underset{(0.0026)}{0.063}\, educ - \underset{(0.0021)}{0.027}\, educ \times female$$

$$SCR=400 \quad R^2=0.229 \quad n=2000$$

En este caso necesitamos contrastar $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 < 0$. Dado que el estadístico *t* es (-0.028/0.0002) =-12.81, se rechaza la hipótesis nula en favor de la hipótesis alternativa para cualquier nivel de significación. Es decir, existe evidencia empírica de que el rendimiento de un año adicional de educación es mayor para hombres que para mujeres.

5.6 Contraste de cambio estructural

Hasta ahora hemos contrastado las hipótesis de que un parámetro, o un subconjunto de parámetros del modelo, son diferentes para dos grupos (mujeres y hombres, por ejemplo). Pero a veces, queremos contrastar la hipótesis nula de que dos grupos tienen la misma función de regresión poblacional, frente a la alternativa de que no es la misma. En otras palabras, queremos contrastar si la misma ecuación es válida para los dos grupos. Existen dos procedimientos para realizar este contraste, denominado *contraste de cambio estructural*: utilizando variables ficticias y realizando regresiones separadas mediante el contraste de Chow.

5.6.1 Utilizando variables ficticias

En este procedimiento, contrastar si hay diferencias entre grupos consiste en realizar un contraste de significación conjunto de la variable ficticia que diferencia entre los dos grupos y de sus interacciones con todas los otros regresores. Por lo tanto, estimamos el modelo con (*modelo no restringido*) y sin (*modelo restringido*) la variable ficticia y todas sus interacciones.

De la estimación de ambas ecuaciones se obtiene el estadístico *F*, ya sea a través de la *SCR* o del R^2. En el siguiente modelo, para la determinación del salario, tanto el término independiente como la pendiente son diferentes para hombres y mujeres:

$$salario = \beta_1 + \delta_1 mujer + \beta_2 educ + \delta_2 mujer \times educ + u \qquad (5\text{-}27)$$

En la figura 5.3, ha sido representada la función de regresión poblacional de este modelo. Como puede verse, si *mujer*=1, se obtiene que

$$salario = (\beta_1 + \delta_1) + (\beta_2 + \delta_2)educ + u \tag{5-28}$$

Entonces, para las mujeres el término independiente es $\beta_1 + \delta_1$ y la pendiente $\beta_2 + \delta_2$. Para *mujer*=0, obtenemos la ecuación (5-1) En este caso, para los hombres el término independiente es β_1 y la pendiente β_2. Por lo tanto, δ_1 mide la diferencia entre los términos independientes para mujeres y hombres y δ_2 mide a su vez la diferencia en el rendimiento la educación entre mujeres y hombres. La figura 5.3 muestra un término independiente y una pendiente menores para mujeres que para hombres. Esto significa que las mujeres ganan menos que los hombres en todos los niveles de la educación, y que la brecha aumenta a medida que *educ* se hace más grande, es decir, un año adicional de educación tiene un rendimiento inferior para mujeres que para hombres.

La estimación (5-27) es equivalente a la estimación de dos ecuaciones de salarios, uno para hombres y otro para las mujeres, por separado. La única diferencia es que (5-27) impone la misma varianza a los dos grupos, mientras que las regresiones por separado no lo hacen. Esta especificación del modelo es ideal, como veremos más adelante, para contrastar la igualdad de pendientes, la igualdad de términos independientes, o la igualdad tanto de términos independientes como de pendientes en los dos grupos.

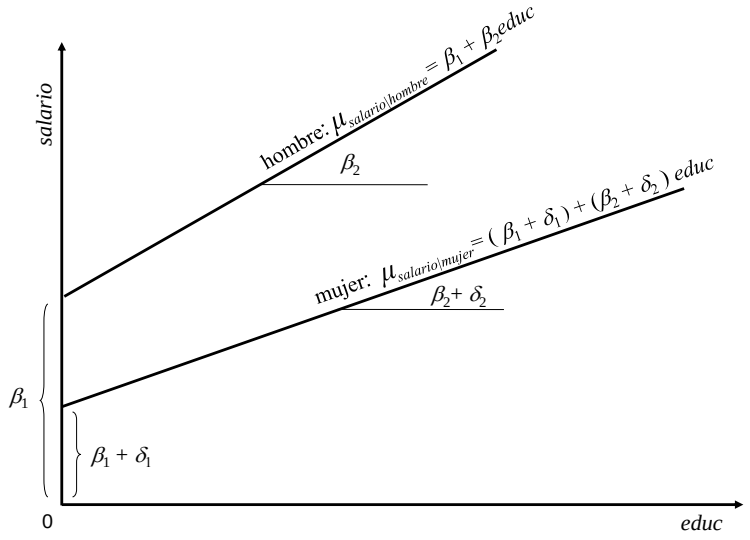


FIGURA 5.3. Pendiente diferente, diferente término independiente.

EJEMPLO 5.12 ¿Es la ecuación de salarios válida tanto para hombres como para mujeres?

Si los parámetros δ_1 y δ_2 son iguales a 0 en el modelo de (5-27), implica que la ecuación para la determinación de los salarios es la misma para hombres y mujeres. Entonces para responder a la cuestión planteada, tomamos (5-27), pero expresando el salario en logaritmos, como el modelo el modelo no restringido. Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ no es cierto}$$

Por lo tanto, el modelo restringido se obtiene aplicando la hipótesis nulaal modelo (5-27). Utilizando la misma muestra que en el ejemplo 5.1 (fichero *wage02sp*), hemos obtenido la siguiente estimación de los modelos (5-27) y (5-17):

$$\ln(wage) = \underset{(0.030)}{1.739} - \underset{(0.055)}{0.332}\ female + \underset{(0.0030)}{0.054}\,educ - \underset{(0.0054)}{0.003}\,educ \times female$$

$$SCR=393 \quad R^2=0.243 \quad n=2000$$

$$\ln(wage) = \underset{(0.026)}{1.657} + \underset{(0.0026)}{0.0525}\,educ$$

$$SCR=433 \quad R^2=0.166 \quad n=2000$$

El estadístico *F* toma el valor

$$F = \frac{\left[SCR_R - SCR_{NR}\right]/q}{SCR_{NR}/(n-k)} = \frac{\left[433-393\right]/2}{393/(2000-4)} = 102$$

Está claro que para cualquier nivel de significanción, las ecuaciones para hombres y mujeres son diferentes.

Cuando contrastamos en el ejemplo 5.1 si había discriminación contra la mujer en España ($H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 < 0$), se asumió que la pendiente de *educ* (modelo (5-6)) era la misma para hombres y mujeres. Ahora también es posible utilizar el modelo (5-27) para contrastar la misma hipótesis nula, pero asumiendo que la pendiente es diferente. Dado que el estadístico *t* es (-0.332/0.0546)=-6.06, entonces se rechaza la hipótesis nula mediante el uso de este modelo más general que el del ejemplo 5.1.

En el ejemplo 5.11 se contrastó si el coeficiente de δ_2 en el modelo de (5-25), tomando log en *wage*, era 0, suponiendo que el término independiente es el mismo para hombres y mujeres. Ahora bien, si tomamos (5-27), tomando log en *wage*, como modelo no restringido, podemos contrastar la misma hipótesis nula, pero asumiendo que el término independiente es diferente para hombres y mujeres. Dado que el estadístico *t* es (0.0027/0.0054)=0.493, entonces no se puede rechazar la hipótesis nula de que no existe interacción entre género y educación.

EJEMPLO 5.13 ¿Tienen los consumidores urbanos el mismo patrón de comportamiento que los rurales con respecto al gasto en pescado?

Para responder a esta pregunta se formula el siguiente modelo, que se tomará como modelo *no restringido*:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + \delta_2 \ln(inc) \times urban + u \qquad (5\text{-}29)$$

Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El modelo restringido correspondiente a esta H_0 es

$$\ln(fish) = \beta_1 + \beta_2 \ln(inc) + u \qquad (5\text{-}30)$$

Utilizando la muestra del ejemplo 5.3 (fichero *demand*), se han estimado los modelos (5-29) y (5-30):

$$\ln(fish) = \underset{(0.627)}{-6.551} + \underset{(1.095)}{0.678}\,urban + \underset{(0.087)}{1.337}\ln(inc) - \underset{(0.152)}{0.075}\ln(inc) \times urban$$

$$SCR=1.123 \quad R^2=0.904 \quad n=40$$

$$\ln(fish) = \underset{(0.542)}{-6.224} + \underset{(0.075)}{1.302}\ln(inc)$$

$$SCR=1.325 \quad R^2=0.887 \quad n=40$$

El estadístico *F* toma el valor

$$F = \frac{\left[SCR_R - SCR_{NR}\right]/q}{SCR_{NR}/(n-k)} = \frac{\left[1.325-1.123\right]/2}{1.123/(40-4)} = 3.24$$

Si miramos en la tabla estadística de la *F* para 2 *grados de libertad* en el numerador y 35 *gl* en el denominador para α=0.10, vemos que $F_{2,36}^{0.10} \simeq F_{2,35}^{0.10} = 2.46$. Como *F*>2.46 rechazamos la H_0. Sin embargo, como $F_{2,36}^{0.05} \simeq F_{2,35}^{0.05} = 3.27$, fracasamos en rechazar H_0 a favor de H_1 para α=0.05 y, por tanto, para α=0.01. Conclusión: no hay una evidencia fuerte de que las familias que viven en las zonas rurales

tengan un patrón de consumo diferente de pescado con respecto a las familias que viven en zonas urbanas.

Ejemplo 5.14 ¿Ha cambiado la estructura productiva de las regiones españolas?

La pregunta que debe responderse es específicamente la siguiente: ¿ha cambiado la estructura productiva de las regiones españolas entre 1995 y 2008? El problema que se plantea es un problema de estabilidad estructural. Para especificar el modelo que se toma como referencia en la estimación, vamos a definir la variable ficticia *y*2008, que toma el valor 1 si el año es 2008 y 0 si el año es 1995.

El modelo de referencia es un modelo de Cobb-Douglas, que introduce parámetros adicionales para recoger los cambios estructurales que puedan haber ocurrido. Su expresión es la siguiente:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + \gamma_2 y2008 + \alpha_2 y2008 \times \ln(k) + \beta_2 y2008 \times \ln(l) + u \qquad (5\text{-}31)$$

Es fácil ver, de acuerdo con la definición de la variable ficticia *y*2008 que las elasticidades de producción/capital en 1995 y 2008 son diferentes. En concreto, toman los siguientes valores:

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 + \alpha_2$$

En el caso de que α_2 sea igual a 0, entonces la elasticidad de la producción/capital es la misma en ambos periodos.

Del mismo modo, las elasticidades de producción/trabajo para los dos periodos vienen dadas por

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 + \beta_2$$

El término independiente de la función Cobb-Douglas es un parámetro que mide la eficiencia. En el modelo de (5-31) se considera la posibilidad de que el parámetro de eficiencia (*PEF*) sea diferente en los dos periodos. Así,

$$PEF(1995) = \gamma_1 \qquad PEF(2008) = \gamma_1 + \gamma_2$$

Si los parámetros α_2, β_2 y γ_2 son cero en el modelo (5-31), la función de producción es la misma en ambos periodos. Por lo tanto, en la estimación de estabilidad estructural de la función de producción, las hipótesis nula y alternativa son:

$$\begin{aligned} &H_0 : \gamma_2 = \ \alpha_2 = \beta_2 \\ &H_1 : H_0 \text{ no es cierta} \end{aligned} \qquad (5\text{-}32)$$

Bajo la hipótesis nula, las restricciones dadas en (5-32) conducen al modelo restringido siguiente:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + u \qquad (5\text{-}33)$$

El fichero *prodsp* contiene información para cada una de las regiones españolas en 1995 y 2008 sobre el valor añadido bruto en millones de euros (*gdp*), la ocupación en miles de puestos de trabajo (*labor*), y el capital productivo en millones de euros (*captot*). También en ese archivo se puede encontrar la variable ficticia *y*2008.

A continuación se muestran los resultados del modelo de regresión no restringido (5-31). Es evidente que no podemos rechazar la hipótesis nula de que cada uno de los coeficientes α_2, β_2 y γ_2, considerados individualmente, sea 0, ya que ninguno de los estadísticos *t* llega a 0.1 en valor absoluto.

$$\ln(gva) = \underset{(0.916)}{0.0559} + \underset{(0.185)}{0.6743}\ln(captot) + \underset{(0.185)}{0.3291}\ln(labor)$$
$$- \underset{(2.32)}{0.1088}\, y2008 + \underset{(0.419)}{0.0154}\, y2008 \times \ln(captot) - \underset{(0.418)}{0.0094}\, y2008 \times \ln(labor)$$

$$R^2 = 0.99394 \qquad n = 34$$

Los resultados del modelo restringido (5-33) son los siguientes:

$$\ln(gva) = -\underset{(0.200)}{0.0690} + \underset{(0.036)}{0.6959}\ln(captot) + \underset{(0.042)}{0.311}\ln(labor)$$

$$R^2 = 0.99392 \qquad n = 34$$

Como puede verse, las R^2 de los dos modelos son prácticamente idénticas ya que difieren sólo a partir del quinto decimal. No es de extrañar, por tanto, que el estadístico *F* para el contraste de la hipótesis nula (5-32) tenga un valor cercano a 0:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k)} = \frac{(0.99394 - 0.99392)/3}{(1-0.99394)/(34-6)} = 0.0308$$

Así pues, la hipótesis alternativa de que exista cambio estructural en la economía productiva de las regiones españolas entre 1995 y 2008 se rechaza para cualquier nivel de significación.

5.6.2 Utilizando regresiones separadas: el contraste de Chow

Este contraste fue introducido por el económetra Chow (1960). Este autor consideró el problema de contrastar la igualdad de dos conjuntos de coeficientes de regresión. En el contraste de Chow el modelo restringido es el mismo que en el caso de uso de variables ficticias para diferenciar entre grupos. Ahora, sin embargo, el modelo no restringido, en lugar de distinguir el comportamiento de dos grupos mediante variables ficticias, consiste simplemente en regresiones separadas. Así, en el ejemplo determinación de los salarios, el modelo no restringido consta de dos ecuaciones:

$$\begin{aligned} &mujer: \quad salario = \beta_{11} + \beta_{21} educ + u \\ &hombre: \quad salario = \beta_{12} + \beta_{22} educ + u \end{aligned} \tag{5-34}$$

Si estimamos ambas ecuaciones por *MCO*, se puede demostrar que la *SCR* del modelo no restringido, SCR_{NR}, es igual a la suma de la *SCR* obtenida de la estimación para las mujeres, SCR_1, y para los hombres, SCR_2. Es decir,

$$SCR_{NR}=SCR_1+SCR_2$$

La hipótesis nula establece que los parámetros de las dos ecuaciones en (5-34) son iguales. Entonces,

$$H_0: \begin{cases} \beta_{11} = \beta_{12} \\ \beta_{21} = \beta_{22} \end{cases}$$
$$H_1: \text{No } H_0$$

Aplicando la hipótesis nula al modelo (5-34), se obtiene el modelo (5-17), que es el modelo restringido. La estimación de este modelo para toda la muestra se suele denominar *regresión agrupada* o *pooled regression* (*P*). Por lo tanto, vamos a considerar que el SCR_R y SCR_P son expresiones equivalentes.

Por lo tanto, el estadístico *F* será la siguiente:

$$F = \frac{\left[SCR_P - \left(SCR_1 + SCR_2\right)\right]/k}{\left[SCR_1 + SCR_2\right]/\left[n-2k\right]} \tag{5-35}$$

Es importante señalar que, bajo la hipótesis nula, deben ser iguales las varianzas de la perturbación para los grupos. Observe que tenemos *k* restricciones: los *k*-1 coeficientes de pendiente (interacciones), más el coeficiente del término independiente. Nótese también que en el modelo no restringido estimamos 2 términos independientes diferentes y 2 coeficientes de pendiente diferentes, por lo que los *gl* del modelo son *n*-2*k*.

Una limitación importante del contraste de Chow es que bajo la hipótesis nula no hay diferencias en absoluto entre los grupos. En la mayoría de los casos, es más interesante permitir diferencias parciales entre ambos grupos, como hemos hecho mediante la utilización de variables ficticias.

El contraste de Chow se puede generalizar a más de dos grupos de un modo natural. Desde el punto de vista práctico, es probablemente más fácil estimar regresiones separadas para cada grupo que utilizar el procedimiento basado en la introducción de variables ficticias en el modelo.

En el caso de tres grupos el estadístico *F* en el contraste de Chow tiene la siguiente configuración:

$$F = \frac{\left[SCR_P - (SCR_1 + SCR_2 + SCR_3)\right] / 2 \times k}{(SCR_1 + SCR_2 + SCR_3) / (n - 3k)} \tag{5-36}$$

Observe que, como regla general, el número de *gl* del numerador es igual al (número de grupos-1)$\times k$, mientras que el número de *gl* del denominador es igual a *n* menos (número de grupos)$\times k$.

EJEMPLO 5.15 Otra forma de abordar la cuestión de la determinación de los salarios por criterio de género

Utilizando la misma muestra que en el ejemplo 5.1 (fichero *wage02sp*), hemos obtenido la estimación de las ecuaciones en (5-34), tomando log en *wage*, para hombres y mujeres, las cuales tomadas conjuntamente dan lugar a la estimación del modelo *no restringido*:

Ecuación para la mujer $\ln(wage) = \underset{(0.042)}{1.407} + \underset{(0.0041)}{0.057}\, educ$

SCR=104 R^2=0.236 n=617

Ecuación para el hombre $\ln(wage) = \underset{(0.031)}{1.739} + \underset{(0.0032)}{0.054}\, educ$

SCR=289 R^2=0.175 n=1383

El modelo restringido, que se estima en el ejemplo 5.4, tiene la misma configuración que las ecuaciones (5-34), pero referido en este caso para toda la muestra. Por lo tanto, es la regresión agrupada (*P*) correspondiente al modelo restringido. El estadístico *F* toma el valor

$$F = \frac{\left[SCR_P - (SCR_F + SCR_M)\right] / k}{SCR_F + SCR_M) / (n - 2k)} = \frac{\left[433 - (104 + 289)\right] / 2}{(104 + 289) / (2000 - 2 \times 2)} = 102$$

El estadístico *F* tiene que ser, y lo es, igual al del ejemplo 5.12. En consecuencia, las conclusiones son las mismas.

EJEMPLO 5.16 ¿El modelo de determinación de los salarios es el mismo para diferentes tamaños de empresa?

En otros ejemplos bien el término independiente o bien la pendiente correspondiente a la variable educación, fue diferente para tres diferentes tamaños de empresa (pequeña, mediana y grande). Ahora consideramos una ecuación completamente diferente para cada tamaño de la empresa. Por lo tanto, el modelo no restringido estará compuesto por tres ecuaciones:

$$\begin{aligned} pequeña &: \ln(wage) = \beta_{11} + \delta_{11}\, female + \beta_{21} edu + u \\ mediana &: \ln(wage) = \beta_{12} + \delta_{12}\, female + \beta_{22} edu + u \\ grande &: \ln(wage) = \beta_{13} + \delta_{13}\, female + \beta_{23} edu + u \end{aligned} \tag{5-37}$$

Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} \\ \delta_{11} = \delta_{12} = \delta_{13} \\ \beta_{21} = \beta_{22} = \beta_{23} \end{cases}$$

$$H_1 : \text{No } H_0$$

Dada esta hipótesis nula, el modelo restringido es el modelo de (5-2).

Las estimaciones de las tres ecuaciones de (5-37), utilizando el fichero *wage02sp*, son las siguientes:

pequeña $$\ln(wage) = \underset{(0.0346)}{1.706} - \underset{(0.0312)}{0.249}\, female + \underset{(0.0038)}{0.040}\, educ$$

$$SCR=121 \quad R^2=0.160 \quad n=801$$

mediana $$\ln(wage) = \underset{(0.0514)}{1.934} - \underset{(0.0390)}{0.422}\, female + \underset{(0.0046)}{0.055}\, educ$$

$$SCR=123 \quad R^2=0.302 \quad n=590$$

grande $$\ln(wage) = \underset{(0.0462)}{1.749} - \underset{(0.0385)}{0.303}\, female + \underset{(0.0044)}{0.055}\, educ$$

$$SCR=114 \quad R^2=0.273 \quad n=609$$

La regresión agrupada (*P*) ya ha sido estimada en el ejemplo 5.1. El estadístico *F* toma el valor

$$F = \frac{\left[SCR_P - (SCR_S + SCR_M + SCR_L)\right]/2\times k}{(SCR_S + SCR_M + SCR_L)/(n-3k)}$$
$$= \frac{\left[393-(121+123+114)\right]/6}{(121+123+114)/(2000-3\times 3)} = 32.4$$

Para cualquier nivel de significación, rechazamos que las ecuaciones para la determinación de los salarios sean las mismas para los tres tamaños de empresa considerados.

EJEMPLO 5.17 ¿Es el modelo Pinkham válido para los cuatro periodos?

En el ejemplo 5.5 se introdujeron variables ficticias temporales y se contrastó si el término independiente era diferente para cada periodo. Ahora, vamos a contrastar si el modelo en su conjunto es válido para los cuatro periodos considerados. Por lo tanto, el modelo restringido estará compuesto por cuatro ecuaciones:

$$\begin{aligned}
&\text{1907-1914} \quad sales_t = \beta_{11} + \beta_{21} advexp_t + \beta_{31}\, sales_{t-1} + u_t \\
&\text{1915-1925} \quad sales_t = \beta_{12} + \beta_{22} advexp_t + \beta_{32}\, sales_{t-1} + u_t \\
&\text{1926-1940} \quad sales_t = \beta_{13} + \beta_{23} advexp_t + \beta_{33}\, sales_{t-1} + u_t \\
&\text{1941-1960} \quad sales_t = \beta_{14} + \beta_{24} advexp_t + \beta_{34}\, sales_{t-1} + u_t
\end{aligned} \tag{5-38}$$

Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} \\ \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} \\ \beta_{31} = \beta_{32} = \beta_{33} = \beta_{34} \end{cases}$$
$$H_1 : \text{No } H_0$$

Dada esta hipótesis nula, el modelo restringido es el siguiente

$$sales_t = \beta_1 + \beta_2 advexp_t + \beta_3\, sales_{t-1} + u_t \tag{5-39}$$

Las estimaciones de las cuatro ecuaciones (5-38) son las siguientes:

1907-1914 $$sales_t = \underset{(603)}{64.84} + \underset{(1.025)}{0.9149}\, advexp + \underset{(0.425)}{0.4630}\, sales_{t-1} \qquad SCR = 36017 \quad n = 7$$

1915-1925 $$sales_t = \underset{(190)}{221.5} + \underset{(0.557)}{0.1279}\, advexp + \underset{(0.425)}{0.9319}\, sales_{t-1} \qquad SCR = 400605 \quad n = 11$$

1926-1940 $$sales_t = \underset{(112)}{446.8} + \underset{(0.115)}{0.4638}\, advexp + \underset{(0.0827)}{0.4445}\, sales_{t-1} \qquad SCR = 201614 \quad n = 15$$

1941-1960 $$sales_t = -\underset{(134)}{182.4} + \underset{(0.241)}{1.6753}\, advexp + \underset{(0.111)}{0.3042}\, sales_{t-1} \qquad SCR = 187332 \quad n = 20$$

La regresión agrupada, estimada en el ejemplo 3.4, es la siguiente:

$$sales_t = \underset{(95.7)}{138.7} + \underset{(0.156)}{0.3288}\, advexp + \underset{(0.0915)}{0.7593}\, sales_{t-1} \qquad SCR = 2527215 \quad n = 53$$

El estadístico *F* toma el valor

$$F = \frac{\left[SCR_P - (SCR_1 + SCR_2 + SCR_3 + SCR_4)\right]/3\times k}{(SCR_1 + SCR_2 + SCR_3 + SCR_4)/(n-4k)}$$
$$= \frac{\left[2527215 - (36017 + 400605 + 201614 + 187332)\right]/9}{(36017 + 400605 + 201614 + 187332)/(53 - 4\times 3)} = 9.16$$

Para cualquier nivel de significación, rechazamos que el modelo (5-39) sea el mismo para los cuatro periodos considerados.

Ejercicios

Ejercicio 5.1 Responda a las dos siguientes cuestiones relativas a un modelo con variables explicativas ficticias:

a) ¿Cuál es la interpretación de los coeficientes de las variables ficticias?
b) ¿Por qué no se deben incluir el mismo número de variables ficticias que categorías?

Ejercicio 5.2 Se han obtenido las siguientes estimaciones de demanda de viviendas para alquiler con una muestra de 560 familias.

$$\hat{q}_i = \underset{(0.11)}{4.17} - \underset{(0.017)}{0.247}\, p_i + \underset{(0.026)}{0.960}\, y_i$$
$$R^2=0.371 \qquad n=560$$
$$\hat{q}_i = \underset{(0.13)}{5.27} - \underset{(0.030)}{0.221}\, p_i + \underset{(0.031)}{0.920}\, y_i + \underset{(0.120)}{0.341}\, d_i y_i$$
$$R^2=0.380$$

donde q_i es el logaritmo del gasto en alquiler de vivienda de la familia *i*-esima, p_i es el logaritmo del precio de alquiler por m^2 en el área que vive la familia *i*-esima, y_i es el logaritmo de la renta familiar disponible *i*-esima y d_i es una variable ficticia que toma el valor uno si la familia reside en un municipio urbano y cero en uno rural.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste, en el primer modelo ajustado, la hipótesis de que la elasticidad del gasto en alquiler de vivienda con respecto a la renta es 1.
b) Contraste si la interacción entre la variable ficticia y la renta es significativa. ¿Existe una diferencia significativa de la elasticidad gasto en alquiler renta entre las áreas rurales y urbanas?

Ejercicio 5.3 En un modelo de regresión lineal con variables ficticias conteste a las siguientes preguntas:

a) Significado e interpretación de los coeficientes de las variables ficticias en modelos con distintas formas funcionales de la variable endógena.
b) ¿Por qué no es conveniente incluir el mismo número de ficticias que de categorías existentes en la variable cualitativa?
c) Exprese cómo se ve afectado un modelo en el que se han introducido variables ficticias en forma aditiva y otro en el que sólo se introducen en forma multiplicativa con respecto a una variable cuantitativa.

Ejercicio 5.4 En el contexto del modelo de regresión lineal múltiple,

a) ¿Qué es una variable ficticia? Ponga un ejemplo de especificación de un modelo econométrico con variables ficticias. Interprete los coeficientes, razonando la respuesta.

b) ¿Qué relación puede existir entre el problema de multicolinealidad y las variables ficticias?

Ejercicio 5.5 Con datos correspondientes a los trabajadores de un departamento de una cierta empresa se ha obtenido la siguiente estimación:

$$salario_i = 500 + 50antigüedad_i + 200niveldeestudios_i + 100hombre_i$$

donde *salari*o es el salario en euros mensuales, *antigüedad* es la antigüedad laboral medida en años, *nivelestudios* es una variable ficticia que toma valor 1 si el trabajador tiene estudios superiores y 0 en caso contrario y hombre es una variable ficticia que toma el valor 1 si el trabajador es hombre y 0 en caso contrario.

a) ¿Qué salario predeciría para una trabajadora con 6 años de antigüedad laboral y con estudios superiores?

b) Suponiendo que todas las mujeres trabajadoras tienen estudios superiores y ninguno de los hombres trabajadores tienen estudios superiores, escriba una hipotética matriz de regresores (**X**) para seis observaciones. En este caso, ¿se plantearía algún problema en la estimación del modelo? Explique su respuesta.

c) Plantee un nuevo modelo econométrico que permita dilucidar si existen diferencias salariales entre los trabajadores con estudios primarios, con estudios medios y con estudios superiores.

Ejercicio 5.6 Considere el siguiente modelo de regresión lineal:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i \qquad (1)$$

donde *y* es el salario mensual de un profesor, *x* es el número de años de experiencia docente y d_1 y d_2 son dos variables ficticias que toman los siguientes valores

$$d_{1i} = \begin{cases} 1 \text{ si el profesor es hombre} \\ 0 \text{ en todos los demás casos} \end{cases} \qquad d_{2i} = \begin{cases} 1 \text{ si el profesor es de raza blanca} \\ 0 \text{ en todos los demás casos} \end{cases}$$

a) ¿Cuál es la categoría de referencia en el modelo?

b) Interprete el significado de γ_1 y γ_2. ¿Cuál es el salario esperado para todas las categorías posibles?

Para mejorar la capacidad explicativa del modelo se consideró la siguiente especificación alternativa

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 (d_{1i} d_{2i}) + u_i \qquad (2)$$

c) ¿Cuál es el significado del término $(d_{1i} d_{2i})$? Interprete el significado de γ_3.

d) ¿Cuál es el salario esperado para todas las categorías posibles en el modelo (2)?

Ejercicio 5.7 Se ha obtenido la siguiente ecuación estimada por mínimos cuadrados ordinarios con una muestra de 36 observaciones:

$$\hat{y}_t = \underset{(0.12)}{1.10} - \underset{(0.34)}{0.96} x_{t1} - \underset{(3.35)}{4.56} x_{t2} + \underset{(0.07)}{0.34} x_{t3}$$

$$\sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 = 109.24 \quad \sum_{t=1}^{n} \hat{u}_t^2 = 20.22$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste la significatividad individual del coeficiente asociado a x_2.
b) Calcule el coeficiente de determinación, R^2, y dé una interpretación del mismo.
c) Contraste la significatividad conjunta del modelo.
d) Dos regresiones adicionales, con la misma especificación, fueron realizadas para los dos grupos, A y B, incluidos en la muestra (n_1=21 y n_2=15). En dichas estimaciones se obtuvieron las siguientes *SCR*, 11.09 y 2.17, respectivamente. Contraste si los grupos A y B tienen un distinto comportamiento.

Ejercicio 5.8 Para explicar el tiempo dedicado a actividades deportivas (*depor*) se ha formulado el siguiente modelo:

$$depor = \beta_1 + \delta_1 mujer + \varphi_1 fumador + \beta_2 edad + u \qquad (1)$$

donde *depor* son los minutos dedicados al día, en promedio, a actividades deportivas en minutos; *mujer* y *fumador* son variables ficticias que toman el valor 1 si la persona es una mujer o si fuma al menos 5 cigarrillos diarios, respectivamente. La variable *edad* está expresada en años.

a) Interprete el significado de δ_1, φ_1 y β_2.
b) ¿Cuál es el tiempo esperado dedicado a actividades deportivas para todas las categorías posibles?
c) Para mejorar la capacidad explicativa del modelo se consideró la siguiente especificación alternativa:

$$\begin{aligned} depor = {} & \beta_1 + \delta_1 mujer + \varphi_1 fumador + \gamma_1 mujer \times fumador \\ & + \delta_2 mujer \times edad + \varphi_2 fumador \times edad + \beta_2 edad + u \end{aligned} \qquad (2)$$

En el modelo (2), ¿cuál es el significado de γ_1?¿Cuál es el significado de δ_2 y φ_2 ?
d) ¿Cuáles son los posibles efectos marginales de *depor* con respecto a la *edad* en el modelo (2)? Detállelos.

Ejercicio 5.9 Utilizando información de las regiones españolas en los años 1995 y 2000 se han estimado varias funciones de producción.

Para el conjunto de los dos periodos se obtuvieron los siguientes resultados

$$\ln(q) = 5.72 + 0.26\ln(k) + 0.75\ln(l) - 1.14 f + 0.11 f \times \ln(k) - 0.05 f \times \ln(l) \qquad (1)$$

$$R^2 = 0.9594 \quad \bar{R}^2 = 0.9510 \quad SCR = 0.9380 \quad n = 34$$

$$\ln(q) = 3.91 + 0.45\ln(k) + 0.60(l) \qquad (2)$$

$$R^2 = 0.9567 \quad \bar{R}^2 = 0.9525 \quad SCR = 1.0007$$

Por otra parte, para cada uno de los años se estimaron separadamente los siguientes modelos:

1995 $$\ln(q) = 5.72 + 0.26\ln(k) + 0.75l \qquad (3)$$

$$R^2 = 0.9527 \quad \bar{R}^2 = 0.9459 \quad SCR = 0.6052$$

2000 $$\ln(q) = 4.58 + 0.37\ln(k) + 0.70l \qquad (4)$$

$$R^2 = 0.9629 \quad \bar{R}^2 = 0.9555 \quad SCR = 0.3331$$

donde *q* es producción, *k* es capital, *l* es empleo y *f* es una variable ficticia que toma el valor 1 para los datos de 1995 y 0 para los del año 2000.

a) Contraste si se produce un cambio estructural entre 1995 y 2000.
b) Compare los resultados de las estimaciones (3) y (4) con la estimación (1).
c) Contraste la significatividad global del modelo (1).

Ejercicio 5.10 Con una muestra de 300 empresas del sector de servicios, se estimó la siguiente función de coste (*cost*):

$$cost_i = 0.847 + \underset{(0.025)}{0.899}\, qty_i \qquad SCR = 901.074 \qquad n = 300$$

donde qty_i es la cantidad producida.

Las 300 empresas están distribuidas en tres grandes áreas (100 en cada una). Los resultados obtenidos fueron los siguientes:

$$\text{Área 1: } cost_i = 1.053 + \underset{(0.038)}{0.876}\, qty_i \qquad \hat{\sigma}^2 = 0.457$$

$$\text{Área 2: } cost_i = 3.279 + \underset{(0.096)}{0.835}\, qty_i \qquad \hat{\sigma}^2 = 3.154$$

$$\text{Área 3: } cost_i = 5.279 + \underset{(0.10)}{0.984}\, qty_i \qquad \hat{\sigma}^2 = 4.255$$

a) Calcule una estimación insesgada σ^2 de la función de costes para el conjunto de las 300 empresas.
b) ¿Es la misma función de coste válida para las tres áreas?

Ejercicio 5.11 Para el estudio del gasto en revistas (*rev*) se han formulado los siguientes modelos:

$$\ln(rev) = \beta_1 + \beta_2 \ln(renta) + \beta_3 edad + \beta_4 hombre + u \qquad (1)$$

$$\ln(rev) = \beta_1 + \beta_2 \ln(renta) + \beta_3 edad + \beta_4 hombre + \beta_5 prim + \beta_6 sec + u \qquad (2)$$

donde *renta* es la renta disponible, *edad* es la edad en años, *hombre* es una variable dicotómica que toma el valor 1 si es hombre, *prim* y *sec* son variables ficticias que toman el valor 1 cuando el individuo ha alcanzado, a lo sumo, los niveles primarios y secundarios de estudios, respectivamente.

Con una muestra de 100 observaciones, se han obtenido los siguientes resultados

$$\ln(rev)_i = \underset{(0.124)}{1.27} + \underset{(0040)}{0.756} \ln(renta_i) + \underset{(0.001)}{0.031} edad_i - \underset{(0.022)}{0.017}\, hombre_i$$

$$SCR{=}1.1575 \qquad R^2{=}0.9286$$

$$\ln(rev_i) = \underset{(0.020)}{1.26} + \underset{(0.007)}{0.811} \ln(renta_i) + \underset{(0.0002)}{0.030}\, edad_i + \underset{(0.003)}{0.003}\, hombre_i - \underset{(0.004)}{0.250}\, prim_i + \underset{(0.005)}{0.108}\, sec_i$$

$$SCR{=}0.0306 \qquad R^2{=}0.9981$$

a) ¿Es la educación un factor relevante para explicar el gasto en revistas? ¿Cuál es la categoría de referencia para la educación?
b) En el primer modelo, ¿es mayor el gasto en revistas para hombres que para mujeres? Justifica tu respuesta.

c) Interprete el coeficiente de la variable *hombre* en el segundo modelo. ¿Es mayor el gasto en revistas para hombres que para mujeres? Compare con el resultado obtenido en la parte *a)*.

Ejercicio 5.12 Consideremos que *fruit* es el gasto en frutas en un año, expresado en euros, realizado por un hogar y r_1, r_2, r_3, y r_4 son variables dicotómicas que reflejan las cuatro regiones de un país.

a) Si se realiza una regresión de *fruit* sobre r_1, r_2, r_3, y r_4 sin término independiente, ¿cuál es la interpretación de los coeficientes?

b) Si se realiza una regresión de *fruit* sobre r_1, r_2, r_3, y r_4 y con un término independiente, ¿qué sucedería? ¿Por qué?

c) Si se realiza una regresión de *fruit* sobre r_2, r_3, y r_4 sin término independiente, ¿cuál es la interpretación de los coeficientes?

d) Si se realiza una regresión de *fruit* sobre r_1-r_2, r_2, r_4-r_3, y r_4 sin término independiente, ¿cuál es la interpretación de los coeficientes?

Ejercicio 5.13 Considere el siguiente modelo

$$salario = \beta_1 + \delta_1 mujer + \beta_2 educ + u$$

Ahora, vamos a considerar tres posibilidades de definir la variable ficticia *mujer*:

1) $mujer = \begin{cases} 1 & \text{para mujer} \\ 0 & \text{para hombre} \end{cases}$ 2) $mujer = \begin{cases} 2 & \text{para mujer} \\ 1 & \text{para hombre} \end{cases}$ 3) $mujer = \begin{cases} 2 & \text{para mujer} \\ 0 & \text{para hombre} \end{cases}$

a) Interprete el coeficiente de la variable ficticia para cada definición.
b) ¿Es alguna definición preferible a las otras? Justifique la respuesta.

Ejercicio 5.14 Se considera el siguiente modelo de regresión:

$$salario = \beta_1 + \delta_1 mujer + u$$

donde *mujer* es una variable dicotómica que toma el valor 1 para las mujeres y el valor 0 para los hombres.

Demuestre que aplicando las fórmulas de *MCO* para la regresión simple se obtiene que

$$\hat{\beta}_1 = \overline{salario_H}$$

$$\hat{\delta}_1 = \overline{salario_M} - \overline{salario_H}$$

donde *M* indica mujer y *H* hombre.

Con el fin de facilitar la obtención de la solución, considere que en la muestra hay n_1 mujeres y n_2 hombres: la muestra total es $n = n_1 + n_2$.

Ejercicio 5.15 Los datos de este ejercicio se obtuvieron de un experimento de marketing controlado en las tiendas en París sobre el gasto en café, publicado por CA Bemmaor y Mouchoux D., "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment", *Journal of Marketing Research*, 28 (1991), 202-14. En este experimento se formuló el siguiente modelo para explicar la cantidad vendida de café por semana:

$$\ln(coffqty) = \beta_1 + \delta_1 advert + \beta_2 \ln(coffpric) + \delta_2 advert \times \ln(coffpric) + u$$

donde *coffpric* toma tres valores: 1, que es el precio habitual, 0.95 y 0.85; *advert* es una variable dicotómica que toma valor 1 si se hace publicidad en esa semana, y 0 si no se hace. El experimento duró 18 semanas. El modelo original y otros tres modelos más fueron estimados, utilizando el fichero *coffee2*:

1) $$\underset{}{\ln(coffqty_i)} = \underset{(0.04)}{5.85} + \underset{(0.099)}{0.2565}\, advert_i - \underset{(0.450)}{3.9760}\ln(coffpric_i) - \underset{(0.883)}{1.069}\, advert_i \times \ln(coffpric_i)$$

$$R^2 = 0.9468 \qquad n = 18$$

2) $$\underset{}{\ln(coffqty_i)} = \underset{(0.04)}{5.83} + \underset{(0.057)}{0.3559}\, advert_i - \underset{(0.393)}{4.2539}\ln(coffpric_i)$$

$$R^2 = 0.9412 \quad n = 18$$

3) $$\underset{}{\ln(coffqty_i)} = \underset{(0.04)}{5.88} - \underset{(0.513)}{3.6939}\ln(coffpric_i) - \underset{(0.582)}{2.9575}\, advert_i \times \ln(coffpric_i)$$

$$R^2 = 0.9214 \quad n = 18$$

4) $$\underset{}{\ln(coffqty_i)} = \underset{(0.07)}{5.89} - \underset{(0.674)}{5.1727}\ln(coffpric)$$

$$R^2 = 0.7863 \quad n = 18$$

a) En el modelo (2), ¿cuál es la interpretación del coeficiente de *advert*?
b) En el modelo (3), ¿cuál es la interpretación del coeficiente de *advert*×ln *(coffpric)*?
c) En el modelo (2), ¿tiene el coeficiente de *advert* un efecto positivo significativo al 5% y al 1%?
d) ¿Es el modelo (4) válido para semanas con publicidad y para semanas sin publicidad?
e) En el modelo (1), ¿es el término independiente el mismo para semanas con publicidad y para semanas sin publicidad?
f) En el modelo (3), ¿es la elasticidad de demanda de café/precio diferente en semanas con publicidad y en semanas sin publicidad?
g) En el modelo (4), ¿es la elasticidad de demanda de café/precio inferior a -4?

Ejercicio 5.16 (Continuación del ejercicio 4.39). Utilizando el fichero *timuse03*, se han estimado los siguientes modelos:

$$\underset{}{\widehat{housework}_i} = \underset{(23)}{132} + \underset{(1.497)}{2.787}\, educ_i + \underset{(0.308)}{1.847}\, age_i - \underset{(0.023)}{0.2337}\, paidwork_i \qquad (1)$$
$$R^2 = 0.142 \quad n = 1000$$

$$\underset{}{\widehat{housework}_i} = -\underset{(22.29)}{3.02} + \underset{(1.356)}{3.641} educ_i + \underset{(0.279)}{1.775}\, age_i - \underset{(0.021)}{0.1568}\, paidwork_i + \underset{(2.16)}{32.11}\, female_i \qquad (2)$$
$$R^2 = 0.298 \quad n = 1000$$

$$\underset{}{\widehat{housework}_i} = -\underset{(35.18)}{8.04} + \underset{(2.352)}{4.847}\, educ_i + \underset{(0.502)}{1.333}\, age_i - \underset{(0.032)}{0.0871}\, paidwork_i + \underset{(8.15)}{32.75}\, female_i$$
$$-\underset{(0.546)}{0.1650}\, educ_i \times female_i + \underset{(0.112)}{0.1019}\, age_i \times female_i - \underset{(0.009)}{0.02625}\, paidwork_i \times female_i \qquad (3)$$
$$R^2 = 0.306 \quad n = 1000$$

a) En el modelo (1), ¿existe una compensación estadísticamente significativa entre el tiempo dedicado a trabajo remunerado y el tiempo dedicado a trabajo doméstico?
b) Manteniendo igual todos los demás factores y tomando como modelo de referencia al (2), ¿existe evidencia de que las mujeres dedican más tiempo al trabajo doméstico que los hombres?
c) Compare el R^2 de los modelos (1) y (2). ¿Cuál es su conclusión?
d) En el modelo (3), ¿cuál es el efecto marginal del tiempo dedicado al trabajo doméstico con respecto al tiempo dedicado al trabajo remunerado?
e) ¿Es significativa la interacción entre *paidwork* y género?
f) ¿Son las interacciones entre género y las variables cuantitativas del modelo conjuntamente significativas?

Ejercicio 5.17 Utilizando datos de la Bolsa de Madrid del 19 de noviembre de 2011 (fichero *bolmad11*), se han estimado los siguientes modelos*:*

$$\underset{}{\widehat{\ln(marktval_i)}} = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.6998}\, ibex35_i + \underset{(0.0369)}{0.6749} \ln(bookval_i) \qquad (1)$$
$$SCR=35.69 \quad R^2=0.8931 \quad n=92$$

$$\ln(marktval_i) = \underset{(0.275)}{1.828} + \underset{(0.778)}{0.4236}\, ibex35_i + \underset{(0.0423)}{0.6678}\ln(bookval_i) + \underset{(0.088)}{0.0310}\, ibex35_i \times \ln(bookval_i) \tag{2}$$

$$SCR=35.622 \quad R^2=0.8933 \quad n=92$$

$$\ln(marktval_i) = \underset{(0.310)}{2.323} + \underset{(0.785)}{0.1987}\, ibex35_i + \underset{(0.0405)}{0.6688}\ln(bookval_i) + \underset{(0.089)}{0.0369}\, ibex35_i \times \ln(bookval_i) - \underset{(0.236)}{0.6613}\, services_i - \underset{(0.221)}{0.6698}\, consump_i - \underset{(0.263)}{0.1931}\, energy_i - \underset{(0.207)}{0.3895}\, industry_i - \underset{(0.324)}{0.7020}\, itt_i \tag{3}$$

$$SCR=30.781 \quad R^2=0.9078 \quad n=92$$

$$\ln(marktval_i) = \underset{(0.234)}{1.366} + \underset{(0.0305)}{0.7658}\ln(bookval_i) \tag{4}$$

$$SCR=41.625 \quad R^2=0.8753 \quad n=92$$

Para *finance*=1 $$\ln(markval_i) = \underset{(0.560)}{0.558} + \underset{(0.0702)}{0.9346}\ln(bookval_i) \tag{5}$$

$$SCR=2.7241 \quad R^2=0.9415 \quad n=13$$

donde

- *marktval* es el valor de mercado de una compañía.
- *bookval* es el valor contable de una compañia.
- *ibex35* es una variable ficticia que toma el valor 1 si la compañía está incluida en el selectivo Ibex 35.
- *services*, *consump* (*consumo*), *energy*, *industry* e *itc* (tecnologías de la información y la comunicación) son variables ficticias. Cada uno de ellas toma el valor 1 si la compañía está clasificada en ese sector en la Bolsa de Madrid. La categoría de referencia es el sector financiero (*finance*).

a) En el modelo (1), ¿cuál es la interpretación del coeficiente de *ibex35*?
b) En el modelo (1), ¿es la elasticidad *marktval/bookval* igual a 1?
c) En el modelo (2), ¿es la elasticidad *marktval/bookval* la misma para todas las compañías incluidas en la muestra?
d) ¿Es el modelo (4) válido tanto para las compañías incluidas en el Ibex 35 y para las compañías excluidas?
e) En el modelo (3), ¿cuál es la interpretación del coeficiente de *consump*?
f) ¿Es el coeficiente de *consump* significativamente negativo?
g) ¿Está justificada estadísticamente la introducción de variables ficticias para los diferentes sectores?
h) ¿Es la elasticidad *marktval/bookval* para el sector financiero igual a 1?

Ejercicio 5.18 (Continuación del ejercicio 4.37). Utilizando el fichero *rdspain*, se han estimado las ecuaciones que aparecen en el cuadro adjunto

Las variables que aparecen en el cuadro son las siguientes:

- *rdintens* es el gasto en investigación y desarrollo (I+D) medido como porcentaje de las ventas,
- *sales*, ventas medidas en millones de euros,
- *exponsal* son las exportaciones medidas como porcentaje de las ventas,

- *medtech* e *hightech* son dos variables ficticias que reflejan si la empresa pertenece a un sector de media o de alta tecnología. La categoría de referencia corresponde a las empresas de baja tecnología.
- *workers* es el número de trabajadores de la empresa.

	(1) *rdintens*	(2) *rdintens*	(3) *rdintens*	(4) *rdintens* para *hightech*=1	(5) *rdintens* para *medtech*=1	(6) *rdintens* para *lowtech*=1
exponsal	0.0136	0.0101	0.00968	0.00584	0.0116	0.00977
	(0.00195)	(0.00193)	(0.00189)	(0.00792)	(0.00300)	(0.00169)
workers	0.000433	0.000392	0.000394	0.00196	0.0000563	0.000393
	(0.0000740)	(0.0000725)	(0.000208)	(0.000338)	(0.0000815)	(0.000121)
hightech		1.448	0.976			
		(0.141)	(0.151)			
medtech		0.361	0.472			
		(0.109)	(0.112)			
hightech× *workers*			0.00153			
			(0.000271)			
medtech× *workers*			-0.000326			
			(0.000222)			
término independiente	0.394	0.137	0.143	1.211	0.577	0.142
	(0.0598)	(0.0691)	(0.0722)	(0.313)	(0.103)	(0.0443)
n	1983	1983	1983	296	616	1071
R^2	0.0507	0.0986	0.138	0.113	0.0278	0.0459
SCR	9282.7	8815.0	8425.3	4409.0	2483.6	1527.5
F	52.90	54.06	52.90	18.71	8.776	25.72
df_n	2	4	6	2	2	2
df_d	1980	1978	1976	293	613	1068

Errores estándar entre paréntesis

a) En el modelo (2), manteniéndose igual todos los demás factores, ¿hay evidencia de que el gasto en investigación y desarrollo (expresado como un porcentaje de las ventas) en empresas de alta tecnología sea mayor que en empresas de baja tecnología? ¿Es fuerte la evidencia empírica?

b) En el modelo (2), manteniéndose igual todos los demás factores, ¿hay evidencia de que el gasto en I+D, *rdintens*, en las empresas de tecnología media sea igual al de empresas de baja tecnología? ¿Es fuerte la evidencia empírica?

c) Tomando como modelo de referencia (2), si tuviera que contrastar la hipótesis de que *rdintens* en las empresas de alta tecnología es igual a las empresas de tecnología media, formule un modelo que le permita contrastar esta hipótesis sin necesidad de utilizar la información sobre la matriz de covarianzas de los estimadores.

d) ¿Hay influencia de los trabajadores asociados en *rdintens* con el nivel de tecnología en las empresas?

e) ¿Es el modelo (1) válido para todas las empresas independientemente de su nivel tecnológico?

Ejercicio 5.19 Para explicar la satisfacción general de las personas (*stsfglo*) se estimaron los siguientes modelos utilizando datos del fichero *hdr2010*:

$$stsfglo_i = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i \qquad (1)$$
$$R^2 = 0.642 \quad n = 144$$

$$stsfglo_i = \underset{(0.897)}{2.911} + \underset{(0.00000572)}{0.0000381}\, gnipc_i + \underset{(0.18)}{1.215}\, lifexpec_i$$
$$+ \underset{(0.179)}{1.215}\, dlatam_i - \underset{(0.259)}{0.7901}\, dafrica_i \qquad (2)$$
$$R^2 = 0.748 \quad n = 144$$

$$stsfglo_i = \underset{(1.146)}{0.6984} + \underset{(0.000006)}{0.0000198}\, gnipc_i + \underset{(0.0164)}{0.0724}\, lifexpec_i + \underset{(1.950)}{4.099}\, dafrica_i$$
$$+ \underset{(0.000052)}{0.0000801}\, gnipc_i \times dafrica_i - \underset{(0.0336)}{0.0896}\, lifexpec_i \times dafrica_i \qquad (3)$$
$$R^2 = 0.6840 \quad n = 144$$

donde

- *gnipc* es el producto nacional bruto per cápita expresado en *PPA* (paridad de poder adquisitivo) en dólares americanos de 2008,
- *lifexpec* es la esperanza de vida al nacer, es decir, el número de años que un recién nacido puede esperar vivir,
- *dafrica* es una variable dicotómica que toma el valor 1 si el país se encuentra en África,
- *dlatam* es una variable dicotómica que toma el valor 1 si el país está en América Latina.

a) En el modelo (2), ¿cuál es la interpretación de los coeficientes de *dlatam* y *dafrica*?
b) En el modelo (2), *dlatam* y *dafrica*, individualmente, ¿tienen una influencia significativamente positiva sobre la satisfacción global?
c) En el modelo (2), *dlatam* y *dafrica* ¿tienen una influencia conjunta sobre la satisfacción global?
d) ¿Es la influencia de la esperanza de vida sobre la satisfacción global menor en África que en otras regiones del mundo?
e) ¿Es la influencia de la variable de *gnipc* mayor en África que en otras regiones del mundo en un 10%?
f) ¿Son las interacciones de las personas que viven en África y las variables *gnipc* y *lifexpec* conjuntamente significativas?

Ejercicio 5.20 Las ecuaciones que aparecen en el cuadro adjunto se han estimado utilizando los datos del fichero *timuse03*. Este archivo contiene 1000 observaciones correspondientes a una submuestra aleatoria extraída de la encuesta de uso del tiempo en España que se llevó a cabo en el periodo 2002-2003.

Las variables que aparecen en el cuadro son:

- *educ* son los años de educación alcanzados,
- *sleep* (dormir), *paidwork* (trabajo remunerado) and *unpaidwrk* (trabajo no remunerado se miden en minutos por día,

- *female* (mujer), *workday* (*lunes a viernes*), *spaniard* (español) y *houswife* (ama de casa) son variables ficticias.
  - *a)* En el modelo (1), ¿existe una compensación estadísticamente significativa entre el tiempo dedicado al trabajo remunerado y el tiempo dedicado a dormir?
  - *b)* En el modelo (1), ¿es el coeficiente de *unpaidwk* estadísticamente significativo?
  - *c)* ¿Existe evidencia de que las mujeres duermen más que los hombres?
  - *d)* En el modelo (2), ¿son *workday* y *spaniard* individualmente significativas? ¿Son conjuntamente significativas?
  - *e)* ¿Es el coeficiente de *housewife* estadísticamente significativo?
  - *f)* ¿Son las interacciones de *female* con *educ*, *paidwork* y *unpaidwk* conjuntamente significativas?

	(1) *Sleep*	*(2)* *Sleep*	*(3)* *Sleep*	*(4)* *Sleep*	*(5)* *Sleep*	*(6)* *sleep*
educ	-4.669	-4.787	-4.805	-4.754	-4.782	-4.792
	(0.916)	(0.912)	(0.912)	(0.913)	(0.917)	(0.917)
persinc	0.0238	0.0207	0.0195	0.0210	0.0208	0.0208
	(0.00587)	(0.00600)	(0.00607)	(0.00601)	(0.00601)	(0.00601)
age	0.854	0.879	0.895	0.884	0.879	0.891
	(0.174)	(0.174)	(0.174)	(0.174)	(0.174)	(0.302)
paidwork	-0.258	-0.247	-0.246	-0.248	-0.246	-0.247
	(0.0150)	(0.0159)	(0.0159)	(0.0160)	(0.0210)	(0.0159)
unpaidwk	-0.205	-0.198	-0.188	-0.224	-0.198	-0.198
	(0.0184)	(0.0184)	(0.0196)	(0.0365)	(0.0185)	(0.0184)
female	4.161	3.588	3.981	2.485	3.638	3.727
	(1.465)	(1.467)	(1.493)	(1.975)	(1.691)	(3.287)
workday		-19.31	-19.46	-19.47	-19.30	-19.30
		(7.168)	(7.165)	(7.171)	(7.173)	(7.172)
spaniard		-47.50	-46.88	-47.90	-47.63	-47.51
		(19.99)	(19.98)	(20.00)	(20.10)	(20.00)
houswife			-14.71			
			(10.42)			
unpaidwk ×*female*				0.00607		
				(0.00726)		
paidwork ***×female***					-0.000324	
					(0.00540)	
age× female						-0.00308
						(0.0652)
término independiente	588.9	648.3	646.6	651.9	648.2	647.8
	(13.62)	(24.34)	(24.36)	(24.73)	(24.39)	(26.40)
N	1000	1000	1000	1000	1000	1000
R^2	0.316	0.325	0.326	0.325	0.325	0.325
SCR	9913901.3	9789312.3	9769648.2	9782424.0	9789276.9	9789290.3
F	76.58	59.62	53.27	53.06	52.95	52.95
df_n	6	8	9	9	9	9
df_d	993	991	990	990	990	990

Errores estándar entre paréntesis

Ejercicio 5.21 Para el estudio de la mortalidad infantil en el mundo se han estimado los siguientes modelos a partir de los datos del fichero *hdr2010*:

$$\underset{}{deathinf_i} = \underset{(4.58)}{93.02} - \underset{(0.0002)}{0.00037}\, gnipc_i - \underset{(0.1866)}{0.6046}\, physicn_i - \underset{(0.003)}{0.003}\, contrcep_i \qquad (1)$$

$$SCR=40285 \quad R^2=0.6598 \quad n=108$$

$$\begin{gathered} \underset{}{deathinf_i} = \underset{(5.96)}{78.55} - \underset{(0.0002)}{0.00042}\, gnipc - \underset{(0.1879)}{0.3809}\, physicn_i - \underset{(0.1042)}{0.6989}\, contrcep_i \\ + \underset{(5.05)}{17.92}\, dafrica \\ SCR=35893 \quad R^2=0.6851 \quad n=108 \end{gathered} \tag{2}$$

$$\begin{gathered} \underset{}{deathinf_i} = \underset{(6.76)}{72.58} - \underset{(0.0002)}{0.00044}\, gnipc - \underset{(0.1879)}{0.3994}\, physicn_i - \underset{(0.1234)}{0.5857}\, contrcep_i \\ + \underset{(5.05)}{17.92}\, dafrica - \underset{(0.000826)}{0.0000914}\, gnipc \times dafrica - \underset{(2.2351)}{2.0013}\, physicn \times dafrica \\ - \underset{(0.2716)}{0.2172}\, contrcep_i \times dafrica \\ SCR=34309 \quad R^2=0.7109 \quad n=108 \end{gathered} \tag{3}$$

donde

- *deatinf* es el número de muertes infantiles (de un año o menos) por cada 1000 nacidos vivos en 2008,
- *gnipc* es el producto nacional bruto per cápita expresado en *PPA* en dólares americanos de 2008,
- *physicn* son los médicos por cada 10000 habitantes en el periodo 2000-2009,
- *contrcep* es la tasa de uso de anticonceptivos de cualquier tipo, expresada como % de las mujeres casadas de 15-49 años para el periodo 1990-2008,
- *dafrica* es una variable dicotómica que toma el valor 1 si el país se encuentra en África.

a) En el modelo (1), ¿cuál es la interpretación de los coeficientes de *gnipc*, *physicn* y *contrcep*?
b) En el modelo (2), ¿cuál es la interpretación del coeficiente de *dafrica*?
c) En el modelo (2), manteniendo igual todos los demás factores, ¿tienen los países de África una mortalidad infantil mayor que los países de otras regiones del mundo?
d) ¿Cuál es el efecto marginal de la variable *gnipc* sobre la mortalidad infantil en el modelo (3)?
e) ¿Es la pendiente correspondiente al regresor *contrcep* significativamente mayor para los países de África?
f) ¿Son las pendientes correspondientes a los regresores *gnipc*, *physicn* y *contrcep* conjuntamente diferentes para los países de África?
g) ¿Es el modelo (1) válido para todos los países del mundo?

Ejercicio 5.22 Utilizando una submuestra aleatoria de 2000 observaciones extraídas de las encuestas de uso del tiempo para España llevadas a cabo en los periodos 2002-2003 y 2009-2010 (fichero *timus309*), se han estimado los siguientes modelos para explicar el tiempo que se pasa viendo la televisión:

$$\begin{gathered} watchtv = \underset{(9.46)}{114} - \underset{(0.620)}{3.523}\, educ + \underset{(0.129)}{1.330}\, age - \underset{(0.010)}{0.1111}\, paidwork \\ R^2 = 0.169 \quad n = 2000 \end{gathered} \tag{1}$$

$$\begin{gathered} watchtv = \underset{(9.92)}{127} - \underset{(0.615)}{3.653}\, educ + \underset{(0.129)}{1.291}\, age - \underset{(0.010)}{0.120}\, paidwork - \underset{(4.903)}{25.15}\, female \\ + \underset{(5.25)}{17.14}\, y2009 \qquad R^2 = 0.184 \quad n = 2000 \end{gathered} \tag{2}$$

$$watchtv = \underset{(10.01)}{123} - \underset{(0.615)}{3.583}\,educ + \underset{(0.129)}{1.302}\,age - \underset{(0.012)}{0.1053}\,paidwork - \underset{(4.90)}{24.87}\,female$$
$$+ \underset{(6.115)}{24.54}\,y2009 - \underset{(0.021)}{0.0501}\,y2009 \times paidwork \qquad R^2 = 0.186 \quad n = 2000 \tag{3}$$

donde

- *educ* son los años de educación alcanzados,
- *watchtv* (ver televisión) y *paidwork* (trabajo remunerado) se miden en minutos por día,
- *female* (mujer) es una variable ficticia que toma valor 1 si el entrevistado es una mujer
- *y2009* es una variable ficticia que toma valor 1 si la encuesta se llevó a cabo en el bienio 2008-2009.

a) En el modelo (1), ¿cuál es la interpretación del coeficiente de *educ*?
b) En el modelo (1), ¿hay una compensación estadísticamente significativa entre el tiempo dedicado al trabajo y el tiempo dedicado a ver la televisión?
c) Manteniendo igual todos los demás factores y tomando como modelo (2) como referencia, ¿existe evidencia de que los hombres ven la televisión más que las mujeres? ¿Es fuerte esa posible evidencia?
d) En el modelo (2), ¿cuál es la diferencia estimada del tiempo dedicado a ver la televisión entre las mujeres encuestadas en 2008-2009 y los hombres encuestados en el periodo 2002-2003? ¿Es esta diferencia estadísticamente significativa?
e) En el modelo (3), ¿cuál es el efecto marginal del tiempo dedicado al trabajo remunerado sobre el tiempo dedicado a ver televisión?
f) ¿Existe una interacción significativa entre el año de la encuesta y el tiempo dedicado al trabajo remunerado?

Ejercicio 5.23 Utilizando el fichero *consumsp*, se estimaron los siguientes modelos para analizar si la entrada de España en la Comunidad Europea en 1986 tuvo algún impacto en el comportamiento de los consumidores españoles:

$$conspc_t = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\,incpc_t + \underset{(0.0903)}{0.5771}\,conspc_{t-1}$$
$$R^2=0.9967 \qquad SCR=1891320 \qquad n=56 \tag{1}$$

$$conspc_t = \underset{(108)}{-102.4} + \underset{(0.0879)}{0.3573}\,incpc_t + \underset{(0.0901)}{0.5992}\,conspc_{t-1} + \underset{(92.56)}{148.60}\,y1986_t$$
$$R^2=0.9968 \qquad SCR=1802007 \qquad n=56 \tag{2}$$

$$conspc_t = \underset{(114)}{78.18} + \underset{(0.1100)}{0.5181}\,incpc_t + \underset{(0.1199)}{0.4186}\,conspc_{t-1} + \underset{(456.3)}{819.82}\,y1986_t$$
$$- \underset{(0.2338)}{0.5403}\,incpc_t \times y1986_t + \underset{(0.2182)}{0.5424}\,conspc_{t-1} \times y1986_t$$
$$R^2=0.9972 \qquad SCR=1600714 \qquad n=56 \tag{3}$$

$$conspc_t = \underset{(118)}{117.03} + \underset{(0.0968)}{0.3697}\,incpc_t + \underset{(0.1051)}{0.5823}\,conspc_{t-1} + \underset{(348)}{41.62}\,y1986_t$$
$$+ \underset{(0.0326)}{0.0104}\,incpc_t \times y1986_t$$
$$R^2=0.9968 \qquad SCR=1798423 \qquad n=56 \tag{4}$$

$$\underset{}{conspc_t} = \underset{(114)}{120.1} + \underset{(0.0854)}{0.3750}\, incpc_t + \underset{(0.0890)}{0.5758}\, conspc_{t-1} + \underset{(0.0087)}{0.0141}\, incpc_t \times y1986_t \qquad (5)$$

$$R^2=0.9968 \quad SCR=1798927 \quad n=56$$

donde el consumo (*conspc*) y la renta disponible (*incpc*) se expresan en euros constantes per cápita, tomando 2008 como año de referencia.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Compruebe en el modelo (6) si la propensión marginal al consumo a corto plazo se redujo en 1986 y años sucesivos.

b) ¿Son las interacciones de *y1986* con las variables cuantitativas del modelo significativas en forma conjunta?

c) Estime si hubo un cambio estructural en la función de consumo en 1986 y siguientes años.

d) Comprobar si el coeficiente de $conspc_{t-1}$ cambió en 1986.

e) ¿Existe una brecha entre el consumo que se realizaba antes de 1986 con respecto al consumo en 1986 y años sucesivos?

6 RELAJACIÓN DE LOS SUPUESTOS EN EL MODELO LINEAL CLÁSICO

6.1 Relajación de los supuestos del *MLC*: una panorámica

En los capítulos 2 y 3 se formuló el modelo de regresión lineal, simple y múltiple, incluyendo el conjunto de supuestos estadísticos denominados supuestos del modelo lineal clásico (*MLC*). Ahora, vamos a examinar los problemas que plantea el incumplimiento de cada uno de los supuestos del *MLC*, así como los métodos alternativos que se plantean para estimar el modelo lineal.

Supuestos sobre la forma funcional

En el supuesto 1 se postula cuál es el modelo poblacional:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{6-1}$$

Este supuesto especifica cuál es la variable endógena y la forma funcional con que aparece en la ecuación, cuáles son las variables explicativas y sus respectivas formas funcionales. Además, se establece que el modelo es lineal en los parámetros.

Cuando se estima un modelo poblacional diferente se comete un error de especificación. Las consecuencias de este tipo de errores se examinan en el epígrafe 6.2.

Supuestos sobre los regresores

Sobre los regresores se formularon los supuestos 2, 3, y 4. En el modelo de regresión lineal múltiple, en el supuesto 2 se postulaba que los valores de $x_2, x_3, \cdots, x_k$ son fijos en repetidas muestras, es decir, los regresores son no estocásticos. Ésta es un supuesto razonable cuando los regresores se obtienen a partir de variables controladas experimentalmente. En cambio, es menos admisible en variables obtenidas mediante observación de carácter pasivo, como sería el caso de la renta en la función del consumo.

Cuando los regresores son estocásticos, la relación estadística entre los regresores y la perturbación aleatoria es un punto crucial en la elaboración de un modelo econométrico. Por ello se formuló el supuesto alternativo 2*: los regresores $x_2, x_3, \cdots, x_k$ se distribuyen independientemente de la perturbación aleatoria. Cuando asumimos este supuesto alternativo, la inferencia, *condicionada* a la matriz de los regresores, lleva a unos resultados que son prácticamente coincidentes con el caso en que la matriz **X** es fija. En otras palabras, en el caso de independencia entre los regresores y la perturbación aleatoria, el método de mínimos cuadrados ordinarios sigue siendo el método óptimo para la estimación del vector de coeficientes.

En el supuesto 3 se postulaba que la matriz de regresores **X** no contiene errores de medida. En el caso de que los tuviera se plantea un problema econométrico muy grave, cuya solución es compleja.

El supuesto 4 establece que no existe relación lineal exacta entre los regresores, o, en otras palabras, establece que no existe multicolinealidad perfecta en el modelo. Este supuesto es necesario para el cálculo del vector de estimadores mínimos cuadráticos. La multicolinealidad perfecta no se suele presentar en la práctica. En cambio, sí es frecuente que entre los regresores exista una relación aproximadamente lineal, en cuyo caso los estimadores que se obtengan serán en general poco precisos, aunque siguen conservando la propiedad de ser estimadores *ELIO*. En otras palabras, la relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre el regresando, lo que determina que las varianzas de los estimadores sean elevadas. Cuando se presenta una relación aproximadamente lineal entre los regresores, se dice que existe multicolinealidad no perfecta. El epígrafe 6.3 se dedica a examinar la detección de la multicolinealidad (no perfecta), así como algunas de las posibles soluciones.

Supuesto sobre los parámetros

En el supuesto 5 se asumió que los parámetros $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ son no aleatorios. El análisis del mundo real puede sugerir que esta constancia de los coeficientes no sea razonable. Así, en los modelos que utilizan datos de series temporales, puede quedar de manifiesto que a lo largo del tiempo se han producido cambios en los patrones de comportamiento, lo que implicaría naturalmente cambios en los coeficientes de regresión. Sobre esta cuestión, en el epígrafe 5.6 se ha examinado el contraste de cambio estructural que permite determinar si se ha producido algún cambio en los parámetros a lo largo del tiempo.

Supuestos sobre la perturbación aleatoria

En el supuesto 6 se asumió que $E(\mathbf{u})=\mathbf{0}$. Este supuesto no es contrastable empíricamente en el caso general de modelos con término independiente.

Antes de pasar a otros supuestos sobre la perturbación aleatoria u_i conviene remarcar que ésta es una variable no observable. La información sobre u_i la obtenemos indirectamente a través de los residuos, que son los que tendremos que utilizar para realizar contrastes acerca del comportamiento de las perturbaciones. Sin embargo, la utilización de los residuos para realizar contrastes sobre las perturbaciones plantea el siguiente problema. Cuando se cumplen los supuestos del *MLC*, las perturbaciones aleatorias son homoscedásticas y no autocorrelacionadas, pero en cambio los residuos son heteroscedásticos y están autocorrelacionados, bajo dichos supuestos. Esta circunstancia ha de tenerse en cuenta en el diseño de los contrastes estadísticos sobre los supuestos de homoscedasticidad y no autocorrelación.

Si no se cumplen los supuestos 7 de homoscedasticidad y/o 8 de no autocorrelación los estimadores obtenidos por mínimos cuadrados siguen siendo lineales, insesgados, pero no óptimos.

Los supuestos de homoscedasticidad y no autocorrelación formuladas en el tema 3 se pueden formular conjuntamente indicando que la matriz de covarianzas de las perturbaciones aleatorias es una matriz escalar, es decir,

$$E(\mathbf{uu}') = \sigma^2\mathbf{I} \tag{6-2}$$

Cuando no se cumple uno, o los dos, de los supuestos señalados, entonces la matriz de covarianzas será menos restrictiva. Así, consideraremos la siguiente matriz de covarianzas de las perturbaciones:

$$E(\mathbf{uu}') = \sigma^2\mathbf{\Omega} \tag{6-3}$$

donde la única restricción que se impone a $\mathbf{\Omega}$ es que sea una matriz definida positiva.

Cuando la matriz de covarianzas es una matriz no escalar, como (6-3), entonces pueden obtenerse unos estimadores lineales, insesgados y óptimos mediante la aplicación del método de mínimos cuadrados generalizados (*MCG*). La expresión de estos estimadores es la siguiente:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \tag{6-4}$$

En la práctica, no se suele aplicar directamente la formula (6-4). En su lugar se aplica un procedimiento en dos etapas, que conduce exactamente a los mismos resultados.

En epígrafe 6.5 se examinarán los contrastes para determinar si existe o no heteroscedasticidad, así como la particularización del método de *MCG* a este caso concreto. En el epígrafe 6.6 se expondrán procedimientos de contraste, así como el tratamiento de modelos con perturbaciones autocorrelacionadas.

El supuesto 9 de normalidad postulado en el *MLC* permite construir estadísticos para realizar inferencias con distribuciones conocidas. Si el supuesto de normalidad no es adecuado, entonces los contrastes solo tendrán una validez aproximada. En el epígrafe 6.4 se expone un contraste de normalidad de las perturbaciones que se utiliza para determinar si este supuesto es aceptable o no.

6.2 Errores de especificación

Como hemos indicado se produce un error de especificación cuando se estima un modelo diferente del modelo poblacional. El problema en las ciencias sociales, y en particular en economía, es que generalmente no conocemos el modelo poblacional.

Teniendo en cuenta esta observación, consideraremos tres tipos de errores de especificación:

- Inclusión de una variable irrelevante
- Exclusión de una variable relevante.
- Forma funcional incorrecta

6.2.1 Consecuencias de la especificación errónea

A continuación, examinaremos las consecuencias en los estimadores *MCO* de cada tipo de especificación errónea.

Inclusión de una variable irrelevante

Supongamos que el modelo poblacional es el siguiente:

$$y = \beta_1 + \beta_2 x_2 + u \tag{6-5}$$

Por lo tanto, la *función de regresión poblacional* (*FRP*) – parte sistemática de este modelo- viene dada por

$$\mu_y = \beta_1 + \beta_2 x_2 \tag{6-6}$$

Ahora supongamos que la *función de regresión muestral* (*FRP*) estimada es la siguiente:

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} + \tilde{\beta}_3 x_{3i} \tag{6-7}$$

Este es el caso de inclusión de una variable irrelevante: específicamente en (6-7) hemos introducido la variable irrelevante x_3. ¿Cuál son los efectos de la inclusión de una variable irrelevante en los estimadores obtenidos por *MCO*?

Puede demostrarse que los estimadores correspondientes a (6-7) son insesgados, es decir,

$$E(\tilde{\beta}_1) = \beta_1 \qquad E(\tilde{\beta}_2) = \beta_2 \qquad E(\tilde{\beta}_3) = 0$$

Sin embargo, las varianzas de estos estimadores serán más grandes que las obtenidas al estimar (6-5) donde se ha omitido (correctamente) x_3.

Este resultado es generalizable: si incluimos una o más variables irrelevantes, entonces los estimadores *MCO* son insesgados, pero con varianzas más grandes que cuando no se incluyen variables irrelevantes en el modelo estimado.

Exclusión de una variable relevante

Supongamos que el modelo poblacional es el siguiente:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \tag{6-8}$$

Entonces la *FRP* viene dada por

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6-9}$$

Ahora supongamos que la *FRM* estimada, debido a nuestra ignorancia o a la no disponibilidad de datos, es la siguiente

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \tag{6-10}$$

Éste es un caso de *exclusión de una variable relevante*: específicamente en (6-10) hemos omitido la variable relevante x_3. ¿Es $\tilde{\beta}_2$, obtenido mediante aplicación de *MCO* a (6-10), un estimador insesgado de β_2?

Como se muestra en el apéndice 6.1 el estimador $\tilde{\beta}_2$ es sesgado. El sesgo es

$$Bias(\tilde{\beta}_2) = \beta_3 \frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2) x_{3i}}{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} \tag{6-11}$$

Este sesgo es nulo si, de acuerdo con (6-11), la covarianza entre x_2 y x_3 es 0. Es importante advertir que la *ratio*

$$\frac{\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2}$$

es justamente la pendiente ($\hat{\delta}_2$) en la regresión de x_3 sobre x_2. Es decir,

$$\hat{x}_2=\hat{\delta}_1+\hat{\delta}_2\hat{x}_2=\hat{\delta}_1+\frac{\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2}\hat{x}_2 \tag{6-12}$$

Así pues, de acuerdo con (6-72) - en el apéndice 6.1- y (6-12), podemos decir que

$$E(\tilde{\beta}_2)=\beta_2+\beta_3\hat{\delta}_2 \tag{6-13}$$

En consecuencia, el sesgo es igual a $\beta_3\hat{\delta}_2$. En el cuadro 6.1 puede verse un resumen del signo del sesgo en $\tilde{\beta}_2$ cuando se omite x_2 en la ecuación estimada. Para la mejor comprensión del contenido de este cuadro debe tenerse en cuenta que el signo de $\hat{\delta}_2$ tiene el mismo signo que la correlación muestral entre x_2 y x_3.

CUADRO 6.1. Resumen del sesgo en $\tilde{\beta}_2$ cuando se omite x_2 en le ecuación estimada.

	Corr(x_2,x_3)>0	*Corr*(x_2,x_3)<0
β_3>0	Sesgo positivo	Sesgo negativo
β_3<0	Sesgo negativo	Sesgo positivo

Forma funcional incorrecta

Si utilizamos una forma funcional diferente del modelo poblacional verdadero, entonces los estimadores *MCO* estarán sesgados.

En resumen, si hay exclusión de variables relevantes y/o se ha utilizado una forma funcional incorrecta, entonces los estimadores *MCO* estarán sesgados y además serán también inconsistentes. En consecuencia los procedimientos convencionales de inferencia quedarán invalidados en estos dos casos.

6.2.2 Contrastes de especificación: el contraste RESET

Para contratar si se han incluido en el modelo variables irrelevantes, se pueden aplicar los contrates de exclusión examinados en el capítulo 4.

Para contrastar la exclusión de variables relevantes o la utilización de una forma funcional incorrecta, puede aplicarse el contraste RESET (Regression Equation Specification Error Test). Este contraste es un contraste general para errores de especificación propuesto por Ramsey (1969). Para explicarlo, consideraremos que el modelo *inicial* es el siguiente:

$$y=\beta_1+\beta_2x_2+\beta_3x_3+u \tag{6-14}$$

Ahora, vamos a introducir un modelo *aumentado* en el cual aparecen dos nuevas variables (z_1 y z_2):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 z_1 + \alpha_2 z_2 + u \tag{6-15}$$

Teniendo en cuenta la especificación de los dos modelos, las hipótesis nula y alternativa serán las siguientes:

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = 0 \\ &H_1 : H_0 \text{ no es cierta} \end{aligned} \tag{6-16}$$

La cuestión clave para construir este contraste es determinar las variables o regresores *z* que se deben introducir. En el caso de exclusión de variables relevantes, las variables *z* serán los regresores omitidos o también cuadrados o potencias de nuevos regresores. El contrate a aplicar sería similar a los contrastes de exclusión, pero con los papeles invertidos: el modelo restringido es ahora el modelo *inicial*, mientras que el modelo no restringido se corresponde con el modelo *aumentado*.

En el contraste para formas funcionales incorrectas, consideremos, por ejemplo, que se ha especificado (6-14) en lugar de la verdadera relación:

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + u \tag{6-17}$$

En el modelo (6-17) existe una relación multiplicativa entre los regresores. Ramsey, tuvo en cuenta que una aproximación por series de Taylor de una relación multiplicativa daría lugar a una expresión que incluiría potencias y productos cruzados de las variables explicativas. Por esta razón, este autor sugiere la inclusión, en el modelo aumentado, de potencias de los valores predichos de la variable independiente (que son, por supuesto, combinaciones de potencias y productos cruzados de las variables explicativas):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 \hat{y}^2 + \alpha_2 \hat{y}^3 + u \tag{6-18}$$

donde las $\hat{y}$ son los valores ajustados por *MCO* correspondientes al modelo (6-14). Los superíndices indican las potencias a las que estos valores predichos están elevados. No se incluye la primera potencia porque sería perfectamente colineal con el resto de los regresores del modelo inicial l.

Los pasos implicados en el contraste RESET son los siguientes:

Paso 1. Se estima el modelo *inicial* y se calculan los valores ajustados, $\hat{y}_i$.

Paso 2. Se estima el modelo *aumentado* (6-18), el cual puede incluir una o más potencias de $\hat{y}_i$.

Paso 3. Tomando el R^2_{inic} correspondiente al modelo inicial y el R^2_{aum} correspondiente al modelo aumentado, se calcula el estadístico *F*:

$$F = \frac{(R^2_{aum} - R^2_{inic}) / r}{(1 - R^2_{aum}) / (n - h)} \tag{6-19}$$

donde *r* es el número de nuevos parámetros que se han añadido al modelo inicial, y *h* es el número de parámetros del modelo aumentado, incluido el término independiente.

Bajo la hipótesis nula, este estadístico se distribuye como sigue:

$$F \mid H_0 \sim F_{r,n-h} \tag{6-20}$$

Paso 4. Para un nivel de significación α, y designando por $F_{r,n-h}^{\alpha}$ el correspondiente valor en la tabla de la *F*, la decisión a tomar es la siguiente

Si $F \geq F_{r,n-h}^{\alpha}$ se rechaza H_0

Si $F < F_{r,n-h}^{\alpha}$ no se rechaza H_0

En consecuencia, valores elevados de este estadístico conducirán a rechazar el modelo inicial.

En el contraste RESET se contrasta una hipótesis nula contra una hipótesis alternativa que no indica cuál debería ser la especificación correcta del modelo. Así pues este contraste es un contraste de especificación que puede indicar que existe algún tipo de especificación errónea pero sin dar ninguna pista de cuál es la especificación correcta.

EJEMPLO 6.1 Especificación errónea en un modelo de determinación de los salarios

Utilizando una de la Encuesta de Estructura Salarial para España en 2006 (archivo *wage06sp*) se estimó el siguiente modelo para explicar los salarios:

$$\underset{}{wage_i} = \underset{(1.55)}{4.679} + \underset{(0.146)}{0.681}educ_i + \underset{(0.071)}{0.293}tenure_i$$

$$R^2=0.249 \quad n=150$$

donde educación (*educ*) y antigüedad en la empresa (*tenure*) están medidos en años y el salario (*wage*) en euros por hora.

Considerando que podía haber un problema de forma funcional incorrecta, se estimó un modelo aumentado. En este modelo aumentado – además de *educ*, *tenure*, y el término independiente - $wage_i^{\ 2}$ y $wage_i^{\ 3}$, obtenidos a partir de la estimación del modelo inicial, fueron incluidos como regresores. El estadístico *F* calculado utilizando y R_{inic}^2 y R_{augm}^2, de acuerdo a (6-18), es igual a 4.18. Dado que $F_{2,145}^{0.05} \simeq F_{2,60}^{0.05} = 3.15$, se rechaza, para los niveles α=0.05 y α=0.10, que la forma lineal sea la adecuada para explicar la determinación de los salarios. Por el contrario, dado que $F_{2,145}^{0.01} \simeq F_{2,60}^{0.01} = 4.98$, la H_0 no se rechaza para α=0.01.

6.3 Multicolinealidad

6.3.1 Planteamiento

La multicolinealidad perfecta no se suele presentar en la práctica, salvo que se diseñe mal el modelo como veremos en el epígrafe siguiente. En cambio, sí es frecuente que entre los regresores exista una relación aproximadamente lineal, en cuyo caso los estimadores que se obtengan serán en general poco precisos, aunque siguen conservando la propiedad de ser estimadores *ELIO*. En otras palabras, la relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre el regresando, lo que determina que las varianzas de los estimadores sean elevadas. Cuando se presenta una relación aproximadamente lineal entre los regresores, se dice que existe *multicolinealidad no perfecta*. Es importante señalar que el problema de multicolinealidad, surge porque no existe información suficiente para obtener una estimación precisa de los parámetros del modelo.

Para analizar este problema, vamos a examinar la varianza de un estimador. En el modelo de regresión lineal múltiple, el estimador de la varianza de un coeficiente de pendiente cualquiera – por ejemplo, de $\hat{\beta}_j$ - se puede formular de la siguiente forma:

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{6-21}$$

donde $\hat{\sigma}^2$ es el estimador insesgado de σ^2, n es el tamaño de la muestra, S_j^2 es la varianza muestral del regresor X_j y R_j^2 es el coeficiente de determinación obtenido al efectuar la regresión de X_j sobre el resto de los regresores del modelo.

El último de estos cuatro factores que determinan el valor de la varianza de $\hat{\beta}_j$ es el que se refiere a la multicolinealidad. Decimos que la multicolinealidad surge al estimar β_j cuando R_j^2 está "próximo" a 1 pero no hay una cota que se pueda fijar para concluir que la multicolinealidad es realmente un problema para la precisión de los estimadores. Aunque el problema de la multicolinealidad no puede definirse claramente, es cierto que, al estimar β_j, es mejor que la variable x_j tenga menos correlación con las otras variables independientes. Si un R_j^2 es igual a 1, tendríamos multicolinealidad perfecta y ninguna posibilidad de obtener estimaciones de los coeficientes. Cuando uno o más de los R_j^2 se aproximan a 1 la multicolinealidad tiene una cierta gravedad. En este caso, se presentan los siguientes problemas al realizar inferencias con el modelo:

a) Las varianzas de los estimadores son muy grandes.
b) Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.

6.3.2 Detección

Como la multicolinealidad es un problema *muestral*, ya que va asociada a la configuración concreta de la matriz de los regresores, no existen contrastes estadísticos, propiamente dichos, que sean aplicables para su detección. (Recuerde que los contrastes estadísticos van referidos a parámetros poblacionales). En cambio, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida la multicolinealidad afecta gravemente a las inferencias realizadas con un modelo. Estas reglas no son siempre fiables, siendo en algunos casos muy discutibles. En cualquier caso, se van a exponer algunas medidas que son útiles para detectar el grado de multicolinealidad: el *factor de agrandamiento de la varianza* (*FAV*) y la *tolerancia*, y el *número de condición* y el *coeficiente de descomposición de la varianza*.

Factor de agrandamiento de la varianza (FAV) y tolerancia

Con objeto de explicar el significado de estas medidas, supongamos que no existe ningún tipo de relación lineal entre el regresor x_j y el resto de regresores del modelo, es decir, el regresor x_j es *ortogonal* con el resto de los regresores. Entonces, R_j^2 será 0 y la varianza de $\hat{\beta}_j$ será igual a

$$\text{var}(\beta_j^*) = \frac{\hat{\sigma}^2}{nS_j^2} \tag{6-22}$$

El cociente entre (6-20) y (6-21) es precisamente el factor de agrandamiento de la varianza (*FAV*), cuya expresión será

$$FAV(\hat{\beta}_j) = \frac{1}{1-R_j^2} \tag{6-23}$$

Al estadístico *FAV* calculado de acuerdo a (6-23) se le denomina a veces "*FAV centrado*" para distinguirlo del "*FAV no centrado*" el cual tiene interés en los modelos sin término independiente. El programa E-views ofrece ambos estadísticos.

La tolerancia, que es la inversa de *FAV*, se define como,

$$Tolerancia(\hat{\beta}_j) = \frac{1}{FAV} = 1-R_j^2 \tag{6-24}.$$

Así, pues, el $FAV(\hat{\beta}_j)$ es la *ratio* entre la varianza observada y la que habría sido en caso de que x_j estuviera incorrelacionado con el resto de regresores del modelo. Dicho de otra forma, el *FAV* muestra en qué medida se «agranda» la varianza del estimador como consecuencia de la no ortogonalidad de los regresores. Se puede ver fácilmente que cuanto más elevado sea el FAV (o cuanto más baja sea la tolerancia), más elevada será la varianza de $\hat{\beta}_j$.

El procedimiento consiste en elegir a cada regresor como variable dependiente, y calcular la regresión sobre el resto de los regresores. De esta forma se obtendrían *k* valores del *FAV*. Si alguno de ello es elevado, es un indicio de multicolinealidad. Desafortunadamente, sin embargo, no hay ningún indicador teórico para determinar si el *FAV* es "alto". Tampoco, existe ninguna teoría que nos diga que hacer en caso de que exista multicolinealidad.

El *FAV* y la tolerancia son medidas utilizadas ampliamente. Algunos autores consideran que existe un problema grave de multicolinealidad cuando el *FAV* de algún coeficiente es mayor de 10, es decir, cuando el *FAV* >10, o análogamente cuando la *tolerancia*<0.10, pero esta regla no tiene una justificación científica.

El problema que tiene el *FAV* (o la tolerancia) es que no suministra ninguna información que pueda utilizarse para tratar el problema.

EJEMPLO 6.2 Analizando la multicolinealidad en el caso del absentismo laboral

En el ejemplo 3.1 se formuló y estimó, utilizando el fichero *absent*, un modelo para explicar el absentismo laboral en función de las variables edad, antigüedad y salario.

En cuadro 6.2 se ofrece información de la tolerancia y del *FAV* de cada variable. Según estos estadísticos la multicolinealidad no parece afectar al *salario* pero si tiene un cierto grado de importancia en las variables *edad* y *antigüedad*. En todo caso el problema de multicolinealidad de este modelo no parece ser serio ya que todos los *FAV* están por debajo de 5.

CUADRO 6.2. Tolerancia y *FAV*.

	Estadísticos de colinealidad	
	Tolerancia	*FAV*
edad	0.2346	4.2634
antigüedad	0.2104	4.7532
salario	0.7891	1.2673

Número de condición y el coeficiente de descomposición de la varianza

Este método, desarrollado por Belsey *et al*. (1982), está basado en la descomposición de la varianza de cada coeficiente de regresión en función de los raíces características λ_h de la matriz X'X y de los correspondientes vectores características asociados. No se discutirá aquí sobre los raíces y vectores característicos, ya que van más allá del objeto de este libro, pero en todo caso veremos su aplicación.

El *número de condición* es una medida estándar del mal condicionamiento de una matriz, e indica la sensibilidad potencial de una matriz inversa calculada con respecto a pequeños cambios en la matriz de partida (X'X en el caso de la regresión). Cuanto más cerca está la matriz de ser singular, más pequeños son los valores característicos. El número de condición (κ) se define como la raíz cuadrada de la mayor raíz característica ($\lambda_{\max}$) dividida por la más pequeña ($\lambda_{\min}$):

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{6-25}$$

Cuando no hay multicolinealidad en absoluto, todas las raíces características y el número de condición será igual a 1. Al crecer la multicolinealidad, las raíces características serán más grandes y más pequeñas que 1 (las raíces características próximas a 0 indican que existe un problema de multicolinealidad), y el número de condición crecerá. Una regla práctica de carácter informal establece que si el número de condición es mayor que 15, entonces la multicolinealidad es un problema, y si es mayor que 30 la multicolinealidad es un problema muy serio.

La varianza $\hat{\beta}_j$ según las contribuciones que aporta cada una de las raíces características puede expresarse del siguiente modo:

$$\text{var}(\hat{\beta}_j) = \sigma^2 \sum_h \frac{u_{jh}^2}{\lambda_h} \tag{6-26}$$

Así, la proporción de la contribución de λ_h a la varianza de $\hat{\beta}_j$ es igual a

$$\phi_{jh} = \frac{\dfrac{u_{jh}^2}{\lambda_h}}{\displaystyle\sum_{h=0}^{k} \frac{u_{jh}^2}{\lambda_h}} \tag{6-27}$$

Valores elevados de ϕ_{jh} indican que, como consecuencia de la multicolinealidad, existe una inflación de la varianza. Dado que las raíces características próximas a 0 indican un problema de multicolinealidad, es importante prestar una especial atención a las raíces características más pequeñas. Las contribuciones correspondientes a la raíz característica más pequeña pueden dar una clave de cuáles son los regresores que están implicados en el problema de multicolinealidad.

EJEMPLO 6.3 Analizando la multicolinealidad de los factores que determinan el tiempo dedicado al trabajo doméstico

Con objeto de analizar los factores que influyen sobre el tiempo dedicado al trabajo doméstico (*houswork*), se formuló el siguiente modelo en ejercicio 3.17, utilizando el archivo *timuse03*:

$$houswork = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

donde *educ* son los años de educación alcanzada, *hhinc* es la renta de la familia en euros por mes. Las variables *houswork* y *paidwork* están medidas en minutos por día.

El cuadro 6.3 proporciona información sobre las raíces características, ordenadas de la más pequeña a la mayor, y las proporciones de descomposición de la varianza para cada raíz característica están calculadas según (6-26). El número de condición es igual a

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{542.14}{7.06E-06}} = 8782$$

Como puede verse, el número de condición es muy elevado, lo que indicaría que el problema de multicolinealidad es muy importante.

Como puede verse en el cuadro 6.3[3] las proporciones más elevadas asociadas a la raíz característica más pequeña, que es la responsable de la multicolinealidad en este modelo, corresponden a los regresores *educ* y *age*. Estos dos regresores están inversamente correlacionados. Las proporciones más elevadas asociadas a la segunda raíz característica más pequeña corresponden a los regresores educación alcanzada y renta del hogar, que están positivamente correlacionadas.

CUADRO 6.3. Raíces características y proporciones de descomposición de la varianza.

Raíces características	7.03E-06	0.000498	0.025701	1.861396	542.1400

Proporciones de descomposición de la varianza

	Associated Eigenvalue				
Variable	1	2	3	4	5
C	0.999995	4.72E-06	8.36E-09	1.23E-13	1.90E-15
EDUC	0.295742	0.704216	4.22E-05	2.32E-09	3.72E-11
HHINC	0.064857	0.385022	0.209016	0.100193	0.240913
AGE	0.651909	0.084285	0.263805	5.85E-07	1.86E-08
PAIDWORK	0.015405	0.031823	0.007178	0.945516	7.80E-05

6.3.3 Soluciones

En principio, el problema de la multicolinealidad está relacionado con deficiencias en la información muestral. El diseño no experimental de la muestra es, a menudo, el responsable de estas deficiencias. Veamos a continuación algunas de las soluciones propuestas para resolver el problema de la multicolinealidad.

Eliminación de variables

La multicolinealidad puede atenuarse si se eliminan los regresores que son más afectados por la multicolinealidad. El problema que plantea esta solución es que los estimadores del nuevo modelo serán sesgados en el caso de que el modelo original fuera el correcto. Sobre esta cuestión conviene hacer la siguiente reflexión. El investigador está interesado en que un estimador sea insesgado (o, si no puede ser, que tenga un sesgo pequeño) y tenga una varianza reducida. El error cuadrático medio (*ECM*) recoge ambos factores. Así, para el estimador $\hat{\beta}_j$, el *ECM* se define de la siguiente manera:

[3] En el cuadro 6.3 las raíces características están ordenadas de menor a mayor lo mismo que las raíces características asociadas (*associated eigenvalue*) las proporciones de descomposición de la varianza. Conviene advertir que en el E-views las raíces características están ordenadas de mayor a menor. Por otra parte, el número de condición está definido de forma diferente a la que es usual en los manuales de econometría la cual hemos seguido.

$$ECM(\hat{\beta}_j) = \left[sesgo(\hat{\beta}_j) \right]^2 + var(\hat{\beta}_j) \qquad (6\text{-}28)$$

Si un regresor es eliminado del modelo, el estimador de un regresor que se mantiene (por ejemplo, $\hat{\beta}_j$) será sesgado, pero, sin embargo, su *ECM* puede ser menor que el correspondiente al modelo original, debido a que la omisión de una variable puede hacer disminuir suficientemente la varianza del estimador. En resumen, aunque la eliminación de una variable no es una práctica que en principio sea aconsejable, en ciertas circunstancias puede tener su justificación cuando contribuye a disminuir el *ECM*.

Aumento del tamaño de la muestra

Teniendo en cuenta que un cierto grado de multicolinealidad acarrea problemas cuando aumenta ostensiblemente las varianzas muestrales de los estimadores, las soluciones deben ir encaminadas a reducir estas varianzas. Esta solución no siempre es viable, puesto que los datos utilizados en las contrastaciones empíricas proceden generalmente de fuentes estadísticas diversas, interviniendo en contadas ocasiones el investigador en la recogida de información.

Por otro lado, cuando se trata de diseños experimentales, se puede incrementar directamente la variabilidad de los regresores sin necesidad de incrementar el tamaño de la muestra.

Utilización de información extramuestral

Otra posibilidad es la utilización de información extramuestral, bien estableciendo restricciones sobre los parámetros del modelo, bien aprovechando estimadores procedentes de otros estudios.

El establecimiento de restricciones sobre los parámetros del modelo reduce el número de parámetros a estimar y, por tanto, palia las posibles deficiencias de la información muestral. En cualquier caso, para que estas restricciones sean útiles deben estar inspiradas en el propio modelo teórico o, al menos, tener un significado económico.

En general, un inconveniente de esta forma de proceder es que el significado atribuible al estimador obtenido con datos de corte transversal es muy diferente del obtenido con datos temporales, en el caso de que se combinen ambos tipos de información. A veces, estos estimadores pueden resultar realmente «extraños» o ajenos al objeto de estudio.

Utilización de ratios

Si en lugar del regresando y de los regresores del modelo original se utilizan *ratios* con respecto al regresor que tenga mayor colinealidad, puede hacer que la correlación entre los regresores del modelo disminuya. Una solución de este tipo resulta muy atractiva, por su sencillez de aplicación. Sin embargo, las transformaciones de las variables originales del modelo que se estima utilizando *ratios* pueden provocar otro tipo de problemas. Suponiendo admisibles los supuestos del *MLC* con respecto a las perturbaciones originales del modelo, esta transformación modificaría implícitamente las propiedades del modelo, de tal manera que las perturbaciones del modelo transformado utilizando *ratios* ya no serían perturbaciones homoscedásticas, sino heteroscedásticas.

6.4 Contraste de normalidad

Los contrastes de significatividad *F* y *t* construidos en el capítulo 4 están basados en el supuesto de normalidad de las perturbaciones. Sin embargo, no es usual realizar contrastes de normalidad, quizás debido a que a menudo no se dispone de una muestra suficientemente grande -por ejemplo, 50 o más observaciones- que es necesaria para realizar contrastes sobre este supuesto. De todas formas, recientemente los contrastes sobre normalidad están recibiendo un interés creciente tanto en los estudios teóricos como aplicados.

Vamos a examinar a continuación un contraste para verificar el supuesto de normalidad de las perturbaciones en un modelo econométrico. Este contraste fue propuesto por Bera y Jarque, y está basado en los estadísticos de asimetría y curtosis de los residuos.

El estadístico de asimetría es un momento de tercer orden estandarizado, aplicado a los residuos, y su expresión es la siguiente:

$$\gamma_{1(\hat{u})} = \frac{\sum \hat{u}_i^3 / n}{\left[\sum \hat{u}_i^2 / n\right]^{3/2}} \tag{6-29}$$

En una distribución simétrica, como es el caso de la distribución normal, el coeficiente de asimetría es 0.

El estadístico de curtosis, que es un momento de cuarto orden estandarizado, tiene la siguiente expresión cuando se aplica a los residuos:

$$\gamma_{2(\hat{u})} = \frac{\sum \hat{u}_i^4 / n}{\left[\sum \hat{u}_i^2 / n\right]^{2}} \tag{6-30}$$

En una distribución normal estándar, es decir, en una distribución $N(0,1)$, el coeficiente de curtosis es igual a 3.

El estadístico de Bera y Jarque (*BJ*) viene dado por

$$BJ = \left[\frac{n}{6}\left(\gamma_{1(\hat{u})}\right)^2 + \frac{n}{24}\left(\gamma_{2(\hat{u})} - 3\right)^2\right] \tag{6-31}$$

En una distribución normal teórica, la anterior expresión tomará un valor nulo, ya que los coeficientes de asimetría y curtosis toman respectivamente los valores de 0 y 3. El estadístico *BJ* tomará valores elevados en la medida que el coeficiente de asimetría se aleje de 0 y de que el coeficiente de curtosis se aleje de 3. Bajo la hipótesis nula de normalidad, el estadístico *BJ* tiene la siguiente distribución:

$$BJ \xrightarrow[n\to\infty]{} \chi_2^2 \tag{6-32}$$

Con la indicación de $n \to \infty$, se quiere señalar que es un contraste asintótico, es decir, que tiene validez cuando la muestra sea suficientemente grande.

EJEMPLO 6.4 ¿Es aceptable la hipótesis de normalidad en el modelo para analizar la eficiencia de la Bolsa de Madrid?

En el ejemplo 4.5, utilizando el fichero *bolmadef*, se analizó la eficiencia del mercado de la Bolsa de Madrid en 1992, mediante un modelo que relaciona la tasa de rendimiento de un día sobre la tasa de rendimiento del día precedente. Ahora, vamos a realizar contrastes de normalidad sobre las perturbaciones de este modelo. Dada la poca proporción de varianza explicada con este modelo (véase

ejemplo 4.5), el contraste de normalidad de las perturbaciones es prácticamente equivalente a contrastar la normalidad de la variable endógena.

En el cuadro 6.4 se muestran los coeficientes de asimetría, curtosis y el estadístico de Bera y Jarque, aplicado a los residuos del modelo estimado. El coeficiente de asimetría (-0.04) no está muy alejado del valor 0 correspondiente a una distribución $N(0,1)$. Por otra parte, el coeficiente de curtosis (4.43) es algo diferente del valor 3 que toma en la distribución normal. En este caso, se rechaza el supuesto de normalidad para los niveles usuales de significación, ya que el estadístico de Bera y Jarque toma el valor de 21.02, que es más grande que $\chi_2^{2(0.01)}$ =9.21.

CUADRO 6.4. Contraste de normalidad en el modelo de la Bolsa de Madrid.

coeficiente de asimetría	*coeficiente de curtosis*	*estadístico Bera y Jarque*
-0.0421	4.4268	21.0232

El hecho de que se rechace con tanta contundencia el supuesto de normalidad puede parecer paradójico, ya que los valores de curtosis y, especialmente, de asimetría no difieren de forma sustancial de los valores que toman estos coeficientes en una distribución normal. Sin embargo, las discrepancias son suficientemente significativas porque están avaladas por un tamaño de muestra elevado (247 observaciones). Si *n* (el tamaño de la muestra) hubiera sido de 60 en lugar de 247, el estadístico *BJ*, calculado según (6-30) y utilizando los mismos coeficientes de asimetría y curtosis, toma el valor de 5.1068, que es más pequeño que $\chi_2^{2(0.01)}$ =9.21. Dicho de otra forma, con los mismos coeficientes, pero con una muestra menor, no proporcionan suficientes evidencias empíricas para rechazar la hipótesis nula de normalidad. Obsérvese que esto se debe a que el estadístico *BJ* crece proporcionalmente con el tamaño de la muestra, pero los grados de libertad (2) permanecen inalterables.

6.5 Heteroscedasticidad

El supuesto de homoscedasticidad (supuesto 7 del *MLC*) postula que las perturbaciones tienen una varianza constante, es decir,

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \tag{6-33}$$

Suponiendo que solo hay una variable independiente, el supuesto de homoscedasticidad significa que la variabilidad en torno a la línea de regresión es la misma a lo largo de toda la muestra de las *x*; es decir, que no aumenta o disminuye cuando *x* varía, como puede verse en la figura 2.7, parte a) del capítulo 2. En la figura 6.1 se ha representado el diagrama de dispersión correspondiente a un modelo en que las perturbaciones son homoscedásticas. Si el supuesto de homoscedasticidad no se cumple se dice que existe heteroscedasticidad, o que las perturbaciones son heteroscedásticas. En la figura 2.7, parte b) se representó un modelo con perturbaciones heteroscedásticas en el que la dispersión aumentaba al incrementarse el valor de *x*. En la figura 6.2 se ha representado el diagrama de dispersión correspondiente a un modelo en el que la dispersión de las perturbaciones crece al crecer *x*.

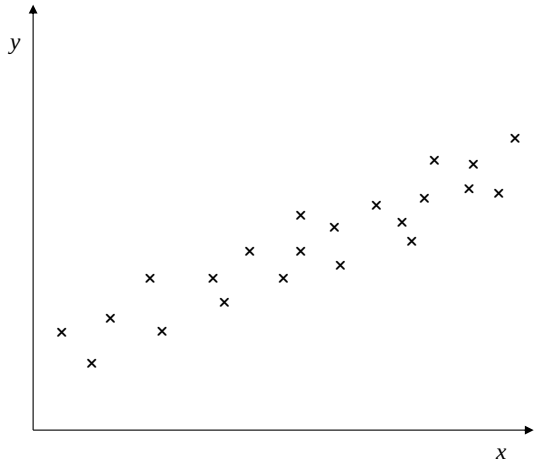


FIGURA 6.1. Diagrama de dispersión correspondiente a un modelo con perturbaciones homoscedásticas.

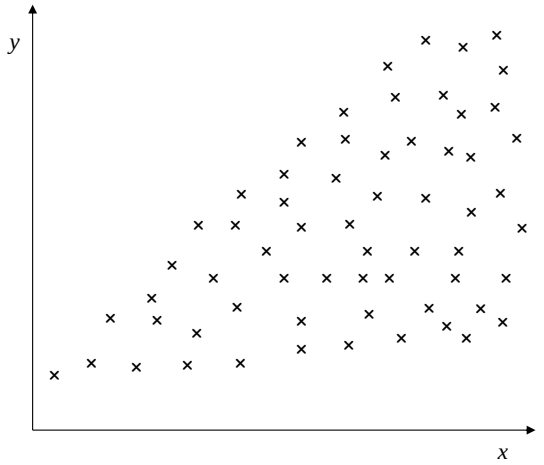


FIGURA 6.2. Diagrama de dispersión correspondiente a un modelo con perturbaciones heteroscedásticas.

6.5.1 Causas de la heteroscedasticidad

En los modelos estimados con datos de corte transversal, como por ejemplo en los estudios de demanda basados en encuestas de presupuestos familiares, es frecuente que se presenten problemas de heteroscedasticidad. De todas formas, la heteroscedasticidad también se puede presentar en modelos estimados con series temporales.

Vamos a considerar ahora algunos factores que pueden causar que las perturbaciones de un modelo sean heteroscedásticas:

a) Influencia del tamaño de una variable explicativa en el tamaño de la perturbación. Examinemos este factor utilizando un ejemplo. Supongamos un modelo en el que el gasto en hoteles es una función lineal de la renta disponible. Si se dispone de una muestra representativa de la población de un país se puede comprobar la gran variabilidad de la renta percibida por las distintas familias. Lógicamente, las familias con rentas bajas tienen pocas posibilidades de efectuar un gasto elevado en hoteles, pudiéndose esperar en este caso que las oscilaciones en el gasto de unas familias a otras no sean importantes. En cambio, en las familias con rentas altas se puede esperar una mayor variabilidad en este tipo de gasto. En efecto, las familias con rentas elevadas pueden optar entre gastar en hoteles una parte substancial de su renta o no gastar prácticamente nada. El diagrama de la figura 6.2 puede ser adecuado para representar lo que sucede en un modelo para explicar la demanda de un bien de lujo como es el caso del gasto en hoteles.

b) La presencia de valores anómalos (outliers) puede causar heteroscedasticidad. Un *outlier* es una observación generada aparentemente por una población diferente a la que ha generado el resto de las observaciones muestrales. Cuando el tamaño de muestra es pequeño la inclusión o exclusión de una observación de este tipo puede alterar substancialmente los resultados del análisis de regresión y causar heteroscedasticidad.

c) Transformación de los datos. Como hemos visto en un epígrafe previo una de las soluciones para resolver el problema de la multicolinealidad consistía en transformar el modelo tomado ratios con respecto a una variable (digamos, X_{ji}), es decir, dividiendo

ambos miembros del modelo por X_{ji}. En consecuencia, la perturbación será ahora u_i/X_{ji}, en lugar de u_i. Suponiendo que u_i cumple el supuesto de homoscedasticidad, las perturbaciones del modelo transformado (u_i/X_{ji}) ya no serán homoscedásticas sino heteroscedásticas.

6.5.2 Consecuencias de la heteroscedasticidad

Cuando existe heteroscedasticidad el método de mínimos cuadrados ordinarios (*MCO*), ya no es el más adecuado, ya que en ese caso los estimadores obtenidos no son óptimos, es decir, los estimadores de *MCO* no son *ELIO*.

Por otra parte, los estimadores obtenidos por *MCO* en el caso de que exista heteroscedasticidad, además de no ser *ELIO*, presentan el siguiente problema. La estimación de la matriz de covarianzas de los estimadores obtenida aplicando la fórmula usual no es válida cuando existe heteroscedasticidad. Consecuentemente, los estadísticos *t* y *F* basados en dicha estimación de la matriz de covarianzas darán lugar a inferencias erróneas.

6.5.3 Contrastes de heteroscedasticidad

Vamos a examinar dos contrastes de heteroscedasticidad: Breusch-Pagan-Godfrey y White. Ambos contrastes son asintóticos y tienen la forma de un contraste de multiplicadores de Lagrange (*ML*).

Contraste de Breusch-Pagan-Godfrey (*BPG*)

Breusch-Pagan (1979) desarrollaron un contraste para heteroscedasticidad y Godfrey desarrolló otro. Dada su similitud, se les conoce como contraste de heteroscedasticidad de Breusch-Pagan-Godfrey (*BPG*)

El contraste *BPG* es un contraste asintótico, es decir, válido solamente para muestras grandes. Las hipótesis nula y alternativa de este contraste pueden formularse de la siguiente forma:

$$\begin{aligned} &H_0 : E\left(u_i^2\right) = \sigma^2 \quad \forall i \\ &H_1 : \sigma_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} \end{aligned} \tag{6-34}$$

donde las z_i pueden ser todas o algunas de las x_i del modelo.

Tomando como referencia la anterior H_1, entonces H_0 puede expresarse como

$$H_0 : \alpha_2 = \alpha_3 = \cdots \alpha_m = 0 \tag{6-35}$$

Los pasos que se requieren en este contraste son los siguientes:

Paso 1. Se estima el modelo original y se calculan los residuos mínimo-cuadráticos.

Paso 2. Se realiza la siguiente regresión auxiliar, tomando como regresando al cuadrado de los residuos obtenidos en la estimación del modelo original ($\hat{u}_i^2$), ya que no se conocen ni σ_i^2 ni u_i^2:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} + \varepsilon_i \tag{6-36}$$

En la regresión auxiliar debe aparecer un término independiente, aunque el modelo original se haya estimado sin él. De acuerdo con la expresión (6-36), en la regresión auxiliar hay *m* regresores.

Paso 3. Designando por R^2_{ra} al coeficiente de determinación de la regresión auxiliar, se calcula el estadístico nR^2_{ra}.

Bajo la hipótesis nula, este estadístico (*BPG*) tiene la siguiente distribución:

$$BPG = nR^2_{ra} \xrightarrow[n\to\infty]{} \chi^2_m \tag{6-37}$$

Paso 4. Para un nivel de significación α, y designando por $\chi^{2(\alpha)}_m$ al valor en la tabla de la χ^2, la decisión a tomar es la siguiente:

Si $BPG > \chi^{2(\alpha)}_m$ se rechaza la H_0

Si $BP \leq \chi^{2(\alpha)}_m$ no se rechaza la H_0

En este contraste valores elevados del estadístico corresponden a una situación de heteroscedasticidad, es decir, al rechazo de la hipótesis nula.

EJEMPLO 6.5 Aplicación del contraste de Breusch-Pagan-Godfrey

Aplicamos a continuación este contraste a una submuestra de 10 observaciones, que se han utilizado para estimar los gastos en hostelería (*hostel)* en función de la renta disponible (*renta*). Los datos aparecen en el cuadro 6.5.

CUADRO 6.5. Datos de *hostel* y *renta*.

i	*hostel*	*renta*
1	17	500
2	24	700
3	7	250
4	17	430
5	31	810
6	3	200
7	8	300
8	42	760
9	30	650
10	9	320

Paso 1. Se aplican *MCO* al modelo

$$hostel = \beta_1 + \beta_2 renta + u$$

y, utilizando los datos del cuadro 6.5, se obtiene el siguiente modelo estimado:

$$\underset{}{hostel_i} = \underset{(3.48)}{-7.427} + \underset{(0.0065)}{0.0533}\, renta_i$$

Los residuos correspondientes a este modelo ajustado aparecen en el cuadro 6.6.

CUADRO 6.6. Residuos de la regresión de *hostel* sobre *renta*.

i	1	2	3	4	5	6	7	8	9	10
$\hat{u}_i$	-2.226	-5.888	1.100	1.505	-4.751	-0.234	-0.565	8.913	2.777	-0.631

Paso 2. La regresión auxiliar a estimar será la siguiente:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 renta_i + \eta_i$$

Aplicando *MCO* al anterior modelo se obtiene la siguiente estimación:

$$\hat{u}_i^2 = -23.93 + 0.0799 renta_i \; ; \qquad R^2=0.5045$$

Paso 3. A partir del valor de R^2 se obtiene el siguiente valor del estadístico *BPG*:

$$BPG=nR^2 =10(0.56)=5.05.$$

Paso 4. Dado que $\chi_1^{2(0.01)}$ =3.84, se rechaza la hipótesis nula de homoscedasticidad para un nivel del 5%, ya que *BPG*>3.84, pero no para el nivel de significación del 1%.

Tenga en cuenta que la validez de este contraste es asintótica. Sin embargo, la muestra utilizada en este ejemplo es muy pequeña.

Contraste de White

En el contraste de White no se especifican las variables que determinan la heteroscedasticidad. Este es un contraste no constructivo ya que no da ningún tipo de indicación del esquema de heteroscedasticidad cuando la hipótesis nula es rechazada

El contraste de White está basado en el hecho de que los errores estándar son válidos asintóticamente si se sustituye el supuesto de homoscedasticidad por el supuesto más débil de que la perturbación al cuadrado, u^2, está incorrelacionada con todos los regresores, sus cuadrados y los productos mixtos entre ellos. Teniendo en cuenta este hecho, White propuso hacer la regresión auxiliar de $\hat{u}_i^2$, puesto que u_i^2 es desconocido, con respecto a todos los factores que se acaban de mencionar. Si los coeficientes de la regresión auxiliar son conjuntamente no significativos, entonces podemos admitir que las perturbaciones son homoscedásticas. De acuerdo con el supuesto adoptado, el contraste de White es asintótico.

La aplicación del contraste de White puede plantear problemas en modelos con muchos regresores. Por ejemplo, si el modelo original tiene 5 variables independientes, la regresión auxiliar de White tiene 16 regresores (a menos que algunos sean redundantes), lo que implica que le regresión se realiza con una pérdida de 16 grados de libertad. Por esta razón, cuando el modelo tiene muchos regresores se aplica a menudo una versión *simplificada* del contraste de White. En esta versión simplificada se omiten los productos cruzados de la regresión auxiliar.

Los pasos que se requieren en este contraste son los siguientes:

Paso 1. Se estima el modelo original y se calculan los residuos mínimo-cuadráticos.

Paso 2. Se realiza la siguiente regresión auxiliar, tomando como regresando al cuadrado de los residuos obtenidos en la estimación del modelo original:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2\psi_{2i} + \alpha_3\psi_{3i} + \cdots + \alpha_m\psi_{mi} + \varepsilon_i \qquad (6\text{-}38)$$

Los regresores de la regresión auxiliar anterior ψ_{ji} son los regresores del modelo original, los cuadrados de los regresores y los productos cruzados de los regresores.

En cualquier caso, es necesario eliminar las posibles redundancias que se produzcan (es decir, regresores que aparezcan repetidos). Por ejemplo, no pueden aparecer simultáneamente como regresores el término independiente (que es un 1 para todas las observaciones) y el cuadrado de dicho regresor, ya que son idénticos. La introducción simultánea de estos dos regresores daría lugar a una situación de multicolinealidad perfecta.

En la regresión auxiliar debe aparecer un término independiente, aunque el modelo original se haya estimado sin él. De acuerdo con la expresión (6-38), se ha considerado que en la regresión auxiliar hay *m* regresores sin incluir el término independiente.

Paso 3 Designando por R^2_{ra} al coeficiente de determinación de la regresión auxiliar, se calcula el estadístico nR^2_{ra}.

Bajo la hipótesis nula, este estadístico (*W*) tiene la siguiente distribución:

$$W = nR^2_{ra} \xrightarrow[n\to\infty]{} \chi^2_m \qquad (6\text{-}39)$$

Con el estadístico nR^2_{ra} se contrasta la significatividad global del modelo (6-38).

Paso 4. Es similar al paso 4 en el contraste de Breusch-Pagan-Godfrey.

EJEMPLO 6.6 Aplicación del contraste de White

Este contrate se va aplicar a los datos del cuadro 6.5.

Paso 1. Este paso es igual que en el contraste de Breusch-Pagan-Godfrey.

Paso 2. Como existen dos regresores en el modelo original (término independiente y *renta*), los regresores de la regresión auxiliar son

$$\psi_{1i} = 1 \qquad \forall i$$
$$\psi_{2i} = 1 \times renta_i$$
$$\psi_{3i} = renta_i^2$$

En consecuencia, el modelo a estimar será

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 renta_i + \alpha_3 renta_i^2 + \eta_i$$

Aplicando *MCO* al anterior modelo, utilizando datos del cuadro 6.5, se obtiene la siguiente estimación:

$$\hat{u}_i^2 = 14.29 - 0.10 renta_i + 0.00018 renta_i^2 \qquad R^2=0.56$$

Paso 3. A partir del valor de R^2 se obtiene el estadístico *W*:

$$W=nR^2 = 10(0.56) = 5.60.$$

El número de grados de libertad es 2.

Paso 4. Dado que $\chi_2^{2(0.10)}=4.61$, se rechaza la hipótesis nula de homoscedasticidad para un nivel del 10% ya que $W=nR^2>4.61$, pero no para niveles de significación del 5% y del 1%.

Téngase en cuenta que la validez de este contraste también es asintótica.

EJEMPLO 6.7 Contrastes de heteroscedasticidad en la determinación del valor de las acciones de los bancos españoles

Para explicar el valor de mercado (*marktval*) de los bancos españoles en función de su valor contable (*bookval*) se han formulado dos modelos, uno lineal (ejemplo 2.8) y el otro doblemente logarítmico (ejemplo 2.10).

Heteroscedasticidad en el modelo lineal

El modelo lineal viene dado por

$$marktval=\beta_1+\beta_2 bookval+u$$

Utilizando datos de 20 bancos y entidades de seguros (fichero *bolmad95*) se han obtenido los siguientes resultados:

$$\underset{}{marktval} = \underset{(30.85)}{29.42} + \underset{(0.127)}{1.219}\, bookval$$

En el gráfico 6.1 se ha representado el diagrama de dispersión entre los residuos en valor absoluto (en ordenadas) y la variable *bookval* (en abscisas). Del examen de este gráfico se desprende que

los valores absolutos de los residuos, que son indicativos de la dispersión de esta serie, crecen al incrementarse los valores de la variable *bookval*. En otras palabras, este gráfico constituye un indicio, pero no una prueba formal, de la existencia de heteroscedasticidad de las perturbaciones asociada a la variable *bookval*.

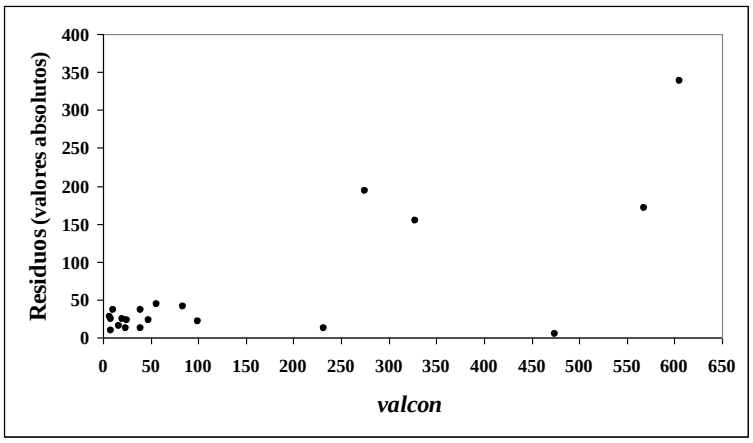


GRÁFICO 6.1. Diagrama de dispersión entre los residuos en valor absoluto y la variable *bookval* en el modelo lineal.

El estadístico de Breusch-Pagan-Godfrey toma el siguiente valor:

$$BPG = nR_{ra}^2 = 20\times 0.5220 = 10.44$$

Como $\chi_1^{2(0.01)}$ =6.64<10.44, se rechaza la hipótesis nula de homoscedasticidad para un nivel de significación del 1%, y, en consecuencia para α=0.05 y para α=0.10.

Vamos a aplicar a continuación el contraste de White. En este caso, en la regresión auxiliar se incluyen como regresores el término independiente, la variable *bookval*, y el cuadrado de esta variable. El estadístico de White toma el siguiente valor,

$$W = nR_{ra}^2 = 20\times 0.6017 = 12.03$$

Como $\chi_2^{2(0.01)}$ =9.21<12.03, se rechaza la hipótesis nula de homoscedasticidad para un nivel de significación del 1%.

Heteroscedasticidad en el modelo doblemente logarítmico

La estimación del modelo doblemente logarítmico con la misma muestra ha sido la siguiente:

$$\ln(marktval) = \underset{(0.265)}{0.676} + \underset{(0.062)}{0.9384}\ln(bookval)$$

En el gráfico 6.2 se ha representado el diagrama de dispersión entre los residuos en valor absoluto (en ordenadas), obtenidos al estimar el modelo el modelo anterior, y la variable ln(*bookval*) (en abscisas). Como puede verse, los dos residuos más grandes corresponden a dos bancos con valor contable pequeño. Aun no teniendo en cuenta estos dos casos, no parece que exista una relación entre los residuos y la variable explicativa del modelo.

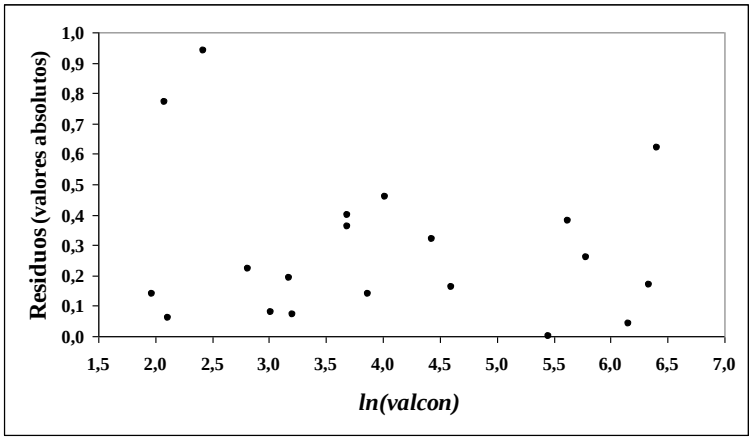


GRÁFICO 6.2. Diagrama de dispersión entre los residuos en valor absoluto y la variable ln(*bookval*) en el modelo doblemente logarítmico

Los resultados de los dos contrastes de heteroscedasticidad aplicados se presentan en el cuadro 6.7.

CUADRO 6. 7. Contrastes de heteroscedasticidad en el modelo doblemente logarítmico para explicar el valor de mercado de los bancos españoles.

Contraste	*Estadístico*	*Valores tablas*
Breusch-Pagan	$BP = nR_{ra}^2 = 1.05$	$\chi_2^{2(0.10)} = 4.61$
White	$W = nR_{ra}^2 = 2.64$	$\chi_2^{2(0.10)} = 4.61$

Como puede verse, ambos contrastes son concluyentes en que no se puede rechazar la hipótesis nula de homoscedasticidad frente a la hipótesis alternativa de que la varianza de las perturbaciones está asociada a la variable explicativa del modelo.

Una conclusión importante de este caso es la siguiente. Cuando en la estimación de un modelo econométrico con datos de corte transversal hay unidades de muy distinto tamaño, los problemas de escala pueden provocar heteroscedasticidad en las perturbaciones. Estos problemas pueden resolverse en muchas ocasiones utilizando modelos logarítmicos.

EJEMPLO 6.8 ¿Existe heteroscedasticidad en la demanda de servicios de hostelería?

En general, en la demanda de bienes alimenticios no suele aparecer heteroscedasticidad en las perturbaciones. En cambio, en la demanda de bienes de lujo la heteroscedasticidad suele ser mucho más frecuente, debido a que en la demanda de estos bienes puede haber una disparidad muy grande en el comportamiento de los hogares con rentas elevadas, frente a los hogares con rentas bajas en los que es muy improbable que exista tal disparidad dado lo reducido de la renta.

A la vista de las consideraciones anteriores, vamos a estimar un modelo en el que se explica el logaritmo del gasto en servicios de hostelería -ln(*hostel*)- en función del logaritmo de la renta disponible -ln(*inc*)- y de otras variables demográficas y sociales.

La especificación utilizada para la estimación de la demanda de los servicios de hostelería es la siguiente:

$$\ln\ hostel\ = \beta_1 + \beta_2 \ln(inc) + \beta_3 secstud + \beta_4 terstud + \beta_5 hhsize + u \qquad (6\text{-}40)$$

donde *inc* es la renta disponible del hogar, *hhsize* es el número de miembros del hogar, y *secstud* y *terstud* son dos variables ficticias que el valor 1 si han completado estudios secundarios y terciarios respectivamente.

Los resultados de la regresión obtenidos son los siguientes (fichero *hostel*):

$$\ln(hostel)_i = \underset{(2.26)}{-16.37} + \underset{(0.324)}{2.732}\ln(inc)_i + \underset{(0.258)}{1.398}\, secstud_i + \underset{(0.333)}{2.972}\, terstud_i - \underset{(0.088)}{0.444}\, hhsize_i$$

A la vista de estos resultados, puede afirmarse que los servicios de hostelería son un bien de lujo, ya que la elasticidad demanda/renta para este bien es muy elevada (2.73). Esto quiere decir que, si la renta se incrementa en un 1%, el gasto en servicios de hostelería aumentará, en promedio, en un 2.73%. Como puede verse las familias en las que el sustentador principal tiene estudios medios (*secstud*) o, en mayor

medida, estudios superiores (*terstud*), realizan un mayor gasto en servicios de hostelería que cuando el sustentador principal solamente tiene estudios primarios. Por el contrario, este gasto disminuye al aumentar el tamaño del hogar (*hhsize*).

En el gráfico 6.3 se ha representado el gráfico de dispersión entre los residuos en valor absoluto y la variable ln(*inc*), ya que, en los modelos de demanda, en los que aparece la renta (o una transformación de la misma) como variable explicativa, es esta variable la principal candidata, por no decir la única, para explicar la hipotética heteroscedasticidad en las perturbaciones. Como puede verse en el gráfico, la dispersión de los residuos es más reducida para las rentas bajas, que en las rentas medias o altas.

Vamos a aplicar a continuación los dos contrastes de heteroscedasticidad que se han expuesto en este apartado.

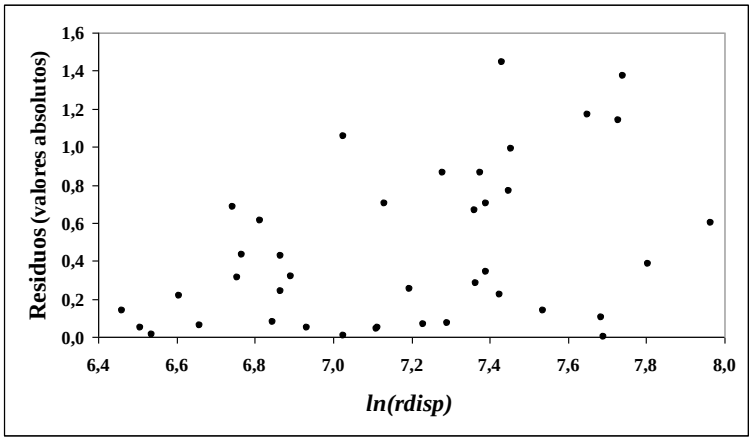


GRÁFICO 6.3. Diagrama de dispersión entre los residuos en valor absoluto y la variable ln(*inc*) en la estimación del modelo de hostelería.

Los resultados de los dos contrastes de heteroscedasticidad examinados se presentan en el cuadro 6.8.

CUADRO 6. 8. Contrastes de heteroscedasticidad en el modelo de demanda de servicios de hostelería.

Contraste	*Estadístico*	*Valores tablas*
Breusch-Pagan	$BP= nR^2_{ra}$ =7.83	$\chi_2^{2(0.05)}$ =5.99
White	$W= nR^2_{ra}$ =12.24	$\chi_2^{2(0.01)}$ =9.21

En el contraste de *BPG* se rechaza la hipótesis nula de homoscedasticidad para un nivel de significación de α=0.05 pero no para un nivel de α=0.01.

En la aplicación del contraste de White, dado que hay muchas variables dicotómicas en el modelo, la inclusión de los productos cruzados en la regresión auxiliar puede dar lugar a serios problemas de multicolinealidad. Por esta razón, en la regresión auxiliar no se han incluido los productos cruzados. Como es lógico, entre los regresores de la regresión auxiliar no figuran los cuadrados de *secstud* y *terstud*, ya que los cuadrados de estos regresores son ellos mismos por tratarse de variables dicotómicas. Dado el valor obtenido en el estadístico de White, se rechaza la hipótesis nula de homoscedasticidad para un nivel de significación de α=0.01. En consecuencia, el contraste de White es más concluyente en el rechazo del supuesto de homoscedasticidad.

6.5.4 Estimación de la matriz de covarianzas consistente bajo heteroscedasticidad

Cuando existe heteroscedasticidad y aplicamos *MCO*, no podemos realizar inferencias correctas si utilizamos la matriz de covarianzas asociada a las estimaciones por *MCO*, ya que esta matriz no es un estimador consistente de la matriz de covarianzas de los coeficientes. En consecuencia, los estadísticos *t* y *F* basados en dicha matriz de covarianzas estimada conducen a inferencias erróneas.

Por tanto, si existe heteroscedasticidad y ha sido aplicado el método de *MCO*, para realizar inferencias debería buscarse un estimador de la matriz de covarianzas que sea consistente bajo el supuesto de heteroscedasticidad. White propuso un estimador que es consistente bajo este supuesto. No obstante, es importante tener en cuenta que este estimador no trabaja bien en pequeñas muestras, ya que es una aproximación asintótica.

La mayoría de los paquetes econométricos permiten calcular desviaciones estándar de los estimadores por el procedimiento de White. Utilizando estos errores estándar consistentes se pueden hacer contrastes correctos bajo el supuesto de heteroscedasticidad.

EJEMPLO 6.9 Errores estándar consistentes en la determinación del valor de las acciones de los bancos españoles (Continuación ejemplo 6.7)

En la siguiente ecuación estimada del modelo lineal las desviaciones típicas de los estimadores son calculadas por el procedimiento de White y, por tanto, son consistentes bajo el supuesto de heteroscedasticidad:

$$\underset{}{marktval} = \underset{(18.67)}{29.42} + \underset{(0.249)}{1.219}\,bookval$$

Como puede comprobarse, el error estándar del coeficiente de *bookval* pasa de 0.127 aplicando el procedimiento usual a 0.249 en el procedimiento de White. De todas formas, el nivel de significación crítico sigue siendo muy bajo, ya que su valor se sitúa en 0.0001. En consecuencia, se sigue manteniendo la significatividad de la variable *bookval* para todos los niveles usuales. Por el contrario, el término independiente, que no tiene especial relevancia en el modelo, tiene ahora un error estándar (18.67), que es inferior al obtenido con el procedimiento usual (30.85).

Si aplicamos el procedimiento de White al modelo doblemente logarítmico se obtienen los siguientes resultados:

$$\ln(marktval) = \underset{(0.3218)}{0.676} + \underset{(0.0698)}{0.9384}\ln(bookval)$$

En este caso, el error estándar del coeficiente ln(*bookval*) es prácticamente el mismo en los dos procedimientos.

De los anteriores resultados pueden obtenerse las siguientes conclusiones. En la determinación del valor de las acciones de los bancos españoles, las perturbaciones del modelo lineal son fuertemente heteroscedásticas. Por ello, al realizar una estimación consistente, la desviación típica casi se duplica con respecto al procedimiento usual. Por el contrario, en el modelo doblemente logarítmico, que no está afectado por la heteroscedasticidad, apenas hay diferencias entre los errores estándar que se obtienen por ambos procedimientos.

6.5.5 Tratamiento de la heteroscedasticidad

Para realizar la estimación de un modelo con perturbaciones heteroscedásticas es necesario conocer o, en caso de que no se conozca, estimar el esquema de heteroscedasticidad. Así, supongamos que la desviación típica de las perturbaciones sigue el siguiente esquema:

$$\sigma_i = f\left(x_{ji}\right) \tag{6-41}$$

Como se ha indicado en el epígrafe 6.1, la aplicación del método de *MCG* permite obtener estimadores *ELIO* cuando las perturbaciones son heteroscedásticas. Conocido el esquema (6-41), la aplicación de *MCG* se realiza en dos fases. En la primera etapa se transforma el modelo original dividiendo ambos miembros por la desviación estándar. Por lo tanto, de acuerdo con (6-40), el modelo transformado vendrá dado por

$$\frac{y_i}{f\left(x_{ji}\right)} = \beta_1 \frac{1}{f\left(x_{ji}\right)} + \beta_2 \frac{x_{1i}}{f\left(x_{ji}\right)} + \beta_3 \frac{x_{2i}}{f\left(x_{ji}\right)} + \cdots + \beta_k \frac{x_{ki}}{f\left(x_{ki}\right)} + \frac{u_i}{f\left(x_{ji}\right)} \tag{6-42}$$

Puede verse fácilmente que las perturbaciones del modelo anterior, ($u_i/f(x_{ji})$), son homoscedásticas. Por ello, en la segunda etapa se aplican *MCO* al modelo transformado, ya que se obtendrán estimadores *ELIO*. Dado que, al dividir por $f(x_{ji})$, se está ponderando cada observación por el inverso del valor que toma esta función, al procedimiento anterior se le denomina frecuentemente *mínimos cuadrados ponderados* (*MCP*). En este caso, el factor de ponderación es $1/f(x_{ji})$.

Si no se conoce la función $f(x_{ji})$, es necesario proceder a su estimación. En ese caso, el método de estimación no será exactamente *MCG*, ya que la aplicación de este método implica el conocimiento de la matriz de covarianzas, o al menos el conocimiento de una matriz que sea proporcional a ésta. Cuando se estima la matriz de covarianzas, además de los parámetros, se dice que se aplican *MCG factibles*. En el caso de perturbaciones heteroscedásticas, a la particularización del método de *MCG* factibles, se le denomina *MCP* en dos etapas. En la primera etapa se estima la función $f(x_{ji})$, mientras que en la segunda etapa se aplica *MCO* al modelo transformado utilizando las estimaciones de $f(x_{ji})$.

Para ver como se puede aplicar el método de *MCP* en dos etapas, vamos a partir de la siguiente relación, que simplemente define la varianza de las perturbaciones, en el caso de heteroscedasticidad,

$$E\left(u_i^2\right)=\sigma_i^2 \tag{6-43}$$

Por lo tanto, la perturbación al cuadrado se puede hacer igual, como en el modelo de regresión, a su esperanza más una variable aleatoria, es decir,

$$u_i^2=\sigma_i^2+\varepsilon_i \tag{6-44}$$

Como las perturbaciones no son observables, se puede establecer una relación análoga a la anterior utilizando los residuos en lugar de las perturbaciones. Por lo tanto, se tiene que

$$\hat{u}_i^2=\sigma_i^2+\eta_{2i} \tag{6-45}$$

Es preciso tener en cuenta que la relación anterior no tiene exactamente las mismas propiedades que (6-44), debido a que los residuos están correlacionados y son heteroscedásticos, aunque las perturbaciones cumplan con todos los supuestos del *MLC*. Sin embargo, en grandes muestras las propiedades son las mismas.

Si utilizamos los residuos como regresando, en lugar de los residuos al cuadrado, habrá que tomar valores absolutos, ya que la desviación estándar solo toma valores positivos. Si se tiene en cuenta (6-45), se puede establecer la siguiente relación:

$$\left|\hat{u}_i\right|=\sigma_i^2+\eta_{2i}=f\left(x_{ji}\right)+\eta_{2i} \tag{6-46}$$

Dado que la función $f(x_{ji})$ será en general desconocida, se suelen ensayar distintas funciones. A continuación, presentamos algunas de las funciones más usuales:

$$\begin{aligned}
|\hat{u}_i| &= \alpha_1 + \alpha_2 x_{ji} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \sqrt{x_{ji}} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \frac{1}{x_{ji}} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \ln(x_{ji}) + \eta_{2i}
\end{aligned} \tag{6-47}$$

A la vista de los resultados, se selecciona aquella forma funcional con la que se obtenga un mejor ajuste (un coeficiente de determinación más elevado o un estadístico AIC más pequeño). Para la transformación del modelo se contemplan dos circunstancias, según cuál sea la significatividad del término independiente. Si este coeficiente es estadísticamente significativo, se transforma el modelo dividiendo por los valores ajustados de la ecuación seleccionada. Si no es estadísticamente significativo, se transforma el modelo dividiendo por el regresor correspondiente a la ecuación seleccionada. Así, si la ecuación seleccionada fuera la segunda de (6-47), no siendo significativo el término independiente, el modelo transformado sería el siguiente:

$$\frac{y_i}{\sqrt{x_{ji}}} = \beta_1 \frac{1}{\sqrt{x_{ji}}} + \beta_2 \frac{x_{2i}}{\sqrt{x_{ji}}} + \beta_3 \frac{x_{3i}}{\sqrt{x_{ji}}} + \cdots + \beta_k \frac{x_{ki}}{\sqrt{x_{ji}}} + \frac{u_i}{\sqrt{x_{ji}}} \tag{6-48}$$

Obsérvese que en el caso de que el término independiente no sea significativo, en la transformación del modelo no intervienen parámetros estimados, pero si lo harán en el caso de que sea significativo dicho término independiente. Como los estimadores de los modelos (6-47) no son insesgados, aunque sí consistentes, no es conveniente realizar transformaciones con valores ajustados -en cuyo cálculo intervienen $\hat{\alpha}_1$ y $\hat{\alpha}_2$- salvo que sea muy fuerte (por ejemplo, superior al 1%) la significatividad del término independiente.

EJEMPLO 6.10 Aplicación de mínimos cuadrados ponderados en la demanda de servicios de hostelería (Continuación 6.8)

Dado que los dos contrastes aplicados al modelo para explicar el gasto de los servicios de hostelería indican que las perturbaciones son heteroscedásticas, vamos a aplicar el método de mínimos cuadrados ponderados para estimar del modelo (6-40).

En primer lugar, se estiman los cuatro modelos (6-47), utilizando como regresando a los residuos en valor absoluto $|\hat{u}_i|$ obtenidos en la estimación del modelo (6-40) por *MCO*. Los resultados de estas estimaciones se presentan a continuación:

$$|\hat{u}_i| = \underset{(0.143)}{0.0239} + \underset{(2.73)}{0.0003}\,inc \qquad R^2 = 0.1638$$

$$|\hat{u}_i| = \underset{(-1.34)}{-0.4198} + \underset{(2.82)}{0.0235}\sqrt{inc} \qquad R^2 = 0.1733$$

$$|\hat{u}_i| = \underset{(5.39)}{0.8857} - \underset{(-2.87)}{532.1}\frac{1}{inc} \qquad R^2 = 0.1780$$

$$|\hat{u}_i| = \underset{(-2.46)}{-2.7033} + \underset{(2.88)}{0.4389}\ln(inc) \qquad R^2 = 0.1788$$

En los resultados anteriores debajo de cada coeficiente aparece el estadístico *t*.

La forma funcional seleccionada es la que utiliza ln(*inc*) como regresor, ya que para ella se obtiene el ^{o} *11111111111111111111Q11* más elevado. Dado que el coeficiente del término independiente no es estadísticamente significativo al 1% y siguiendo la recomendación hecha, se van a aplicar *MCP*, tomando como ponderación 1/ln(*inc*). En la estimación por *MCP* se han obtenido los siguientes resultados:

$$\ln(hostel)_i = \underset{(2.15)}{-16.21} + \underset{(0.309)}{2.709}\ln(inc)_i + \underset{(0.247)}{1.401}secstud_i + \underset{(0.326)}{2.982}terstud_i - \underset{(0.085)}{0.445}hhsize_i$$

$$R^2=0.914 \qquad n=40$$

Comparando con la estimación por MCO, hecha en el ejemplo 6.5, puede verse que las diferencias son muy pequeñas, lo que es indicativo de la robustez del modelo.

6.6 Autocorrelación

El supuesto de *no autocorrelación*, o de *no correlación serial*, (supuesto 8 del *MLC*) postula que las perturbaciones con diferentes subíndices no están correlacionadas entre sí:

$$E(u_i u_j) = 0 \qquad i \neq j \tag{6-49}$$

Es decir, las perturbaciones correspondientes a diferentes periodos de tiempo, o a individuos diferentes, no están correlacionadas entre sí. En la figura 6.3 se muestra un gráfico que corresponde a perturbaciones que no están autocorrelacionados. El eje *x* es el tiempo. Como se puede observar, las perturbaciones se distribuyen aleatoriamente por encima y por debajo de la línea 0 (media teórica de *u*). En la figura, cada perturbación está unida por una línea a la perturbación del período siguiente: en total esta línea cruza la línea 0 en 13 ocasiones.

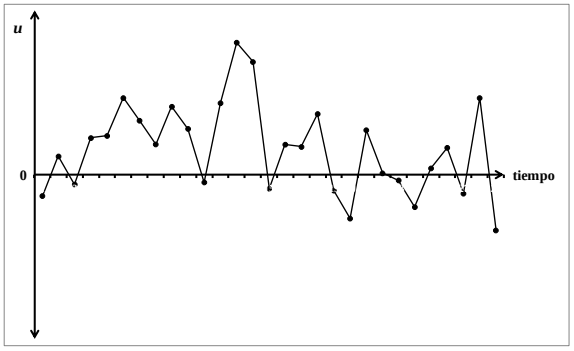


FIGURA 6.3. Gráfico de perturbaciones no autocorrelacionadas.

La transgresión del supuesto de no autocorrelación se produce con bastante frecuencia en los modelos que utilizan datos de series temporales. Hay que señalar también que la autocorrelación puede ser tanto positiva como negativa. La autocorrelación positiva se caracteriza por dejar una estela a lo largo del tiempo, debido a que el valor de cada perturbación se encuentra próximo al valor de la perturbación que le precede. La autocorrelación positiva se produce mucho más frecuentemente en la práctica que la negativa. En la figura 6.4 se muestra un gráfico que corresponde a las perturbaciones que están positivamente autocorrelacionadas. Como puede verse, la línea que une las perturbaciones sucesivas cruza la línea 0 en sólo 4 veces.

Por el contrario, las perturbaciones afectadas por autocorrelación negativa presentan una configuración de dientes de sierra, y a menudo cada perturbación tiene el signo opuesto de la perturbación que le precede. En la figura 6.5 el gráfico corresponde a perturbaciones que están negativamente autocorrelacionadas. Ahora la línea 0 es cruzada en 21 ocasiones por la línea que une las perturbaciones sucesivas.

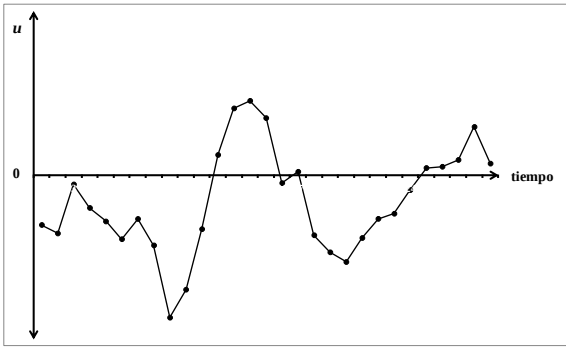


FIGURA 6.4. Gráfico de perturbaciones autocorrelacionadas positivamente.

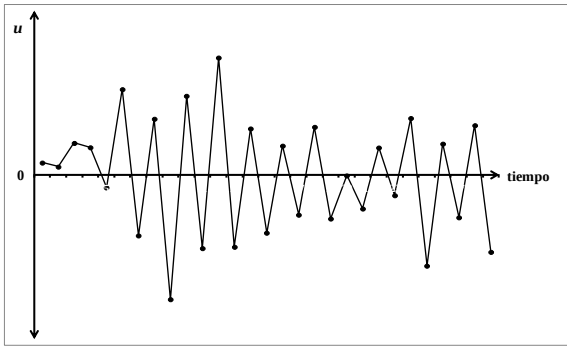


FIGURA 6.5. Gráfico de perturbaciones autocorrelacionadas negativamente.

6.6.1 Causas of autocorrelación

Existen varias causas para la presencia de autocorrelación en un modelo. Veamos a continuación algunas de ellas.

a) Sesgo de especificación. Puede deberse al uso de una forma funcional incorrecta o a la omisión de una variable relevante.

Supongamos que la forma funcional correcta para determinar el *salario* en función de los años de experiencia (*exp*) es la siguiente:

$$salario = \beta_1 + \beta_2 exp + \beta_3 exp^2 + u$$

En vez de este modelo se ajusta el siguiente:

$$salario = \beta_1 + \beta_2 exp + v$$

En el segundo modelo de la perturbación tiene un componente sistemático ($v = \beta_3 exp^2 + u$). En la figura 6.5 se ha representado un diagrama de dispersión (generado por el primer modelo) y la función ajustada del segundo modelo. Como puede verse, para valores de *exp* bajos se sobrestiman los salarios; para valores intermedios de *exp* se subestiman los salarios; por último, para valores elevados de *exp* el modelo ajustado sobrestima de nuevo a los salarios. Este ejemplo ilustra un caso en el que el uso de una forma funcional incorrecta provoca autocorrelación positiva.

Por otra parte, la omisión de una variable relevante en el modelo podría inducir autocorrelación positiva si esa variable tiene, por ejemplo, un comportamiento cíclico.

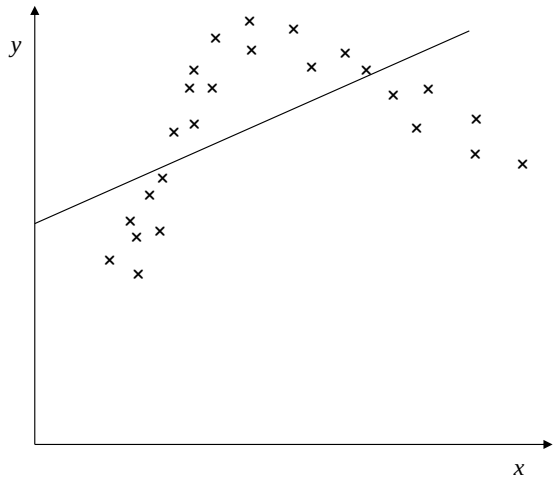


FIGURA 6.6. Perturbaciones autocorrelacionadas debidas a un sesgo de especificación.

b) Inercia. El término de perturbación en una ecuación de regresión refleja la influencia de las variables que afectan a la variable dependiente no incluidas en la ecuación de regresión. Precisamente, la inercia o la persistencia de los efectos de las variables excluidas del modelo -e incluidas en *u*- es probablemente la causa más frecuente de que exista autocorrelación positiva. Como es bien sabido, las series temporales macroeconómicas tales como el *PIB*, la producción, el empleo y los índices de precios tienden a moverse conjuntamente: en periodos de expansión estas series tienden a aumentar de forma más o menos paralela, mientras que en los tiempos de contracción del ciclo tienden a disminuir también en una forma paralela. Por esta razón, en las regresiones con datos de series temporales, es muy probable las observaciones sucesivas de la perturbación dependan de los valores previos. Con ello, este comportamiento cíclico puede producir autocorrelación en las perturbaciones.

c) Transformación de datos. A modo de ejemplo consideremos el siguiente modelo que explica el consumo en función de la renta:

$$cons_t = \beta_1 + \beta_2 inc_t + u_t \tag{6-50}$$

Para la observación *t*-1, tenemos que

$$cons_{t-1} = \beta_1 + \beta_2 inc_{t-1} + u_{t-1} \tag{6-51}$$

Si restamos (6-51) de (6-50), obtenemos

$$\Delta cons_t = \beta_2 \Delta inc_t + \Delta u_t \tag{6-52}$$

donde $\Delta cons_t = cons_t - cons_{t-1}$, $\Delta inc_t = inc_t - inc_{t-1}$ y $v_t = \Delta u_t = u_t - u_{t-1}$.

A la ecuación (6-50) se le conoce como ecuación en forma de *niveles*, mientras que a la ecuación (6-52) se le conoce como ecuación en forma de *primeras diferencias*. En el análisis empírico se utilizan ambas especificaciones. Si la perturbación no está autocorrelacionada en (6-50), la perturbación en (6-52), que es igual a $v_t = u_t - u_{t-1}$, sí que lo estará, ya que v_t y v_{t-1} tienen un elemento en común (u_{t-1}). En cualquier caso, conviene advertir que el modelo (6-52) tal como está especificado puede plantear otros problemas econométricos que no serán examinados aquí.

6.6.2 Consecuencias de la autocorrelación

Las consecuencias de la autocorrelación para *MCO* son similares a las de la heteroscedasticidad. Por lo tanto, si las perturbaciones están autocorrelacionadas, el estimador por *MCO* no es *ELIO*, ya que se puede encontrar otro estimador insesgado alternativo que tenga menor varianza. Además de no ser *ELIO*, el estimador obtenido por *MCO* bajo el supuesto de autocorrelación presenta el problema de que la estimación de la matriz de covarianzas de los estimadores calculada por las fórmulas usuales de *MCO* está sesgada y, por consiguiente, los estadísticos *t* y *F* basados en esta matriz de covarianzas puede llevar a inferencias erróneas.

6.6.3 Contrastes de autocorrelación

Para realizar contrastes de autocorrelación hay que especificar la hipótesis alternativa que defina un esquema de autocorrelación de las perturbaciones. A continuación, se van a examinar tres de los más conocidos contrastes. En dos de ellos (el contraste de Durbin y Watson y el contraste *h* de Durbin) la hipótesis alternativa es

un esquema autorregresivo de primer orden, mientras que en el tercero, denominado contraste de Breusch-Godfrey, es un contraste general de autocorrelación aplicable a esquemas autorregresivos de orden más elevado.

Contraste de Durbin y Watson

El contraste *d* de Durbin y Watson fue propuesto por estos económetras en el año 1950. Para referirse a este estadístico es también usual la denominación de *DW*.

Durbin y Watson proponen el siguiente esquema sobre las perturbaciones aleatorias u_t:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \rightarrow NID(0, \sigma^2) \tag{6-53}$$

El esquema propuesto para las u_t es un esquema autorregresivo de primer orden, ya que las perturbaciones aparecen como regresando y también como regresor con un periodo de desfase. En la terminología usual del análisis de series temporales, al esquema (6-53) se le denomina *AR*(1), es decir, un proceso autorregresivo de orden 1. El coeficiente de este esquema es ρ al que se exige que sea menor que 1 en valor absoluto con objeto de que las perturbaciones no tengan un carácter explosivo, al crecer indefinidamente *n*. La variable ε_t es una variable aleatoria para la que se postula una distribución normal e independiente (esto es lo que quiere decir *NID*) con media 0 y varianza σ^2. En consecuencia, sobre la variable ε_i se postulan los mismos supuestos que se postularon para u_i en los supuestos del *MLC*. A la variable que goza de estas propiedades se le suele denominar variable ruido blanco.

Según que el valor de ρ sea positivo o negativo la autocorrelación será positiva o negativa. La autocorrelación positiva es, con mucha diferencia, la que se presenta con mucha más frecuencia en la práctica. Por otra parte, casi siempre se realizan contrastes de una sola cola, es decir, se toma como hipótesis alternativa o la autocorrelación positiva o la autocorrelación negativa.

La construcción de un contraste de autocorrelación de las perturbaciones presenta el problema de que éstas no son observables, por lo que el contraste se tiene que basar en los residuos obtenidos por *MCO*. Esta circunstancia plantea problemas, ya que, bajo la hipótesis nula de que las perturbaciones no están autocorrelacionadas, los residuos en cambio sí lo están. Durbin y Watson, en la construcción de su contraste, sí tuvieron en cuenta esa circunstancia.

Veamos ahora como se aplica este contraste. Tomando como referencia el esquema definido en (6-53), Durbin y Watson formulan las siguientes hipótesis nula y alternativa de autocorrelación positiva

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned} \tag{6-54}$$

Así pues, bajo la hipótesis nula se verifica que $u_t = \varepsilon_t$, es decir, el modelo cumple los supuestos del *MLC*.

El estadístico que utilizan Durbin y Watson para el contraste de las hipótesis (6-54) es el estadístico *d*, o *DW*, definido de la siguiente forma:

$$d = DW = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})}{\sum_{t=1}^{n}\hat{u}_t^2} \tag{6-55}$$

La distribución del estadístico *d*, que es simétrica con una media igual a 2, es muy complicada, ya que depende de la forma concreta de la matriz de regresores **X**, del tamaño de la muestra (*n*) y del número de regresores (*k*) excluido el término independiente

De todas formas, Durbin y Watson, para diferentes niveles de significación, tabularon dos valores (d_L y d_U) para cada valor de *n* y de *k*. Las reglas para contrastar autocorrelación positiva son las siguientes:

$$\begin{aligned} &\text{Si } d < d_L && \text{, existe autocorrelación positiva.} \\ &\text{Si } d_L \le d \le d_U && \text{, no es concluyente el contraste.} \\ &\text{Si } d > d_U && \text{, no existe autocorrelación positiva.} \end{aligned} \tag{6-56}$$

Como puede verse, existen unos valores en los que el contraste no es concluyente. Esto se debe al efecto que la configuración concreta de la matriz **X** tiene en la distribución de *d*.

Si se desea realizar el contraste de autocorrelación negativa, la hipótesis alternativa es la siguiente:

$$H_1 : \rho < 0 \tag{6-57}$$

Para aplicar el contraste de autocorrelación negativa se tiene en cuenta que el estadístico *d* tiene una distribución simétrica con un recorrido entre 0 y 4. Las reglas, por lo tanto, son las siguientes:

$$\begin{aligned} &\text{Si } d > 4 - d_L && \text{, existe autocorrelación negativa.} \\ &\text{Si } 4 - d_U \le d \le 4 - d_L && \text{, no es concluyente el contraste.} \\ &\text{Si } d < 4 - d_U && \text{, no existe autocorrelación negativa.} \end{aligned} \tag{6-58}$$

El contraste de Durbin y Watson no es aplicable cuando entre los regresores haya variables endógenas desfasadas.

Para su aplicación a datos trimestrales, Wallis consideró el siguiente esquema autorregresivo de cuarto orden:

$$u_t = \rho_4 u_{t-4} + \varepsilon_i \qquad |\rho_4| < 1 \qquad \varepsilon_t \to NID(0, \sigma^2) \tag{6-59}$$

El anterior esquema es similar a (6-53), con la diferencia de que la perturbación del segundo miembro está retardada 4 periodos. El estadístico de contraste de Wallis es similar a (6-55), pero teniendo en cuenta que ahora el retardo es de 4 periodos. Este autor diseñó unas tablas *ad hoc* para contrastar el modelo (6-59).

EJEMPLO 6.11 Autocorrelación en el modelo para determinar la eficiencia de la Bolsa de Madrid

En el ejemplo 4.5 se formuló un modelo para determinar la eficiencia de la bolsa de Madrid. Para tener una primera impresión, el gráfico 6.4 muestra los residuos estandarizados[4] correspondientes a

[4] Los residuos estandarizados son igual a los residuos divididos por $\hat{\sigma}$.

la estimación de este modelo, utilizando el fichero *bolmadef*. El estadístico *DW* es igual a 2.04. (El estadístico *DW* aparece en la salida de cualquier paquete econométrico). Como las tablas publicadas no recogen los valores significativos para un tamaño de muestra de 247, utilizaremos los correspondientes a n=200 y k'=1. (En la nomenclatura de este contraste se utiliza k' para referirse al número total de regresores excluido el término independiente). Como el tamaño de la muestra es muy elevado utilizaremos un nivel de significación α=0.01, es decir del 1%. En la tabulación realizada por Durbin y Watson los valores inferior y superior, que corresponden a las anteriores especificaciones, son los siguientes:

$$d_L = 1.664 \quad ; \quad d_U = 1.684$$

Puesto que DW=2.04>d_U, se acepta la hipótesis nula de que las perturbaciones no están autocorrelacionadas, para un nivel de significación del 1%, frente a la hipótesis alternativa de autocorrelación positiva según el esquema **¡Error! No se encuentra el origen de la referencia.**.

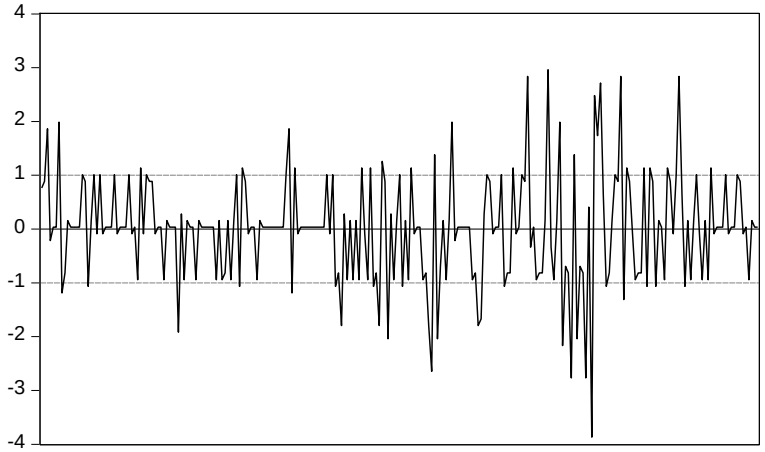


GRÁFICO 6.4. Residuos estandarizados en la estimación del modelo para determinar la eficiencia de la Bolsa de Madrid

EJEMPLO 6.12 Autocorrelación en el modelo sobre la demanda de pescado

En el ejemplo 4.9 se estimó el modelo (4-44), utilizando el fichero *fishdem*, para explicar la demanda de pescado en España. En el gráfico 6.5 se muestran los residuos estandarizados correspondientes a la estimación de este modelo. Del examen del gráfico no se desprende que exista un esquema de autocorrelación apreciable. En este sentido, conviene señalar que, sobre un total de 28 observaciones, la línea que une los puntos de los residuos cruza el eje 0 en 11 ocasiones, lo que es indicio de una cierta aleatoriedad de la distribución de los residuos.

El valor del estadístico *DW*, para el contraste del esquema **¡Error! No se encuentra el origen de la referencia.**, es 1.202. Para n=28 y k'=3, y para un nivel de significación del 1%, se obtienen los siguientes valores en la tabla tabulada por Durbin y Watson:

$$d_L = 0.969 \quad ; \quad d_U = 1.415$$

Dado que d_L<1.202<d_U, no hay evidencias suficientes ni para aceptar la hipótesis nula, ni para rechazarla.

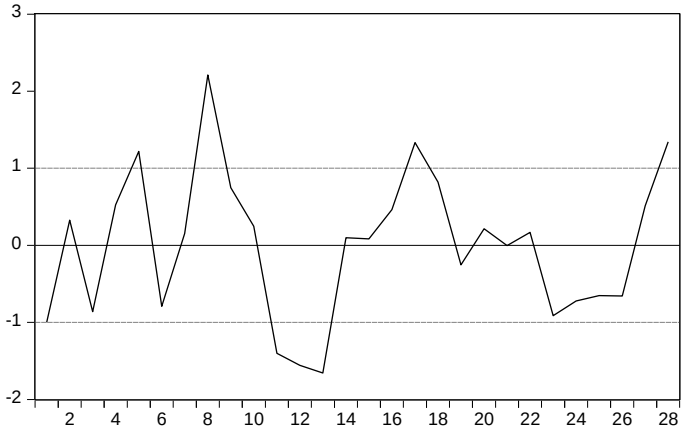


GRÁFICO 6.5. Residuos estandarizados en la estimación del modelo de demanda de pescado

Contraste h de Durbin

Durbin propuso en 1970 un estadístico, al que denominó *h*, para contrastar las hipótesis (6-54) en el caso de que haya una o más variables endógenas desfasadas, que aparezcan como variables explicativas del modelo. La expresión del estadístico *h* es la siguiente:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n\text{var}\ \hat{\beta}_j}} \tag{6-60}$$

donde $\hat{\rho}$ es el coeficiente de correlación entre $\hat{u}_i$ y $\hat{u}_{i-1}$, n es el tamaño de la muestra, y var $\hat{\beta}_j$ es la varianza correspondiente al coeficiente de la variable endógena desfasada.

El estadístico $\hat{\rho}$ puede estimarse utilizando la siguiente aproximación: $DW = d \simeq 2(1-\hat{\rho})$. En el caso de que aparezcan como regresores la variable endógena con distintos desfases se seleccionará la varianza correspondiente al coeficiente de la variable endógena con menor desfase.

Bajo los supuestos (6-54), el estadístico *h* tiene la distribución:

$$h \xrightarrow[n\to\infty]{} N(0,1) \tag{6-61}$$

La región crítica se encuentra, pues, en las colas de la distribución normal: en la cola de la derecha para la autocorrelación positiva y en la cola de la izquierda para la autocorrelación negativa.

El contraste (6-60) no se puede calcular cuando $n\text{var}\ \hat{\beta}_j \geq 1$. En ese caso Durbin propone como alternativa estimar una regresión auxiliar, en la que se toma como regresando los residuos mínimo cuadráticos y como regresores los mismos del modelo original y, además, los residuos desfasados un periodo. Si el coeficiente correspondiente a los residuos desfasados no fuera significativo, se rechaza la hipótesis alternativa.

EJEMPLO 6.13 Autocorrelación en el caso de Lydia E. Pinkham

En el ejemplo 5.5 se examinó el caso Lydia E. Pinkham en el que se estimó un modelo para explicar las ventas de un extracto herbal, utilizando el fichero *pinkham*. Con objeto de tener una primera impresión, en el gráfico 6.6 se muestra el gráfico de los residuos estandarizados de este modelo. Como

puede observarse, no parece que los residuos se distribuyan de forma totalmente aleatoria, ya que, por ejemplo, a partir de 1936 los residuos toman valores positivos durante 8 años consecutivos.

El contraste de autocorrelación apropiado para este modelo es el estadístico *h* de Durbin, debido a la presencia de la variable endógena desfasada $sales_{t-1}$. El estadístico *h* es:

$$h = \hat{\rho}\sqrt{\frac{n}{1-n\text{var}\ \hat{\beta}_j}} = \left[1-\frac{d}{2}\right]\sqrt{\frac{n}{1-n\text{var}\ \hat{\beta}_j}} = \left[1-\frac{1.2012}{2}\right]\sqrt{\frac{53}{1-53\times 0.0814^2}} = 3.61$$

Dado este valor de *h*, se rechaza la hipótesis nula de no autocorrelación, ya que la hipótesis nula se rechaza para α=0.01 e, incluso, para α=0.001, de acuerdo a la tabla de la normal.

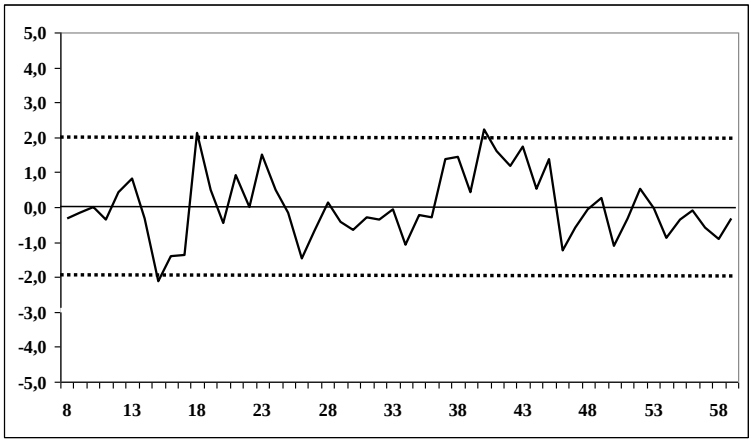


GRÁFICO 6.6. Residuos estandarizados en la estimación del modelo del caso Lydia E. Pinkham

Contraste de Breusch–Godfrey (BG)

El contraste de Breusch–Godfrey (1978) es un contraste general de autocorrelación aplicable a esquemas autorregresivos de un orden superior, y puede utilizarse cuando hay regresores estocásticos tales como el regresando retardado. Este es un contraste asintótico al que también se conoce como el contraste general de *ML* (multiplicadores de Lagrange) para autocorrelación.

En el contraste *BG* se asume que las perturbaciones u_t siguen un proceso autorregresivo de orden *p*, *AR*(*p*):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to NID(0,\sigma^2) \qquad (6\text{-}62)$$

Este es simplemente una extensión del esquema *AR*(1) del contraste de Durbin y Watson.

Las hipótesis nula y alternativa a contrastar son:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$
$$H_1 : H_0 \text{ no es cierto}$$

El contraste *BG* implica los siguientes pasos:

Paso 1. Se estima el modelo original y se calculan los residuos por *MCO* ($\hat{u}_i$).

Paso 2. Se estima la regresión auxiliar en la que se toma como regresando a los residuos ($\hat{u}_i$) y como regresores a los regresores del modelo original y los residuos retardados 1, 2, .. y *p* periodos:

$$\hat{u}_t = \alpha_1 + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + \gamma_1 \hat{u}_{t-1} + \cdots + \gamma_1 \hat{u}_{t-p} + \varepsilon_i \qquad (6\text{-}63)$$

La regresión auxiliar debería tener un término independiente, aunque el modelo original no lo tuviera. De acuerdo con (6-63), en la regresión auxiliar hay *k+p* regresores además del término independiente.

Paso 3. Designando por R^2_{ar} al coeficiente de determinación de la regresión auxiliar, se calcula el estadístico nR^2_{ar}.

Bajo la hipótesis nula, el estadístico *BG* se distribuye del siguiente modo:

$$BG = nR^2_{ar} \xrightarrow[n\to\infty]{} \chi^2_{k+p} \qquad (6\text{-}64)$$

El estadístico *BG* se utiliza para realizar un contraste global del modelo (6-63). Para este propósito se puede utilizar también el estadístico *F*, aunque en este caso solo tiene validez asintótica, como ocurre con el estadístico *BG*.

Paso 4. Para un nivel de significación α, y designando por $\chi^{2(\alpha)}_{k+p}$ al correspondiente valor en la tabla χ^2, la decisión a tomar es la siguiente:

Si $BG > \chi^{2(\alpha)}_{k+p}$ Se rechaza H_0

Si $BG \leq \chi^{2(\alpha)}_{k+p}$ No se rechaza H_0

Como un caso particular el contraste *BG* puede aplicarse a datos trimestrales utilizando un esquema *AR*(4).

EJEMPLO 6.14 Autocorrelación en un modelo para explicar los gastos de los residentes en el extranjero

Para explicar los gastos de los residentes en el extranjero (*turimp*), se estimó el siguiente modelo utilizando datos trimestrales de la economía española (archivo *qnatacsp*):

$$\ln(turimp_t) = \underset{(3.43)}{-17.31} + \underset{(0.276)}{2.0155}\ln(gdp_t)$$

R^2=0.531 *DW*=2.055 *n*=49

donde *gdp* es el producto interior bruto.

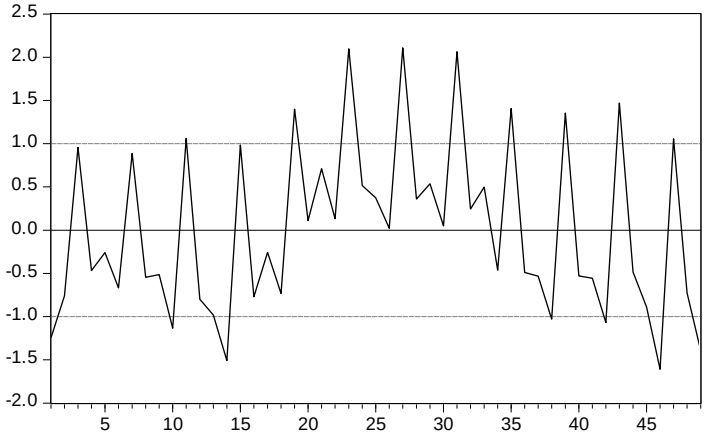


GRÁFICO 6.7. Residuos estandarizados en el modelo para explicar los gastos de los residentes en el extranjero.

El gráfico 6.7 muestra los residuos estandarizados correspondientes a este modelo. Como puede verse, parece que los residuos no están distribuidos de forma aleatoria, porque por ejemplo, se observan picos cada 4 trimestres, lo que es un indicativo de que la autocorrelación sigue un esquema *AR*(4).

El estadístico *BG*, calculado para un esquema *AR*(4), es igual a nR_{ar}^2 =36.35. Dado este valor de *BG*, se rechaza la hipótesis de no autocorrelación para α=0.01, ya que $\chi_5^{2(\alpha)}$ =15.09. En la regresión auxiliar, en la que se han utilizado como regresores $\hat{u}_{t-1}, \hat{u}_{t-2}, \hat{u}_{t-3}$ y $\hat{u}_{t-4}$, el único que ha resultado significativo ha sido $\hat{u}_{t-4}$.

6.6.4 Errores estándar *HAC*

Como una extensión de los errores estándar consistentes para heteroscedasticidad de White, examinados en la sección 6.5.2, Newey y West propusieron un método conocido como errores estándar *HAC* (heteroskedasticity and autocorrelation consistent) que permiten corregir los errores estándar de *MCO* no solamente en situaciones de autocorrelación sino también en caso de heteroscedasticidad. Recuerde que el procedimiento de White fue diseñado específicamente para heteroscedasticidad. Es importante resaltar que el procedimiento Newey y West es válido, estrictamente hablando, para grandes muestras y puede no ser apropiado para pequeñas muestras. Puede considerarse que un tamaño de 50 observaciones es razonablemente grande.

EJEMPLO 6.15 Errores estándar HAC en el caso de Lydia E. Pinkham (Continuación del ejemplo 6.13)

Dada la existencia de autocorrelación en el modelo del caso Lydia E. Pinkham, se han calculado los errores estándar de acuerdo con el procedimiento de Newey y West, lo que permitirá realizar correctamente contrastes de hipótesis sobre los parámetros. En el cuadro 6.9 aparecen los estadísticos *t* obtenidos por el procedimiento convencional y por el procedimiento *HAC*, así como la ratio entre ambos. Como puede verse las *t* obtenidas por el procedimiento *HAC* son ligeramente inferiores a las obtenidas por el método convencional, con la excepción del coeficiente de *advexp*, cuya *t* sorprendentemente es mucho mayor cuando se aplica el procedimiento *HAC*. En cualquier caso, al realizar contrates de significatividad de cada uno de los parámetros se obtienen exactamente las mismas conclusiones por ambos procedimientos para los niveles de significación de 0.1, 0.05 y 0.01.

CUADRO 6.9. Estadísticos *t*, convencional y *HAC*, en el caso de Lydia E. Pinkham.

regresor	*t* convencional	*t HAC*	ratio
intercept	2.644007	1.779151	1.49
advexp	3.928965	5.723763	0.69
sales(-1)	7.45915	6.9457	1.07
*d*1	-1.499025	-1.502571	1.00
*d*2	3.225871	2.274312	1.42
*d*3	-3.019932	-2.658912	1.14

6.6.5 Tratamiento de la autocorrelación

Para realizar la estimación de un modelo econométrico, donde las perturbaciones siguen el esquema *AR*(1) vamos a considerar en primer lugar el caso en que ρ es conocido. Este es más bien un supuesto académico que no se presenta en la realidad, pero que es conveniente adoptarlo como supuesto inicial a efectos de exposición. Sea el siguiente modelo de regresión lineal múltiple:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t \tag{6-65}$$

Si en (6-65) se considera un desfase y se multiplican ambos miembros por ρ, se obtiene que

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{2,t-1} + \rho\beta_3 x_{3,t-1} + \cdots + \rho\beta_k x_{k,t-1} + \rho u_{t-1} \tag{6-66}$$

Restando (6-66) de (6-65) se obtiene lo siguiente:

$$y_t - \rho y_{t-1} = \beta_{1,1}(1-\rho) + \beta_2\left(x_{2,t} - \rho x_{2,t-1}\right) + \cdots + \beta_k\left(x_{k,t} - \rho x_{k,t-1}\right) + \left(u_t - \rho u_{t-1}\right) \quad (6\text{-}67)$$

Como puede verse, de acuerdo con el esquema dado en (6-53), el término de perturbación de (6-67) cumple con los supuestos del *MLC*.

El modelo (6-67) se puede estimar directamente por mínimos cuadrados si se conoce el valor de ρ. Los estimadores obtenidos se aproximan al método de M*CG* si la muestra es suficientemente grande. Estrictamente hablando el método de MCG, consiste en transformar las observaciones 2 a *n* según el esquema (6-67) y, además, en transformar la primera observación de la siguiente forma:

$$y_t\sqrt{1-\rho^2} = \beta_1\sqrt{1-\rho^2} + \beta_2\sqrt{1-\rho^2}x_{2t} + \cdots + \beta_k\sqrt{1-\rho^2}x_{kt} + \varepsilon_t \quad (6\text{-}68)$$

Cuando se estima ρ conjuntamente con el resto de los parámetros del modelo, entonces al método de *MCG* se le denomina *MCG factibles*.

En general, en los diferentes métodos para aplicar *MCG* factibles se hace caso omiso de la transformación de la primera observación realizada en (6-68). Los métodos de *MCG* factibles para la estimación de un modelo en que las perturbaciones siguen un esquema *AR*(1) se pueden agrupar en tres bloques: a) métodos en dos etapas; b) métodos iterativos; y c) métodos de rastreo.

A continuación, vamos a exponer dos métodos correspondientes al bloque a), denominados método directo y método de Durbin en dos etapas.

En la primera etapa del método directo y en el método propuesto por Durbin se procede a estimar ρ. En el método directo, ρ se estima fácilmente a partir del estadístico DW, utilizando la aproximación $DW \simeq 2(1-\hat{\rho})$. En el método de Durbin en dos etapas se estima el siguiente modelo de regresión en el que las variables explicativas son los regresores del modelo original, los regresores desfasados un periodo y la variable endógena desfasada un periodo:

$$y_t = \alpha_1 + \alpha_{2,0}x_{2t} + \alpha_{2,1}x_{2,t-1} + \cdots + \alpha_{k0}x_{kt} + \alpha_{k1}x_{k,t-1} + \rho y_{t-1} + \upsilon_t \quad (6\text{-}69)$$

El coeficiente de la variable endógena desfasada es precisamente el parámetro ρ. En la primera etapa, se estima el modelo (6-69) por *MCO*, tomando del mismo la estimación de ρ. En la segunda etapa, aplicable a los dos métodos, se transforma el modelo con la estimación de ρ calculada en la primera etapa de la siguiente forma:

$$y_t - \hat{\rho}y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2\left(x_{2t} - \hat{\rho}x_{2,t-1}\right) + \cdots + \beta_k\left(x_{kt} - \hat{\rho}x_{k,t-1}\right) + \xi_t \quad (6\text{-}70)$$

Aplicando *MCO* al modelo transformado se obtienen las estimaciones de los parámetros. Una exposición de los métodos iterativos y de rastreo puede verse en Uriel, E.; Contreras, D.; Moltó, M. L. y Peiró, A. (1990): *Econometría. El modelo lineal*. Editorial AC. Madrid.

Ejercicios

Ejercicio 6.1 Consideremos el siguiente modelo poblacional**:**

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad (1)$$

En su lugar, se estimó el siguiente modelo estimado:

$$\tilde{y}_i = \tilde{\beta}_2 x_{2i} \tag{2}$$

¿Es $\tilde{\beta}_2$, obtenido al aplicar *MCO* a (2), un estimador insesgado de β_2?

Ejercicio 6.2 Consideremos el siguiente modelo poblacional:

$$y_i = \beta_2 x_i + u_i \tag{1}$$

En su lugar, se estimó el siguiente modelo estimado:

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \tag{2}$$

¿Es $\tilde{\beta}_2$, obtenido al aplicar *MCO* a (2), un estimador insesgado de β_2?

Ejercicio 6.3 Sean los siguientes modelos:

$$imp = \beta_1 + \beta_2 gdp + \beta_3 rpimp + u \tag{1}$$

$$\ln(imp) = \beta_1 + \beta_2 \ln(gdp) + \beta_3 \ln(rpimp) + u \tag{2}$$

donde *imp* es la importación de bienes, *gdp* es el producto interior bruto a precios de mercado, y *rpimp* son los precios relativos importaciones/pib. Las magnitudes *imp* y *gdp* están expresadas en millones de pesetas.

a) Utilizando una muestral del periodo 1971-1997 para España (archivo *importsp*), estime los modelos (1) y (2).
b) Interprete los coeficientes β_2 y β_3 en ambos modelos.
c) Aplique el procedimiento RESET al modelo (1).
d) Aplique el procedimiento RESET al modelo (2).
e) Utilice la especificación más adecuada usando los valores *p* obtenidos en las secciones *b)* y *c).*

Ejercicio 6.4 Considere el siguiente modelo de demanda de alimentos

$$alim = \beta_1 + \beta_2 pr + \beta_3 renta + u$$

donde *alim* es el gasto en alimentos, *pr* son los precios relativos y *renta* es la renta disponible.

El investigador A omite por olvido la variable *renta*, obteniendo la siguiente estimación del modelo:

$$\underset{}{alim_i} = \underset{(11.85)}{89.97} + \underset{(0.118)}{0.107}\ pr_i$$

El investigador B, que es más cuidadoso, obtiene la siguiente estimación del modelo:

$$\underset{}{alim_i} = \underset{(5.84)}{92.05} - \underset{(0.067)}{0.142}\ pr_i + \underset{(0.031)}{0.236}\ renta_i$$

(Entre paréntesis figuran desviaciones típicas)

A lo largo de la discusión entre ambos investigadores acerca de cuál de los dos modelos estimados es el más adecuado, el investigador A trata de justificar su olvido, atribuyendo la omisión de la variable *renta* al problema de la multicolinealidad.

a) En favor de cuál de los investigadores se inclinaría usted, a la vista de los resultados obtenidos. Argumente razonadamente su punto de vista.

b) Obtenga analíticamente la expresión del sesgo de estimación del estimador del parámetro β_2 en el modelo con error de especificación por omisión de variable relevante.

Ejercicio 6.5 Para estimar una función de producción se ha formulado el siguiente modelo

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u$$

donde *output* es la cantidad de *output* producido, *labor* es la cantidad de mano de obra, y *capital* es la cantidad de capital

Se dispone de las siguientes observaciones correspondientes a 9 empresas:

$output_i$	230	140	180	270	300	240	230	350	120
$labor_i$	30	10	20	40	50	20	30	60	40
$capital_i$	160	50	100	200	240	190	160	300	150

Un investigador estima el modelo tomando equivocadamente sólo 8 observaciones, y obtiene los siguientes resultados:

$$\underset{}{output_i} = \underset{(1.956)}{97.259} + \underset{(0.124)}{0.970}\, labor_i + \underset{(0.027)}{0.650}\, capital_i$$

$$R^2 = 0.999; \qquad F=3422$$

Los valores entre paréntesis son los errores estándar de los estimadores y el estadístico *F* corresponde al contraste global del modelo.

Cuando se da cuenta del error cometido, estima el modelo con todas las observaciones (n=9), obteniendo en este caso los siguientes resultados:

$$\underset{}{output_i} = \underset{(32.046)}{75.479} - \underset{(1.742)}{1.970}\, labor_i + \underset{(0.376)}{1.272}\, capital_i$$

$$R^2 = 0.824 \qquad F= 14.056$$

Su desconcierto es grande al comparar ambas estimaciones, y no puede comprender cómo, por utilizar una sola observación más, los resultados obtenidos

llegan a ser tan diferentes. ¿Puede encontrar alguna explicación que pueda justificar estas diferencias?

Ejercicio 6.6 Supongamos que en el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

el *R* cuadrado obtenido en la regresión de x_1 sobre x_2, al que denominaremos $R^2_{1/2}$, es cero.

Por otra parte, si estima los siguientes modelos:

$$y = \lambda_0 + \lambda_1 x_1 + u$$

$$y = \gamma_0 + \gamma_1 x_2 + u$$

a) ¿Será $\hat{\lambda}_1$ igual a $\hat{\beta}_1$ y $\hat{\gamma}_1$ igual a $\hat{\beta}_2$?

b) ¿Será $\hat{\beta}_0$ igual a $\hat{\lambda}_0$ o $\hat{\beta}_0$ igual a $\hat{\gamma}_0$?

c) ¿Será var($\hat{\lambda}_1$) igual a var($\hat{\beta}_1$) y var($\hat{\gamma}_1$) igual a var($\hat{\beta}_2$)?

Ejercicio 6.7 Un analista desea estimar el siguiente modelo utilizando las observaciones del cuadro adjunto:

$$y_i = e^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} x_{4i}^{\beta_4} e^{u_i}$$

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

¿Qué problemas se pueden presentar en la estimación de este modelo con estos datos?

Ejercicio 6.8 En el ejercicio 4.8, utilizando el fichero *airqualy*, se estimó el siguiente modelo:

$$\begin{aligned} airqual_i = {} & \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i \\ & - \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i \end{aligned}$$

$$R^2 = 0.415 \qquad n = 30$$

a) Calcule el estadístico *FAV* para cada coeficiente.

b) ¿Cuál es su conclusión?

Ejercicio 6.9 Para examinar los efectos de los rendimientos de la empresa sobre los salarios de los directores ejecutivos se ha formulado el siguiente modelo:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + \beta_6 age + u$$

donde *roa* es la ratio beneficios/activos expresados en porcentaje, *tenure* es el número de años como consejero delegado en la empresa (=0 si es menos de 6 meses), y *age* es la

edad en años. Los salarios están expresados en miles de dólares, y *sales* (vendes) y *profits* (beneficios) en millones de dólares.

a) Utilizando una muestra de 447 observaciones del fichero *ceoforbes*, estime el modelo por *MCO*.

b) Aplique el contraste de normalidad a los residuos.

c) Utilizando las 60 primeras observaciones, estime el modelo por *MCO*. Compare los coeficientes y el R^2 de esta estimación con los obtenidos en el apartado *a)*. ¿Cuál es su conclusión?

d) Aplique el contraste de normalidad a los residuos obtenidos en el apartado *c)*. ¿Cuál es su conclusión al comparar este resultado con el obtenido en el apartado *b)*?

Ejercicio 6.10 Sea el modelo

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

siendo

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Aplíque *MCG* al modelo [1] para estimar β_i.

b) Calcule la varianza del estimador por *MCG*:

Ejercicio 6.11 Sea el modelo

$$y_i = \beta x_i + u_i \qquad [1]$$

donde

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Estime β del modelo [1] por mínimos cuadrados generalizados.

b) Calcule la varianza del estimador obtenido.

Ejercicio 6.12 Sea el modelo

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

donde la varianza de las perturbaciones es igual a

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

1) Aplicando *MCO* al modelo [1] y teniendo en cuenta los supuestos Gauss-Markov, la varianza del estimador, de acuerdo con (2-16) es

$$\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad [2]$$

2) Aplicando *MCO* al modelo [1] y teniendo en cuenta que $\sigma_i^2 = \sigma^2 x_i$ y los restantes supuestos Gauss-Markov, la varianza del estimador es entonces igual a

$$\frac{\sigma^2 \sum (x_i - \bar{x})^2 x_i}{(\sum (x_i - \bar{x})^2)^2} \qquad [3]$$

3) Aplicando *MCG* al modelo [1] y teniendo en cuenta que $\sigma_i^2 = \sigma^2 x_i$, y los restantes supuestos Gauss-Markov, la varianza del estimador es

$$\frac{\sigma^2}{\sum \frac{(x_i - \bar{x})^2}{x_i}} \qquad [4]$$

a) ¿Son correctas las varianzas [2] y [3]?

b) Demuestre que [4] es menor o igual que [3]. (*Sugerencia*: Aplique la desigualdad Cauchy-Schwarz que dice que $\left[\sum w_i z_i\right]^2 \leq \left[\sum w_i^2\right]\left[\sum z_i^2\right]$ es verdad)

Ejercicio 6.13 Sea el modelo

$$hostel = \alpha_1 + \alpha_2 renta + u$$

donde *hostel* es el gasto en hostelería y *renta* es la renta anual disponible

Se dispone de la siguiente información sobre 9:

familia	*hostel*	*renta*
1	13	300
2	3	200
3	38	700
4	47	900
5	14	400
6	18	500
7	25	800
8	1	100
9	21	600

Las variables *hostel* y *renta* están expresadas en miles de pesetas.

a) Estime el modelo por *MCO.*

b) Aplique el contraste de heteroscedasticidad de White.

c) Aplique el contraste de heteroscedasticidad de Breusch-Pagan-Godfrey.

d) ¿Le aparece adecuado utilizar los anteriores contrastes de heteroscedasticidad en este caso?

Ejercicio 6.14 Con referencia al modelo del ejercicio 4.5, se supone ahora que

$$\text{var}(\varepsilon_i) = \sigma^2 \ln(y_i)$$

a) ¿Son, en este caso, insesgados los estimadores obtenidos por *MCO*?

b) ¿Son eficientes los estimadores *MCO*?

c) ¿Podría sugerir un estimador mejor que *MCO*?

Ejercicio 6.15 Indique cuáles de las siguientes afirmaciones son verdad, justificando las respuestas, cuando existe heteroscedasticidad:

a) Los estimadores *MCO* dejan de ser estimadores *ELI*O.

b) Los estimadores *MCO* $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ son inconsistentes.

c) Los contrastes convencionales *t* y *F* son no válidos.

Ejercicio 6.16 En el ejercicio 3.19, utilizando el archivo *consumsp*, se estimó el modelo de Brown para la economía española en el periodo 1954-2010. Los resultados obtenidos fueron los siguientes:

$$\underset{}{consp c_t} = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

$$R^2=0.997 \qquad SCR=1891320 \qquad n=56$$

Utilizando los residuos del anterior modelo ajustado, se obtuvo la siguiente regresión:

$$(\hat{u}_t^2) = 141568 + 89.71 incpc_t - 149.2 conspc_{t-1}$$
$$-0.183 incpc_t^2 - 0.221 conspc_{t-1}^2 + 0.406 incpc_t \times conspc_{t-1}$$
$$R^2=0.285$$

a) ¿Existe heteroscedasticidad en esta función de consumo?

b) Se obtuvo la siguiente estimación, con errores estándar consistentes para heteroscedasticidad de White:

$$\underset{}{conspc_t} = \underset{(66.92)}{?} + \underset{(0.0669)}{?}\, incpc_t + \underset{(0.0741)}{?}\, conspc_{t-1}$$

¿Puede rellenar los espacios con interrogante? Por favor, hágalo.
Explique la diferencia entre los errores estándar consistentes para la heteroscedasticidad de White y los errores estándar usuales.

c) Contraste si el coeficiente de *incpc* es igual a 5. ¿Qué errores estándar utilizaría en el proceso de inferencia? ¿Por qué?

Ejercicio 6.17 Suponga la siguiente especificación:

$$c_i = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i$$
$$\sigma_i^2 = \sigma^2 h_i^2$$

¿Sería adecuado para eliminar la heteroscedasticidad realizar la siguiente transformación del modelo

$$\frac{c_i}{h_i} = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i \text{ ?}$$

Razone su respuesta.

Ejercicio 6.18 Sea el modelo

$$y = \beta_1 + \beta_2 x + u$$

y se dispone de la siguiente información:

y_i	x_i	$\hat{u}_i$
2	-3	1.37
3	-2	-0.42
7	-1	0.79
6	0	-3.00
15	1	3.21
8	2	-6.58
22	3	4.63

a) Aplique el contraste de heteroscedasticidad de White.

b) Aplique el contraste de heteroscedasticidad de Breusch-Pagan-Godfrey.

c) ¿Por qué la significación obtenida en ambos contrastes es tan diferente?

Ejercicio 6.19 Responda a las siguientes preguntas

a) Explique detalladamente en qué consiste el problema de la heteroscedasticidad en el modelo de regresión lineal.

b) Ilustre brevemente el problema de la heteroscedasticidad con un ejemplo.

c) Proponga soluciones al problema de la heteroscedasticidad.

Ejercicio 6.20 Utilizando una muestra correspondiente a 17 regiones se han obtenido las siguientes estimaciones:

$$\hat{y}_i = -309.8 + 0.76z_i + 3.05h_i \qquad R^2 = 0.989$$

$$\hat{u}_i^2 = -1737.2 - 17.8z_i + 0.09z_i^2 + 0.65z_ih_i + 10.6h_i - 0.31h_i^2 \qquad R^2 = 0.705$$

donde *y* es el gasto en educación, *z* es el *PIB* y *h* es el número de habitantes.

a) ¿Existe un problema de heteroscedasticidad? Detalle el procedimiento de contraste.

b) Suponiendo que se detectara la presencia de heteroscedasticidad en el modelo de regresión, ¿qué solución adoptaría para analizar la significatividad de las variables explicativas del modelo? Razone la respuesta.

Ejercicio 6.21 Utilizando datos de la economía española para el periodo 1971-1997 (archivo *importsp*), se estimó el siguiente modelo para explicar las importaciones (*imp*):

$$\ln(imp_t) = \underset{(3.65)}{-26.58} + \underset{(0.210)}{2.4336}\ln(gdp_t) - \underset{(0.0232)}{0.4494}\ln(rpimp_t)$$

$$R^2=0.997 \quad n=27$$

donde *gdp* es el producto interior bruto a precios de mercado, y *rpimp* son los precios relativos importaciones/pib. Las variables *imp* y *gdp* están expresadas en millones de pesetas

a) Formule y estime la regresión auxiliar para realizar el contraste de heteroscedasticidad de Breusch-Pagan-Godfrey.

b) Aplique el contraste de heteroscedasticidad de Breusch-Pagan-Godfrey utilizando la regresión formulada en la sección *a)*.

c) Formule y estime la regresión auxiliar para realizar el contraste *completo* de White de heteroscedasticidad.

d) Aplique el contraste de heteroscedasticidad *completo* de White utilizando la regresión formulada en la sección *c)*.

e) Formule y estime la regresión auxiliar para realizar el contraste *simplificado* de heteroscedasticidad de White.

f) Aplique el contraste de heteroscedasticidad *simplificado* de White utilizando la regresión formulada en la sección *e)*.

g) Compare los resultados de los contrastes realizados en las secciones *b*), *d*) y *f*).

Ejercicio 6.22 Utilizando datos del archivo *tradocde*, se estimó el siguiente modelo para explicar las importaciones (*impor*) en los países de la OCDE:

$$\ln(impor_i) = \underset{(6.67)}{18.01} + \underset{(0.658)}{1.6425}\ln(gdp_i) - \underset{(0.636)}{0.5151}\ln(popul_i)$$

$$R^2=0.614 \qquad n=34$$

donde *gdp* es el producto interior bruto a precios de mercado, y *popul* es la población de cada país.

a) ¿Cuál es la interpretación del coeficiente de *gdp*?
b) Formule y estime la regresión auxiliar para realizar el contraste de White de heteroscedasticidad.
c) Aplique el contraste de heteroscedasticidad de White utilizando la regresión formulada en la sección *b)*.
d) Contraste si la elasticidad *import*/*gdp* es más grande que 1. Para realizar este contraste, ¿necesita utilizar los errores estándar consistentes para la heteroscedasticidad de White?

Ejercicio 6.23 Explique detalladamente cuál sería el contraste de autocorrelación apropiado en cada situación:

a) Cuando el modelo no tiene variables endógenas retardadas y las observaciones son anuales.
b) Cuando el modelo tiene variables endógenas retardadas y las observaciones son anuales.
c) Cuando el modelo no tiene variables endógenas retardadas y las observaciones son trimestrales.

Ejercicio 6.24 Se han estimado dos modelos alternativos del coste medio de producción anual de automóviles de una determinada marca en el periodo 1980-1999.

$$c = \alpha + \beta p + u \qquad R^2 = 0.848; \quad \bar{R}^2 = 0.812; \quad d = DW = 0.51$$

$$c = \alpha + \beta p + \gamma p^2 + u \qquad R^2 = 0.852; \quad \bar{R}^2 = 0.811; \quad d = DW = 2.11$$

a) Al comparar ambas estimaciones, indique si observa algún problema econométrico. Explíquelo.
b) En función de su respuesta al apartado anterior, ¿Cuál de los dos modelos elegiría?

Ejercicio 6.25 En el periodo 1950-1980 se ha estimado la siguiente función de producción

$$\ln(o_t) = \underset{}{-3.94} + \underset{(0.24)}{1.45}\ln(l_t) + \underset{(0.083)}{0.38}\ln(k_t) + \underset{(0.048)}{}$$

$$R^2 = 0.994 \qquad DW = 0.858 \qquad \hat{\rho} = 0.559$$

donde *o* es la producción, *l* es el trabajo, y *k* es el capital.

(Los números entre paréntesis son las desviaciones estándar de los estimadores).

a) Contraste detalladamente la existencia de autocorrelación.
b) Si el modelo tuviera una variable endógena retardada como variable explicativa indique de qué forma contrastaría la autocorrelación.

Ejercicio 6.26 Utilizando una muestra de 38 observaciones de periodicidad anual se ha estimado la siguiente función de demanda de un producto

$$d_i = 2.47 + \underset{(0.39)}{0.35}\, p_i + \underset{(0.06)}{0.9}\, d_{i-1} \qquad R^2 = 0.98 \qquad DW = 1.82$$

donde *d* es la cantidad demandada y *p* es el precio.

(Los números entre paréntesis son las desviaciones estándar de los estimadores).

a) ¿Existe un problema de autocorrelación? Razone la respuesta.
b) Enumere las condiciones bajo cuales sería adecuado utilizar el contraste de Durbin Watson.

Ejercicio 6.27 Se ha estimado el siguiente modelo de demanda de vivienda con observaciones anuales correspondientes al periodo 1960-1994:

$$\ln(v_t) = \underset{(0.15)}{-0.39} + \underset{(0.05)}{0.3}\ln(r_t) - \underset{(0.02)}{0.67}\ln(p_t) + \underset{(0.04)}{0.70}\ln(v_{t-1})$$

$$R^2 = 0.999 \qquad DW = 0.52$$

donde *v* es el gasto en vivienda, *r* es la renta disponible, *p* es el precio de la vivienda.

(Los números entre paréntesis son las desviaciones estándar de los estimadores).

a) Contraste detalladamente la existencia de autocorrelación.

b) Teniendo en cuenta las conclusiones obtenidas en al apartado *a*), como realizaría los contrastes de significatividad de cada uno de los coeficientes. Razone la respuesta.

Ejercicio 6.28 Conteste a las siguientes preguntas:

a) En un modelo para explicar las ventas se realiza la estimación utilizando datos trimestrales. Explique cómo puede contrastar si existe autocorrelación.

c) Describa detalladamente, introduciendo los supuestos que considere oportunos, cómo estimaría el modelo cuando se rechaza la hipótesis nula de no autocorrelación.

Ejercicio 6.29 En la estimación de la función de consumo keynesiana de la economía francesa se han obtenido los siguientes resultados:

$$consumo_t = \underset{(-0.73)}{-485.22} + \underset{(79.39)}{0.913}\,renta_t$$

$$R^2 = 0.9936 \quad \text{DW=0.4205} \qquad n\text{=}30$$

(Los números entre paréntesis son los estadísticos *t* de los estimadores).

Un investigador considera que se debe centrar la atención en la función de ahorro, en lugar de hacerlo en la función de consumo. En consecuencia, propone el siguiente modelo:

$$ahorro_t = \alpha_1 + \alpha_2 renta + v_t \qquad [1]$$

donde

$$ahorro_t = renta_t - consumo_t$$

Utilizando la información dada en el presente ejercicio, si ello es posible:

a) Obtenga las estimaciones de α_1 y α_2.

b) Estime las varianzas de $\hat{\alpha}_1$ y $\hat{\alpha}_2$.

c) Calcule el estadístico DW (Durbin-Watson) del modelo de ahorro.

d) Calcule el coeficiente R^2 para el modelo de ahorro.

Ejercicio 6.30 Sea el modelo

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t; \quad E\left[\varepsilon_t^2\right] = \sigma^2 \qquad \forall i \end{aligned} \qquad [1]$$

a) Si el modelo [1] se transforma tomando primeras diferencias, ¿bajo qué supuestos resulta ventajosa la estimación por *MCO* del modelo transformado con respecto a la estimación por *MCO* del modelo [1]?

b) ¿Es adecuado utilizar el R^2 para comparar el modelo [1] y el modelo transformado? Explique su respuesta.

Ejercicio 6.31 Sea el modelo

$$y_t = \beta_1 + \beta_2 x_t + u_t \qquad [1]$$

Se obtiene la siguiente muestra de observaciones para las variables *x* e *y*:

y_i	6	3	1	1	1	4	6	16	25	36	49	64
x_i	-4	-3	-2	-1	1	2	3	4	5	6	7	8

a) Estime el modelo [1] por *MCO* y calcule el correspondiente coeficiente de determinación corregido.
b) Calcule el estadístico de Durbin-Watson correspondiente a la estimación realizada en *a*).
c) A la vista del contraste de Durbin y Watson y de la representación de la recta ajustada y de los residuos, ¿es conveniente reformular el modelo [1]? Justifique la respuesta y, en caso de que ésta sea afirmativa, estime el modelo alternativo que se considere más adecuado a los datos.

Ejercicio 6.32 En el siguiente modelo:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t; \qquad \varepsilon_t \sim NI\ \ 0, \sigma^2$$

La siguiente información adicional está también disponible:

$$\rho = 0.5$$

y_i	22	26	32	31	40	46	46	50
x_i	4	6	10	12	13	16	20	22

a) Estime el modelo por *MCO*.
b) Estime el modelo por *MCG* sin la transformación de la primera observación.
c) ¿Cuál de los dos estimadores de β_2 es más eficiente?

Ejercicio 6.33 En un estudio sobre la demanda de un producto se han obtenido los siguientes resultados:

$$\hat{y}_t = \underset{(7.17)}{2.30} + \underset{(0.05)}{0.86}\, x_t$$

$$R^2 = 0.9687 \quad \text{DW=3.4} \quad n = 15$$

(Los números entre paréntesis son los errores estándar de los estimadores.)

Además, se dispone de la siguiente información adicional sobre las regresiones de los errores, en valor absoluto:

1. $\left|\hat{u}_t\right| = \underset{(0.210)}{0.167} + \underset{(0.180)}{0.127}\, x_t$
2. $\left|\hat{u}_t\right| = \underset{(0.098)}{0.231} + \underset{(0.095)}{0.218}\, x_t^{1/2}$

a) Detecte si existe autocorrelación.
b) Detecte si existe heteroscedasticidad.
c) ¿Cuál sería el procedimiento más adecuado para evitar el posible problema de heteroscedasticidad?

Ejercicio 6.34 Utilizando una muestra para el periodo 1971-1997 (archivo *importsp*), se estimó el siguiente modelo, utilizando errores estándar *HAC*, para explicar las importaciones de bienes en España (*imp*):

$$\ln(imp_t) = \underset{(3.65)}{-26.58} + \underset{(0.210)}{2.434}\ln(gdp_t) - \underset{(0.023)}{0.4494}\ln(rpimp_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.73} \quad n = 27$$

donde *gdp* es el producto interior bruto a precios de mercado, y *rpimp* son los precios relativos importaciones/pib. Ambas magnitudes están expresadas en millones de pesetas.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Interprete el coeficiente de *rpimp*.

b) ¿Hay autocorrelación en este modelo?

c) Contraste si la elasticidad *imp/gdp* más cuatro veces la elasticidad *imp/rpimp* es igual a 0. (Información adicional: $\text{var}(\hat{\beta}_2)$ =0.044247; $\text{var}(\hat{\beta}_3)$ =0.000540; y $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =0.004464).

d) Contraste la significación global.

Ejercicio 6.35 Utilizando una muestra para el periodo 1954-2009 (archivo *electsp*), se estimó el siguiente modelo para explicar el consumo de electricidad en España (*conselec*):

$$\ln(conselec_t) = \underset{(0.46)}{-9.98} + \underset{(0.035)}{1.469}\ln(gdp_t)$$

$$R^2 = 0.9805 \quad \text{DW=0.18} \quad n = 37 \qquad (1)$$

donde *gdp* es el producto interior bruto a precios de mercado. La variable *conselec* está expresada en miles de toneladas equivalentes de petróleo (*ktep*) y *gdp* está expresado en millones de pesetas.

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Contraste si hay autocorrelación mediante la aplicación del estadístico Durbin-Watson.

b) Contraste si hay autocorrelación mediante la aplicación del estadístico Breusch-Godfrey para un esquema *AR*(2).

c) También fue estimado el siguiente modelo:

$$\ln(conselec_t) = \underset{(0.75)}{-0.917} + \underset{(0.107)}{0.164}\ln(gdp_t) + \underset{(0.072)}{0.871}\ln(conselec_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.93} \quad n = 36 \qquad (2)$$

Contraste si hay autocorrelación mediante la aplicación del procedimiento que estime oportuno.

d) Contraste si la elasticidad *conselec/gdp* en una situación de equilibrio ($\ln(conselec^e) = \beta_1 + \beta_2 \ln(gdp^e) + \beta_3 \ln(conselec^e)$) es más grande que 1 utilizando un procedimiento adecuado.

Ejercicio 6.36 La curva de Phillips representa la relación entre la tasa de inflación (*inf*) y la tasa de desempleo (*unemp*). Mientras que a corto plazo se ha observado un *tradeoff* estable entre desempleo e inflación, este fenómeno no se ha constatado a largo plazo.

El siguiente modelo refleja la curva de Phillips estática:

$$inf = \beta_1 + \beta_2 unempl + u$$

Utilizando una muestra de la economía española para el periodo 1970-2010 (archivo *phillipsp*), se obtuvieron los siguientes resultados:

$$\underset{}{inf_t} = \underset{(1.79)}{12.59} - \underset{(0.120)}{0.3712}\, unempl_t$$

R^2=0.198; DW=0.219; n=41

(Los números entre paréntesis son los errores estándar de los estimadores.)

a) Interprete el coeficiente de *unempl*.
b) Contraste si hay autocorrelación de primer orden mediante la aplicación del estadístico Durbin-Watson.
c) Utilizando la información que tiene disponible hasta ahora, ¿puede contrastar de forma adecuada el coeficiente de *unempl*?
d) Utilizando los errores estándar *HAC*, contraste la significación del coeficiente de *unempl*.

Ejercicio 6.37 Es importante remarcar que la curva de Phillips es una relación relativa. La inflación es considerada alta o baja en relación a la tasa de inflación esperada y el desempleo es considerado alto o bajo en relación con la denominada tasa natural de desempleo. En la curva *aumentada* de Phillips todo esto se tiene en cuenta:

$$inf_t - inf^e_{t/t-1} = \beta_2 (unempl_t - \lambda_0) + u_t$$

donde λ_0 es la tasa natural de desempleo e $inf^e_{t/t-1}$ es la tasa de inflación esperada en t y formada en t-1. Si consideramos que tasa esperada para t es igual a la inflación en t-1 ($inf^e_{t/t-1} = inf_{t-1}$) y haciendo $\beta_1 = -\beta_2 \lambda_0$, la curva aumentada de Phillips puede expresarse así:

$$inf_t - inf_{t-1} = \beta_1 + \beta_2 unempl_t + u_t$$

a) Utilizando el archivo *phillipsp*, estime el modelo anterior.
b) Interprete el coeficiente de *unempl*.
c) Contraste si hay correlación de segundo orden.
d) Contraste si la tasa natural de desempleo es mayor que 10.

Apéndice 6.1

En primer lugar vamos a expresar el estimador $\tilde{\beta}_2$ teniendo en cuenta que *y* ha sido generada por el modelo (6-8):

$$\begin{aligned}\tilde{\beta}_2 &= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)y_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(\beta_1+\beta_2 x_{1i}+\beta_3 x_{2i}+u_i)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{1i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}\end{aligned} \tag{6-71}$$

Si tomamos esperanza en ambos miembros de (6-71), tenemos que

$$\begin{aligned}E(\tilde{\beta}_2) &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)E(u_i \mid x_2, x_3)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}\end{aligned} \tag{6-72}$$