

# Técnicas de previsión de demanda

Autor:

M. Victoria Esteban González

*Departamento de Economía Aplicada III. Econometría y Estadística  
Facultad de Ciencias Económicas y Empresariales  
Universidad del País Vasco/Euskal Herriko Unibertsitatea*

Queda terminantemente prohibida la reproducción no autorizada de este material docente, y la distribución no autorizada de copias de la misma, así como cualquier otra infracción de los derechos que sobre esta recopilación corresponden a la Profesora M<sup>a</sup> Victoria Esteban junto con el Departamento de Econometría y Estadística de la Facultad de Ciencias Económicas y Empresariales de la UPV/EHU.

©UPV/EHU 2018.





# Presentación

El objetivo de este documento es introducir un conjunto de técnicas estadísticas y econométricas para la estimación de modelos lineales en situaciones donde se cumplen las hipótesis estadísticas de comportamiento habituales. Se pretende introducir al alumno en el análisis de regresión, por lo que se estudia en detalle el Modelo de Regresión Lineal General. El objetivo fundamental del curso es que, al final del mismo, los estudiantes sean capaces de utilizar el modelo de regresión para resolver un problema sencillo que se les plantee: desde la especificación, estimación y validación del modelo hasta contrastar hipótesis de relevancia económica y predecir. Este objetivo se ha de satisfacer tanto desde un punto de vista teórico, resolver cuestiones y explicar resultados ya obtenidos, como práctico: estimar un modelo con una base de datos concreta y realizar los contrastes pertinentes.

Estas notas incluyen cinco temas. En el tema introductorio se incluye un anexo que sirve de recordatorio sobre cómo abordar la descripción gráfica y numérica de una variable. El segundo tema introduce la especificación del Modelo de Regresión Lineal General así como la nomenclatura y conceptos más habituales a manejar en el contexto del análisis de regresión. El tema tres aborda la estimación del modelo de regresión lineal general. El estimador de referencia es el estimador de Mínimos Cuadrados Ordinarios. Se estudiarán sus propiedades y cómo compararlo con otros estimadores de interés. El tema cuatro aborda el contraste de hipótesis, las consecuencias de la existencia de colinealidad o de mala especificación en el modelo para finalizar abordando la predicción de la variable de interés. El tema cinco muestra cómo analizar si alguna de las hipótesis estadísticas de comportamiento habituales no se cumplen y cuáles son las consecuencias de su incumplimiento.

A lo largo de los temas se va mostrando cómo utilizar un software libre, el programa *gretl*, especialmente indicado para el análisis econométrico y que permite un afianzamiento de los contenidos teóricos. Por ello, al final de los temas se incluye una sección que muestra cómo utilizar este programa en relación a los contenidos vistos. En cada tema se muestran ejemplos que ilustran los diferentes escenarios de trabajo así como se recomienda la realización de los ejercicios propuestos. Al término de cada tema se muestra la bibliografía correspondiente. Al final del documento aparece la bibliografía completa.

Las notas tienen como objetivo servir de apoyo al proceso de aprendizaje de los estudiantes de la asignatura *Técnicas de previsión de demanda* del Grado en Marketing. Sin embargo, dada su temática básica de estadística y análisis de regresión pueden ser útiles en asignaturas afines de los Grados en Economía, Administración y Dirección de Empresas, Fiscalidad y Administración Pública. Así mismo sirven de apoyo a estudiantes de master por ejemplo el Master en Ciencias Actariales y Financieras o el Master Universitario en Banca y Finanzas Cuantitativas.

## **Las competencias específicas de la asignatura y la evaluación**

La asignatura de Técnicas de previsión de demanda es una asignatura de 6 créditos ECTS que conlleva 60 horas de trabajo presencial en el aula y 90 horas de trabajo no presencial. La metodología y modalidades docentes a utilizar están sujetas al criterio del docente y pueden variar cada curso académico. Hay que tener en cuenta que la organización de la metodología docente junto con el diseño de los contenidos de los temas del curso van dirigidos a que los alumnos alcancen las siguientes competencias específicas de la asignatura:

- C1. Analizar de forma crítica los elementos básicos del modelo de regresión lineal con el objetivo de comprender la lógica de la modelización econométrica y poder especificar relaciones causales entre variables económicas.
- C2. Aplicar la metodología econométrica básica para estimar y validar relaciones económicas en base a la información estadística disponible sobre las variables y utilizando los instrumentos informáticos apropiados.
- C3. Interpretar razonadamente los resultados obtenidos en la estimación y validación del modelo econométrico con el objetivo de elaborar informes económicos.
- C4. Presentar de forma clara y concisa, tanto oralmente como por escrito, las conclusiones obtenidas en una aplicación empírica.

A lo largo del curso se trabajan las siguientes Competencias Transversales del módulo:

- CT1. Capacidad para emitir juicios razonados apoyándose en los datos obtenidos en el contexto de Investigación de Mercados.
- CT2. Desarrollar las habilidades de aprendizaje para adquirir un alto grado de autonomía, tanto de cara a emprender estudios posteriores como de cara a su propia autoformación en el contexto de Investigación de Mercados.
- CT3. Capacidad para la comunicación escrita y oral con fluidez en el contexto de Investigación de Mercados.
- CT4. Capacidad para el pensamiento analítico y la reflexión crítica en el contexto de Investigación de Mercados.

El sistema actual de docencia dentro del EEES tiene como ejes fundamentales el proceso de enseñanza-aprendizaje y la adquisición no sólo de conocimientos, sino también, y fundamentalmente, de destrezas implica directamente la valoración del trabajo diario del alumno y su evolución en la adquisición de las competencias. La utilización de la evaluación continua en la evaluación de los alumnos implica la realización, junto con otras pruebas y tareas que el docente crea de interés, de test rápidos o de preguntas cortas en relación a todo lo visto en las clases, conceptos teóricos y ejercicios prácticos incluido el software gretl que permitan evaluar al alumno y saber si han adquirido los resultados del aprendizaje alcanzando así las competencias específicas. Parte de las pruebas tendrán componente de sorpresa, es decir sin previo aviso, y parte serán pactadas en cuanto a fecha.

Como se indicaba anteriormente estas notas sirven de apoyo al estudio. Analizan los problemas en profundidad y permiten al alumno profundizar en los temas que conforman el contenido del curso. Así mismo tienen una fuerte vertiente práctica que permitirá al alumno no solo saber sino también saber hacer. En ningún caso deben utilizarse como sustituto de los libros incluidos en la bibliografía. De igual manera se recomienda la realización de ejercicios tanto los recomendados en clase como los que aparecen en la bibliografía. La unión del estudio de los conceptos y la utilización de los mismos en los ejercicios permite adquirir la agilidad necesaria para el dominio de la asignatura y alcanzar las competencias específicas de la misma.

## **Sobre el software gretl**

A lo largo del curso se muestra cómo utilizar un software gretl que permite al alumno un afianzamiento de los contenidos teóricos del curso de Econometría como la puesta en práctica de casos reales con la utilización del software gretl<sup>1</sup>.

**gretl** es software libre especialmente dirigido hacia la práctica de la econometría y la estadística, muy fácil de utilizar. Ha sido elaborado por Allin Cottrell (Universidad Wake Forest) y existen versiones en inglés, castellano y euskera, además de en otros idiomas. Junto con el programa se pueden cargar los datos utilizados como ejemplos de aplicaciones econométricas en los siguientes libros de texto Davidson y Mackinnon (2004), Greene (2008), Gujarati (1997), Ramanathan (2002), Stock y Watson (2003), Verbeek (2004), Wooldridge (2003). Al instalar gretl automáticamente se cargan los datos utilizados en Ramanathan (2002) y Greene (2008). El resto se pueden descargar de la página:

*[http://gretl.sourceforge.net/gretl\\_data.html](http://gretl.sourceforge.net/gretl_data.html)*

en la opción *textbook datasets*. Este curso se estructura sobre casos prácticos presentados en Ramanathan (2002) y en Wooldridge (2003) y ejercicios a resolver con ayuda de gretl.

También da acceso a bases de datos muy amplias, tanto de organismos públicos, como el Banco de España, como de ejemplos recogidos en textos de Econometría. En la página

*[http://gretl.sourceforge.net/gretl\\_espanol.html](http://gretl.sourceforge.net/gretl_espanol.html)*

se encuentra la información en castellano relativa a la instalación y manejo del programa. También hay versiones de esta ayuda en euskera y en inglés.

Una página web interesante sobre las posibilidades del programa para el aprendizaje de Econometría es:

*<http://www.learn econometrics.com/gretl.html>*

---

<sup>1</sup>Acrónimo de *Gnu Regression, Econometric and Time Series* (Biblioteca Gnu de Regresión Econometría y Series Temporales)



# Contenido

<b>1. Introducción a la Econometría</b>	<b>1</b>
1.1. Modelo económico y modelo econométrico . . . . .	3
1.2. Etapas en la elaboración de un modelo econométrico . . . . .	5
1.3. Tipología de datos y variables en Econometría . . . . .	6
1.3.1. Conceptos básicos . . . . .	7
1.3.2. Fuentes de datos . . . . .	10
1.4. Tratamiento de la información con <i>gretl</i> : inclusión de datos en <i>gretl</i> y análisis descriptivo básico . . . . .	11
1.5. Apéndice . . . . .	17
1.5.1. La naturaleza de los datos . . . . .	17
1.5.2. Clasificación de las variables . . . . .	18
1.5.3. Representación gráfica de las variables . . . . .	19
1.5.4. Descripción numérica de los datos . . . . .	24
1.6. Bibliografía del tema . . . . .	26
<b>2. Modelo de Regresión Lineal General: Especificación</b>	<b>29</b>
2.1. Especificación del Modelo de Regresión Lineal General (MRLG): supuestos básicos . . . . .	31
2.1.1. Hipótesis básicas. . . . .	34
2.2. Función de Regresión Poblacional. Interpretación de los coeficientes. . . . .	37
2.2.1. Forma funcional . . . . .	44
2.3. Utilización de variables explicativas cualitativas . . . . .	46
2.3.1. Modelo que recoge sólo efectos cualitativos: comparando medias. Sólo un conjunto de variables ficticias. . . . .	47
2.3.2. Dos o más conjuntos de variables ficticias . . . . .	49
2.3.3. Inclusión de variables cuantitativas . . . . .	50

2.3.4.	Comportamiento estacional . . . . .	51
2.3.5.	Efectos de interacción . . . . .	51
2.4.	Bibliografía del tema . . . . .	53
<b>3.</b>	<b>Modelo de Regresión Lineal General: Estimación</b>	<b>55</b>
3.1.	Estimación por Mínimos Cuadrados Ordinarios (MCO) . . . . .	57
3.1.1.	Método de estimación de Mínimos Cuadrados Ordinarios (MCO) . . . . .	58
3.1.2.	Propiedades de la Función de Regresión Muestral, FRM . . . . .	66
3.1.3.	Medidas de bondad del ajuste . . . . .	67
3.2.	Propiedades de los estimadores MCO . . . . .	71
3.2.1.	Propiedades de los estimadores MCO . . . . .	71
3.2.2.	Estimación de la varianza de las perturbaciones . . . . .	73
3.3.	Estimación del MRLG con <i>gretl</i> : principales resultados . . . . .	80
3.3.1.	Tratamiento de las variables ficticias en <i>gretl</i> . . . . .	82
3.4.	Apéndice . . . . .	87
3.4.1.	Anexo 1. Distintas expresiones de SCT, SCE y SCR . . . . .	87
3.4.2.	Anexo 2. Demostración de la insesgadez de $\hat{\sigma}^2$ . . . . .	88
3.5.	Bibliografía del tema . . . . .	90
<b>4.</b>	<b>Modelo de Regresión Lineal General: Inferencia</b>	<b>91</b>
4.1.	Distribución del estimador MCO. Estimación por intervalo . . . . .	93
4.1.1.	Distribución del estimador de MCO bajo Normalidad . . . . .	93
4.1.2.	Estimación por intervalo . . . . .	93
4.2.	Contraste de hipótesis sobre los coeficientes de la regresión . . . . .	94
4.2.1.	Contraste de restricciones sobre los coeficientes de regresión individuales . . . . .	95
4.2.2.	Contraste de restricciones sobre los coeficientes de regresión . . . . .	97
4.2.3.	Contrastes basados en sumas de cuadrados de residuos . . . . .	103
4.3.	Consecuencias del incumplimiento de algunos supuestos: colinealidad . . . . .	106
4.3.1.	Multicolinealidad exacta . . . . .	107
4.3.2.	Alta colinealidad . . . . .	108
4.4.	Consecuencias del incumplimiento de algunos supuestos: omisión de variables relevantes e inclusión de variables irrelevantes . . . . .	111
4.4.1.	Omisión de variables relevantes . . . . .	111

4.4.2. Inclusión de variables irrelevantes . . . . .	112
4.5. Predicción . . . . .	112
4.6. Contraste de hipótesis y predicción en el MRLG con <i>gretl</i> . . . . .	124
4.6.1. Predicción en <i>gretl</i> . . . . .	132
4.7. Apéndice . . . . .	134
4.7.1. Distribuciones que nos interesan . . . . .	134
4.8. Bibliografía del tema . . . . .	136
<b>5. Modelo de Regresión Lineal General. Diagnóstico y utilización</b>	<b>137</b>
5.1. Forma funcional . . . . .	139
5.2. Sobre constancia de los coeficientes: contraste de cambio estructural . . . . .	141
5.3. Sobre las perturbaciones: contrastes de heterocedasticidad y ausencia de correlación	142
5.3.1. Contraste de heterocedasticidad . . . . .	142
5.3.2. Detección gráfica. . . . .	146
5.3.3. Contraste de White . . . . .	149
5.3.4. Contraste de ausencia de correlación temporal . . . . .	150
5.4. Validación en <i>gretl</i> . . . . .	158
5.4.1. Contraste de Ramsey con <i>gretl</i> . . . . .	158
5.4.2. Contraste de cambio estructural o Chow con <i>gretl</i> . . . . .	159
5.4.3. Contraste de heterocedasticidad con <i>gretl</i> . . . . .	162
5.4.4. Contraste de ausencia de correlación con <i>gretl</i> . . . . .	165
5.5. Bibliografía del tema . . . . .	169
<b>Bibliografía</b>	<b>171</b>



# Figuras

1.1. Gráficos de las observaciones para las variables <i>price</i> y <i>sqft</i> . . . . .	16
1.2. N° de aprobados en tres asignaturas de la Licenciatura en Economía. . . . .	20
1.3. Distribución de frecuencias del gasto en sanidad . . . . .	21
1.4. Distribución de frecuencias relativas y evolución temporal del la demanda doméstica de electricidad . . . . .	23
1.5. Evolución temporal del consumo y renta per cápita junto a la nube de puntos . . . . .	24
2.1. Perturbaciones homocedásticas versus heterocedásticas . . . . .	34
2.2. Relaciones económicas no lineales . . . . .	45
3.1. Función de regresión poblacional y función de regresión muestral . . . . .	58
5.1. Relaciones económicas no lineales . . . . .	139
5.2. Perturbaciones homocedásticas versus heterocedásticas . . . . .	143
5.3. Residuos MCO versus <i>POP</i> . . . . .	147
5.4. Residuos MCO versus <i>POP</i> . . . . .	147
5.5. Residuos MCO y sus cuadrados versus <i>SEN</i> . . . . .	148
5.6. Perturbaciones homocedásticas . . . . .	148
5.7. Residuos MCO frente a una variable ficticia . . . . .	149
5.8. Proceso autorregresivo de orden uno . . . . .	154
5.9. Perturbaciones AR(1) positivo versus AR(1) negativo . . . . .	155
5.10. Variable endógena versus exógena . . . . .	159
5.11. Gasto sanitario real y ajustado . . . . .	163
5.12. Residuos MCO versus <i>RENTA</i> . . . . .	164
5.13. Residuos versus tiempo . . . . .	166
5.14. Residuos en <i>t</i> versus residuos en <i>t-1</i> . . . . .	166

5.15. Variable endógena versus exógena . . . . .	167
5.16. Residuos modelo (6.20) versus tiempo . . . . .	168

# Tablas

1.1. N° de aprobados en tres asignaturas de la Licenciatura en Economía. . . . .	20
1.2. Gasto sanitario por estado . . . . .	22
1.3. Distribución de frecuencias para el gasto en sanidad . . . . .	22
1.4. Demanda doméstica de electricidad . . . . .	22
1.5. Distribución de frecuencias para la demanda doméstica de electricidad . . . . .	23
3.1. Datos de características de viviendas. Fichero 4-1.gdt. . . . .	65
4.1. Observaciones sobre rendimiento y t/i por país . . . . .	116
4.2. Observaciones muestrales de la prima pagada y renta . . . . .	118



# Tema 1

## Introducción a la Econometría

En este tema y siguientes vamos a abordar cómo se relacionan las variables entre sí. De ello se ocupa la Econometría. Así, en estos temas aprenderemos a interpretar la información estadística sobre la realidad económica. La importancia de la Econometría va más allá de la disciplina de la economía. La Econometría es un conjunto de instrumentos de investigación empleados en finanzas, marketing, dirección de empresas, negocios, historia, sociología incluso agronomía.

La herramienta básica es un modelo econométrico que conjuga los esquemas teóricos sobre el funcionamiento de la Economía con las técnicas estadísticas de análisis de datos. Un modelo puede tener una estructura muy compleja, pero nos centramos en el modelo más sencillo, y que da contenido a buena parte de la asignatura, el **modelo de regresión lineal general**. Este modelo explica el comportamiento de una única variable económica mediante un conjunto de variables. En este tema definiremos la disciplina de la Econometría e introduciremos conceptos relacionados con un modelo econométrico: los datos, las variables, los parámetros, entre otros elementos de un modelo.

El desarrollo de la Econometría ha sido enormemente facilitado por el avance en la informática. El curso, con gran componente aplicado necesita complementarse con el aprendizaje de un software econométrico. El paquete econométrico a utilizar es *gretl*; se trata de software de libre uso, fácil de manejar y que tiene acceso a las bases de datos que se estudian en muchos libros de análisis econométrico. El alumno deberá aprender su manejo, en paralelo con los conceptos estadísticos y econométricos, y a interpretar adecuadamente los resultados obtenidos.

### Competencias a trabajar en estas sesiones:

1. Comprender la lógica de la modelización econométrica y las características de los diferentes elementos de los modelos, así como la relevancia de cada uno de los supuestos empleados en la especificación de un modelo.

### Al final de este tema deberíais ser capaces de:

1. Distinguir entre un modelo económico y un modelo econométrico.
2. Conocer las etapas en la realización de un trabajo aplicado.

3. Distinguir los diferentes tipos de datos empleados en el análisis econométrico.
4. Distinguir las diferentes variables implicadas en un modelo econométrico.
5. Distinguir entre parámetros de la relación económica y parámetros de la relación probabilística.
6. Distinguir entre estimador y estimación.

**Bibliografía Recomendada:**

Al final del tema tenéis recogida la bibliografía correspondiente. En particular se os recomienda leer los capítulos correspondientes a la bibliografía básica detallados a continuación:

- Gujarati, D. y Porter, D.C. (2010). Introducción.
- Ramanathan, R. (2002). Cap. 1.
- Wooldridge, J.M. (2006). Cap. 1.

## ¿Qué es la Econometría?

*Econometría en sentido estricto significa medida de la economía. La Econometría se ocupa de formular, cuantificar y valorar las relaciones entre variables económicas, para ello necesita de otras materias como son la Teoría Económica, la Estadística y las Matemáticas.*

*La Econometría se ocupa del estudio de estructuras que permitan analizar características o propiedades de una variable económica utilizando como causas explicativas otras variables económicas. (Novales, 1993)*

### 1.1. Modelo económico y modelo econométrico

Como es sabido la Teoría Económica se ocupa del análisis de la economía, como consecuencia del mismo formula las relaciones existentes entre las variables económicas objeto de estudio. Sin embargo la teoría Económica no se ocupa de cuantificarlas, éste es un cometido específico de la Econometría, que sí tiene como objetivo cuantificar las relaciones entre variables. Unido a este objetivo aparece un pilar clave para la Econometría que es la disponibilidad de información cuantificada sobre las variables que son objeto de estudio, en definitiva lo que llamamos **datos**. Las Matemáticas nos servirán para escribir en términos de ecuaciones las teorías económicas objeto de estudio y la Estadística nos proporciona instrumentos para el tratamiento de datos que nos permiten cuantificar las relaciones y valorar los resultados de acuerdo a criterios establecidos. En ocasiones nos encontraremos con problemas específicos para los que la estadística no tiene solución y por ello necesitaremos desarrollar los instrumentos y métodos apropiados para llevar a cabo los objetivos.

Resumiendo, podríamos decir que los **objetivos de la Econometría** son: verificación de una teoría, estudio del pasado, descripción del presente, predicción del futuro y orientación de la acción política. Para tratar de entender las relaciones entre la Econometría y las otras materias mencionadas en el apartado anterior vamos a desarrollar un ejemplo.

Supongamos que somos el gerente de una empresa y que estamos interesados en la relación existente entre las ventas de un producto de la empresa y su precio, las condiciones de la competencia y el ciclo económico. Un modelo que tiene en cuenta estos supuestos podría ser el siguiente:

$$V_t = f(p_t, pc_t, c_t) \quad (1.1)$$

Siendo  $V$  las ventas de la empresa y  $p$  el precio del producto, la variable  $pc$  es el precio de la competencia y nos sirve para aproximar las condiciones de la competencia. La variable  $c$  recoge el momento del ciclo económico y sirve para aproximar las condiciones de mercado. El subíndice  $t$  denota el tiempo o momento en el que se considera la relación. La ecuación anterior postula que las ventas son función del precio del producto, el precio de la competencia y del ciclo económico. Además la Teoría Económica nos dice que la relación entre ventas y precio es inversa, es decir, a mayor precio menores ventas. Sin embargo será positiva con respecto al precio de la competencia ya que si el precio de la competencia sube y el propio se mantiene es lógico que se espere vender más. De igual manera se venderá más en momentos de auge económico que en momentos de depresión por lo que la relación entre las ventas y el ciclo económico también se esperará que sea positiva.

El gerente también dispondrá de información en forma de cifras o datos sobre cuales eran las ventas correspondientes a los diferentes precios que ha podido alcanzar su producto, el precio de la competencia y el momento del ciclo económico, variable que puede aproximarse a una variable cuantitativa que se mueva con el ciclo económico, por ejemplo el Índice de Producción Industrial.

Por ahora como gerentes de la empresa disponemos de dos informaciones distintas. Por un lado disponemos de un modelo económico que nos relaciona un conjunto de variables y por otro disponemos de observaciones o datos sobre las mismas para un periodo de tiempo dado. El gerente también dispone de un objetivo que es saber como responden las ventas de su producto a cambios en su precio. Para unir ambos conjuntos de información podemos empezar por dar forma a la función. La elección más sencilla sería tomar una relación lineal, que para la ecuación (1.1) determinaría el siguiente modelo:

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 p c_t + \beta_4 c_t \quad (1.2)$$

Los parámetros o coeficientes de cada variable se representan por  $\beta_1$ ,  $\beta_2$  y  $\beta_3$ . El coeficiente  $\beta_2$  le indica al gerente cuanto cambian las ventas si el precio de su producto cambia en una unidad, permaneciendo el resto de variables constantes.

Con los datos disponibles, que supongamos son:

fecha	ventas	precio	p. competencia	IPI
$t$	$V$	$p$	$pc$	$c$
enero 80	1725	12,37	11,23	101,7
febrero 80	1314	11,25	10,75	97,3

podemos relacionar las variables con los valores que han tomado en cada momento siguiendo la ecuación (1.2). Así en enero de 1980 la relación entre las ventas y el resto de variables ha sido:

$$1725 = \beta_1 + 12,37\beta_2 + 11,23\beta_3 + 101,7\beta_4$$

Mientras que en febrero de 1980 fue:

$$1314 = \beta_1 + 11,25\beta_2 + 10,75\beta_3 + 97,3\beta_4$$

Estas relaciones se repetirían para cada mes del que tengamos datos. Como el valor de las variables cambia de un mes a otro, para que las igualdades se cumplan también deben cambiar los valores de los parámetros. Este no es el objetivo del gerente, quién necesita la mejor aproximación posible del valor de las ventas al precio, que resuma toda la información disponible del periodo considerado. Para ello consideraremos que el modelo debe reflejar el comportamiento medio de la relación entre variables manteniéndose la relación entre las variables estable. Para que esto se cumpla y podamos recoger el comportamiento medio incluiremos en el modelo un nuevo elemento al que llamaremos  $u_t$ . Así el modelo especificado será:

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 c p_t + \beta_4 c_t + u_t \quad (1.3)$$

El nuevo elemento deberá ser capaz de mantener la igualdad de la relación para cualquier conjunto de datos, tomando por tanto a veces valores positivos y en otras ocasiones valores negativos; a veces grandes, a veces pequeños. La interpretación del mismo resulta bastante intuitiva: recoge

todos los efectos que afectan a las ventas en cada período muestral y que no están explícitamente recogidos en las variables que el modelo contiene. Si el modelo ha recogido todas las influencias “importantes y sistemáticas” que existen sobre las ventas, el nuevo elemento, que en adelante llamaremos **perturbación** recogerá los efectos no sistemáticos que serán, en general, más erráticos. Por tanto es factible considerar su comportamiento como aleatorio. Así a la perturbación  $u_t$  se le trata como una variable aleatoria cuya distribución de probabilidad es preciso especificar al mismo tiempo que el resto del modelo.

Dado que el modelo recogido por la ecuación (1.3) contiene una variable aleatoria para obtener resultados a partir del mismo necesitaremos de la Estadística. Mediante procedimientos estadísticos podremos cuantificar la relación entre las variables, obteniendo valores numéricos para los coeficientes  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  y  $\beta_4$  que reflejen la información que contienen los datos. De esta forma el modelo general representado por la ecuación (1.3) que en principio puede servir para analizar el comportamiento de cualquier empresa servirá para contestar a las preguntas que el gerente se hace sobre su propia empresa convirtiéndose en un modelo específico válido para la toma de decisiones.

El ejemplo anterior describe una situación muy concreta pero la Econometría es útil en otras muchas situaciones, por ejemplo:

- Para analizar el efecto del impacto de cambios en la política fiscal sobre los indicadores económicos de un país, la demanda interna, los tipos de interés, exportaciones e importaciones, desempleo, grado de morosidad.
- Los directivos de la empresa Mercedes pueden estar interesados en los factores que determinan la demanda de automóviles.
- Para analizar los efectos de la publicidad en las ventas de una empresa.
- Para analizar el impacto en la función de producción de cambios en los factores de producción.
- Analizar si la demanda de tabaco se ve afectada por las campañas anti tabaco.
- Analizar si las campañas publicitarias contra el consumo de alcohol cuando se conduce reduce el número de siniestros.
- Estudiar como afecta el tabaquismo al peso de nacimiento y posterior crecimiento de un bebe.

## 1.2. Etapas en la elaboración de un modelo econométrico

Un estudio econométrico consta de las siguientes etapas, Heij , de Boer, Franses, Kloer y Dijk (2004):

- *Formulación del problema.* Se trata de determinar la cuestión de interés. Debemos plantear de forma precisa las preguntas que nos interesa responder. La teoría económica puede ayudarnos a enfocar el problema, a determinar qué variables están involucradas y cuál puede ser la relación entre ellas.

- *Recolección de datos* estadísticos relevantes para el análisis. En el caso del gerente los datos están disponibles en los balances de la propia empresa. Los resultados del análisis van a depender en gran medida de la calidad de los datos. Sin embargo, no siempre es sencillo obtener los datos relevantes para el análisis. Podemos encontrar problemas como la ausencia de algún dato, cambios en la definición de una variable, fallos en el método de recogida, tener una cantidad insuficiente de datos o no disponer de información relativa a una variable.
- *Formulación y estimación del modelo*. En esta fase hay que dar forma al problema inicial en términos de un modelo. Determinar la variable a *explicar*, en el ejemplo las ventas, y las variables explicativas, en el ejemplo el precio, el precio de la competencia y el ciclo económico; la forma funcional del modelo y la distribución probabilística de la perturbación aleatoria.

El siguiente paso es la estimación de los parámetros desconocidos de la distribución y que son de interés para el análisis. La estimación consiste en utilizar los datos y toda la información relevante para aprender algo sobre los parámetros desconocidos. En la interpretación de los resultados de estimación es importante tener en cuenta que *no conocemos* el valor de los parámetros, por lo que únicamente vamos a hacer afirmaciones del tipo “*con un 95% de confianza, el aumento del impuesto sobre carburantes no afecta al consumo de gasolina*”.

Existen muchos métodos de estimación. La elección entre uno u otro depende de las propiedades del modelo econométrico seleccionado. Es decir, una mala selección del modelo también influye en la validez de las estimaciones. Un curso introductorio de Econometría, como este, se suele centrar en el estudio del modelo de regresión lineal y su estimación mediante *mínimos cuadrados ordinarios*, que son instrumentos sencillos y muy útiles en la práctica.

- *Análisis del modelo*. Se trata de estudiar si el modelo elegido es adecuado para recoger el comportamiento de los datos. Consiste en una serie de contrastes diagnósticos que valoran si el modelo está correctamente especificado, es decir, si los supuestos realizados son válidos. Si es necesario, se modifica el modelo en base a los resultados obtenidos en los contrastes.
- *Aplicación del modelo*. Una vez obtenido un modelo *correcto*, se utiliza para responder a las cuestiones de interés y para la *predicción*. Un modelo correctamente especificado y estimado ha de ser utilizado para predecir. Este concepto implica tanto determinar los valores futuros de la variable endógena como contestar a preguntas del tipo ¿qué pasaría si...?, en definitiva debe servirnos para dar consejos de política económica.

### 1.3. Tipología de datos y variables en Econometría

El modelo econométrico genérico completamente especificado tiene la siguiente forma:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_K X_{Kt} + u_t \quad t = 1, 2, \dots, T \quad (1.4)$$

Donde  $Y$  es la variable a explicar o variable endógena,  $X_2, X_3, \dots, X_K$  son las variables explicativas, o regresores, del modelo. El subíndice que las acompaña indica el número de variables explicativas del modelo, el modelo anterior tiene  $K$ -variables explicativas. Los coeficientes  $\beta_k \quad k = 1, 2, \dots, K$  son los parámetros a estimar, que se suponen constantes. Además es de interés notar que el parámetro  $\beta_1$  acompaña a la variable explicativa  $X_1$  constante e igual a la unidad en todo momento del

tiempo. El subíndice  $t$  hace referencia al tiempo y por tanto  $T$  indica el tamaño de la muestra de observaciones disponible.

La diferencia entre un modelo económico y un modelo econométrico es la perturbación aleatoria que incluimos en el modelo econométrico. A partir de este elemento en el modelo econométrico podemos distinguir dos partes **la parte sistemática** del modelo y **la parte aleatoria**. La primera corresponde al comportamiento medio o estable de la relación y la segunda se corresponde con la perturbación aleatoria,  $u_t$ .

El objetivo sobre el modelo genérico representado por la ecuación (1.4) es conocer los valores de los parámetros desconocidos  $\beta_k$   $k = 1, 2, \dots, K$ . Para llevar a cabo este objetivo utilizaremos métodos estadísticos. Para ello al modelo especificado deberemos de añadir hipótesis sobre el comportamiento probabilístico de la perturbación aleatoria que caractericen su distribución. En general, supondremos que dicha perturbación tiene una distribución centrada en cero, o sea media cero, lo que implica que el comportamiento medio de la variable a explicar está recogido en su totalidad por la parte sistemática del modelo:

$$E(Y_t) = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_K X_{Kt} \quad t = 1, 2, \dots, T \quad (1.5)$$

Además de la media debemos caracterizar también la varianza, covarianzas y distribución de la perturbación.

### 1.3.1. Conceptos básicos

En los puntos anteriores han surgido algunos conceptos que deberían quedar claros para poder referirnos a ellos con propiedad. Revisaremos algunos de ellos.

- **Población y muestra:**

Población son todos los posibles valores que toma la variable objeto de estudio. La muestra sería la parte de la población que vamos a utilizar en el estudio para extraer conclusiones. Por tanto la muestra está contenida en la población y nosotros la utilizaremos para establecer conclusiones que puedan extrapolarse a la población.

- **Datos:**

Los datos son los valores numéricos que toman tanto la variable a explicar como las variables explicativas. Generalmente los obtenemos de series estadísticas cuyas fuentes pueden ser oficiales o privadas. La importancia de los datos está determinada por la unidad de medida. Los podemos clasificar en:

1. *Datos de serie temporal:* Reflejan la evolución de una variable a lo largo del tiempo, según esto la variable estará ordenada cronológicamente con un orden lógico. Las variables medidas en series temporales se denotan con el subíndice  $t$  y este puede referirse a observaciones temporales mensuales, trimestrales, diarias cuatrimestrales, anuales, etc. Ejemplo: el Producto Nacional Bruto (PNB) de 1965-2000. En este caso la población serían todos los posibles valores del PNB a lo largo del tiempo y la muestra el período que vamos a estudiar, de 1965 al 2000.

2. *Datos de sección cruzada o corte transversal*: Son datos atemporales dado que miden el comportamiento de una variable en diferentes unidades y en el mismo momento del tiempo. Ejemplo: ventas de las empresas metalúrgicas en el País Vasco en el año 1999. Esta sería la muestra a utilizar y la población estaría constituida por todas las unidades.
3. *Datos de panel*: es la unión de datos de serie temporal y datos de sección cruzada. Están fuera del objetivo del curso.

- **Variables:**

Una variable es un ente económico que toma diferentes valores. Podemos distinguir entre variables exógenas, aquellas que inciden en el modelo desde el exterior y variables endógenas, aquellas que queremos explicar con el modelo. A las variables exógenas también se las denomina variables explicativas o independientes y a la variable endógena también se le puede denominar como variable a explicar o dependiente. Además debemos tener en cuenta que podemos encontrarnos con relaciones simultáneas como:

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + u_t$$

o como

$$C_t = \beta_1 + \beta_2 Y_t + u_t \quad Y_t = C_t + I_t$$

donde las variables cambian su papel según miremos a una ecuación u otra. Podemos distinguir, entre otros, los siguientes tipos de variables:

1. - *Fijas*: aquellas que toman valores que el investigador puede controlar.
  - *Estocásticas*: aquellas cuyo valor cambia según una ley de probabilidad.
2. - *Cuantitativas*: aquellas que podemos valorar numéricamente. Por ejemplo, la renta disponible de una familia, el precio de un bien, la renta per cápita.
  - *Cualitativas*: aquellas que miden cualidades y que por lo tanto no se miden con un valor numérico y será el investigador el que se lo asigne según un criterio. Por ejemplo, si un individuo está o no casado, si trabaja en turno de noche o no, si tiene estudios superiores o no. En las variables cualitativas es el investigador el que establece el valor de la variable para cada característica. Por ejemplo:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases}$$

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

definen dos variables cualitativas  $S_{1i}$  y  $S_{2i}$  que permiten recoger el sexo del individuo y ver por ejemplo si existe discriminación salarial por sexo en un estudio sobre la función de salario.

- **Los parámetros:**

Los parámetros son los valores que permanecen desconocidos del modelo. En un modelo económico podemos distinguir dos tipos de parámetros:

1. *Los parámetros de la relación económica:* Son las ponderaciones que aplicadas a las variables exógenas nos permiten calcular la endógena.

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 c p_t + \beta_4 c_t + u_t \quad (1.6)$$

En el modelo anterior  $\beta_1, \beta_2, \beta_3$  y  $\beta_4$ .

2. *Los parámetros de la estructura probabilística:* son los parámetros que determinan la estructura de la parte aleatoria del modelo, media y varianza de la perturbación aleatoria y de la variable endógena.

• **Modelo:**

Hemos visto que un modelo no es más que un conjunto de relaciones entre variables económicas y que representamos mediante relaciones matemáticas. Clasificación de los modelos:

1. - *Modelos exactos:* aquellos que determinan exactamente el valor de una variable conocido el valor de otra-s:

$$Y = \beta_1 + \beta_2 X$$

- *Modelos estocásticos:* aquellos que incluyen alguna variable aleatoria:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad u \sim (m(u), V(u))$$

2. - *Modelos uniecuacionales:* aquellos que se componen de una única ecuación:

$$C_t = \beta_1 + \beta_2 Y_t + u_t$$

- *Modelos multiecuacionales:* aquellos que se componen de más de una ecuación. Por ejemplo cuando una variable influye en otra-s y a la vez es influida por éstas:

$$C_t = \beta_1 + \beta_2 Y_t + u_t \quad Y_t = C_t + I_t$$

3. - *Modelos estáticos:* Cuando el tiempo no aparece de forma explícita en la ecuación y todas las variables se miden en el mismo momento.

- *Modelos dinámicos:* Aquellos que tienen variables definidas en diferentes momentos del tiempo o el tiempo aparece como variable explícita en la ecuación. Un ejemplo de los primeros sería:

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 C_{t-1} + u_t$$

mientras que un ejemplo de los segundos sería el siguiente modelo no explícitamente dinámico, generalmente llamado estático histórico

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 t + u_t$$

donde el parámetro  $c$  recoge la tendencia de la variable endógena a lo largo del tiempo.

4. - *Modelos basados en series temporales:* pueden ser dinámicos u estáticos.
  - *Modelos basados en datos de corte transversal:* son siempre estáticos.

- **Parámetro, estimador y estimación:**

En el modelo:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1, 2, \dots, T$$

tenemos diferentes **parámetros desconocidos**. En la parte aleatoria aparecerían los que caracterizan a la distribución probabilística de la perturbación aleatoria y en la parte sistemática aparecen  $\beta_1$  y  $\beta_2$ . Todos son parámetros desconocidos. Los llamaremos parámetros poblacionales ya que lo que nosotros hemos especificado es un modelo general que debería recoger el comportamiento medio de las variables en la población. Para obtener resultados del modelo anterior nosotros lo aplicamos a la muestra, de tamaño  $T$ . Nuestro objetivo es determinar el valor de estos parámetros poblacionales desconocidos de la muestra. Para aproximarnos a ese valor utilizamos técnicas estadísticas, en concreto estimadores. Un **estimador** no es más que una fórmula que nos dice como debemos obtener los valores numéricos de  $\beta_1$  y  $\beta_2$  mediante la muestra. Al valor finalmente obtenido en la muestra le llamamos **estimación**. En concreto la notación matemática para estos conceptos, aplicada al parámetro  $\beta_2$  sería:

$$\begin{array}{ll} \beta_2 & \text{parámetro poblacional} \\ \hat{\beta}_2 & \text{estimador} \\ 0,5 & \text{estimación} \end{array}$$

donde por ejemplo:

$$\hat{\beta}_2 = \frac{\sum_{t=1}^T (Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^T (X_t - \bar{X})^2} = 0,5$$

Los estimadores van a ser variables aleatorias con distribución a determinar ya los que exigiremos ciertas propiedades que van a determinar esta distribución.

- **Estructura:**

Cuando estudiamos la relación entre las variables económicas especificamos un modelo econométrico. En la especificación elegimos la forma funcional del modelo y las variables explicativas a incluir así como las propiedades de la perturbación. Una vez que el modelo está totalmente especificado le estimaremos y tendremos unos valores para los parámetros. A la relación resultante le llamamos estructura. **Un modelo especificado** sería:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1, 2, \dots, T$$

mientras que **una estructura** para ese modelo dada una muestra de tamaño  $T$  podría ser:

$$\hat{Y}_t = 20 + 5X_t$$

Notar que un modelo puede tener diferentes estructuras según los valores que las variables exógena y endógena tomen en la muestra.

### 1.3.2. Fuentes de datos

Encontrar y recopilar datos no es siempre sencillo. En ocasiones es muy costoso coleccionar los datos adecuados a la situación y manejarlos. Sin embargo, esta tarea se ha visto favorecida en los últimos

años por la mejora en la recogida de datos y el hecho de que muchos organismos permiten acceder a sus bases de datos en la *World Wide Web*. Algunos organismos que publican datos macroeconómicos son:

- Instituto Vasco de Estadística (EUSTAT): <http://www.eustat.es>.
- Banco de España: <http://www.bde.es> → Estadísticas. También publica el **Boletín estadístico mensual** y el Boletín de coyuntura mensual.
- Instituto Nacional de Estadística (INE): <http://www.ine.es> → Inebase o Banco tempus. Están disponibles, por ejemplo, los resultados de la encuesta de población activa, la Contabilidad Nacional o el **boletín estadístico mensual**. Además, en *enlaces* se encuentran otras páginas web de servicios estadísticos.
- EUROSTAT: Es la Oficina Estadística de la Unión Europea, se encarga de verificar y analizar los datos nacionales recogidos por los Estados Miembros. El papel de Eurostat es consolidar los datos y asegurarse de que son comparables utilizando una metodología homogénea. La información en términos de tablas estadísticas, boletines estadísticos e informativos, incluso documentos de trabajo papers se puede encontrar en la dirección: <http://europa.eu.int/comm/eurostat>.
- Organización para la Cooperación y Desarrollo Económico (OCDE): <http://www.oecd.org>, Statistical portal, statistics. Están disponibles algunas series de las publicaciones **Main Economic Indicators** (mensual) o Comercio internacional.
- Fondo Monetario Internacional (FMI): <http://www.imf.org>. Para obtener datos sobre un amplio conjunto de países también se puede consultar su publicación **Estadísticas Financieras Internacionales** (mensual y anual).

Muchos manuales de Econometría incluyen una base de datos que se analizan en el texto como ilustración a la materia. En este curso utilizaremos principalmente los datos incluidos en Ramanathan (2002) y Wooldridge (2006) que están accesibles como archivos de muestra en *gretl*.

## 1.4. Tratamiento de la información con *gretl* : inclusión de datos en *gretl* y análisis descriptivo básico

*gretl* es un programa que permite obtener de manera sencilla mediante ventana resultados estadísticos y econométricos. Una vez ejecutado el programa *gretl* en la ventana principal aparece un menú de ventanas que nos permite diferentes posibilidades. En la pantalla principal, una vez abierto *gretl* nos aparecen las siguientes pestañas:

*Archivo Herramientas Datos Ver Añadir Muestra Variable Modelo Ayuda*

Pero solo tres de ellas están activas, las distinguimos porque las no activas aparecen en gris mientras que las activas están en negrita. Las activas son *Archivo*, *Herramientas* y *Ayuda*. En la primera leemos datos. Empezaremos viendo como leer datos. Dependiendo del origen de éstos si están en una archivo de muestra incluido en *gretl*, si están disponibles en papel, en la web o en un archivo

propio procederemos de una manera u otra.

- Para leer **datos incluidos en la base del programa gretl** :

Pinchar *Archivo* → *Abrir archivo de datos* → *Archivo de muestra* → Aquí seleccionamos la base de datos que necesitemos, por ejemplo *ETM* → y ahora seleccionamos el archivo, por ejemplo *monthly-crsp.gdt*

Aparecerán las variables de la muestra y en la barra superior se habrán activado las etiquetas mencionadas anteriormente. Por ejemplo en *Datos* podremos ver las observaciones y sus características. Algunas de las opciones que contiene la etiqueta *Datos* son las siguientes:

Mostrar valores  
Editar los valores  
Información del conjunto de datos  
Estructura del conjunto de datos

Para obtener lo que necesitamos sólo tenemos que pinchar la etiqueta correspondiente y la variable o variables a estudiar. Por ejemplo para ver la estructura del conjunto de datos pinchamos en esta etiqueta y obtendremos una pantalla en la que aparecerá seleccionado el tipo de datos con el que estamos trabajando, en este caso *Serie temporal*. Pinchamos adelante y veremos la frecuencia, *mensual*, y el inicio y final de la muestra *1968:1 a 1998:12*. La etiqueta estructura del conjunto de datos es muy útil cuando necesitamos cambiar alguno de ellos por ejemplo si añadimos nuevas observaciones.

En el menú inicial aparece también la etiqueta *Ver* con, entre otras, las siguientes opciones:

Gráficos  
Gráficos múltiples  
Estadísticos principales  
Matriz de correlación

- Para hacer **Gráficos**:

Pinchar *Ver* → *Gráficos* → *Gráficos de series temporales*. Seleccionar las variables que se quieren incluir en el gráfico y pinchar *Aceptar*.

Para guardar el gráfico: situar el ratón sobre el gráfico y pinchar con el botón derecho. Elegir opción. Podemos guardarlos en postscript (.eps) o .png, etc. En la ventana que aparece para guardarlo escribir la dirección de la carpeta donde queremos guardar el gráfico y ponerle un nombre por ejemplo CRSPVW.

Dentro de las opciones que aparecen al pinchar con el botón derecho está la opción *Editar*. En esta opción se pueden modificar los ejes, los nombres de las variables, incluso el tipo de línea y color utilizada para representar la serie de observaciones, entre otras posibilidades.

- Para obtener los **Estadísticos principales** de las variables de la muestra:

Pinchar en *Ver* → *Estadísticos principales*.

La ventana de output mostrará la media, moda, valor máximo y mínimo de la serie, desviación típica, coeficiente de variación, curtosis y asimetría. Podemos obtener los estadísticos para una única serie o para el conjunto de ellas seleccionándolo previamente.

Si queremos guardar el output pinchamos en el icono del diskette arriba a la izquierda y seleccionamos cómo queremos que lo guarde, texto plano, Word o Latex y en la ventana damos el nombre que deseemos al fichero de resultados, por ejemplo *estadVW* para la serie CRSP o *estadmuestra* para el conjunto y a continuación damos la dirección de la carpeta donde queremos que nos guarde el fichero de resultados.

En el menú inicial también aparece la etiqueta *Variable* para trabajar con una única serie de la muestra. Algunas de las opciones que incluye esta etiqueta son:

- Buscar
- Mostrar valores
- Estadísticos principales
- Contraste de Normalidad
- Distribución de frecuencias
- Gráfico de frecuencias (simple, contra la normal, contra la gamma)
- Gráfico de series temporales
- Editar atributos
- etc.

• **Obtener datos que están en el servidor:**

Queremos estudiar una serie que se encuentra en el servidor, *Crédito más de 5 años a hogares*. Esta serie aparece publicada en la base de datos del Banco de España con el código BE182704.

Pinchar *Archivo* → *Bases de datos* → *Sobre servidor*

En el listado de bases de datos que aparece vamos a *bde18 Banco de España (Tipo de interés)* y pinchamos en *Obtener listado de series* comprobando que contienen la serie que queremos.

*Series* → *Mostrar*

Para representarla gráficamente: *Series* → *Representar*

Para importar los datos a *gretl* situamos el cursor sobre la serie de interés, BE182704, y vamos a

*Series* → *Importar*

Además tenemos opción de hacer lo siguiente:

- Añadir o cambiar información sobre la variable: en menú *Variable* → *Editar atributos*. En esta ventana podremos cambiar también el nombre de la serie utilizado en los gráficos.
- Añadir notas explicativas: en menú *Datos* → *Editar información*
- Consultar las notas informativas: en menú *Datos* → *Leer información* o en *Datos* → *Descripción*

- Para **crear un conjunto de datos**:

Pinchar *Archivo* → *Nuevo conjunto de datos* y completar la información que pide sobre:

número de observaciones  
 estructura del conjunto de datos (serie temporal o sección cruzada)  
 frecuencia

A la pregunta *¿Desea empezar a introducir los valores de los datos usando la hoja de cálculo de gretl ?* contestar *Sí*

- Introducir el nombre de la variable. El máximo de caracteres que acepta es 15, no usar acentos ni la letra ñ. Pinchar *Aceptar*.
- En la hoja de cálculo situarnos en la primera celda y teclear la observación correspondiente, a continuación pinchar *intro*. Si nos saltamos alguna observación podemos insertar una fila en el lugar correspondiente con solo situarnos en la celda posterior e ir a *observación* → *insertar obs*. Una vez introducidas todas las variables pinchar *Aplicar*.
- Para guardar los datos: en menú *Archivo* → *Guardar datos*. Dar nombre al conjunto de datos, por ejemplo *Azar* y se grabará automáticamente con la extensión *gdt*.

Si en otro momento queremos usar este conjunto de datos solo habrá que clicar en él dos veces para que se active.

- Si queremos añadir variables en menú: Pinchar en la etiqueta *Añadir* tenemos las siguientes posibilidades:
  - Logaritmos de las variables seleccionadas
  - Cuadrados de las variables seleccionadas
  - Retardos de las variables seleccionadas
  - Primeras diferencias de las variables seleccionadas
  - Diferencias del logaritmo las variables seleccionadas
  - Diferencias estacionales de las variables seleccionadas
  - Variable índice
  - Tendencia temporal
  - Variable aleatoria (uniforme, normal, chi cuadrado y t-Student) Por ejemplo para crear una variable normal de media 0 y desviación 1 haremos *nombre de la variable 0 1*
  - Variables ficticias, etc.
  - Definir una nueva variable. Esta opción podemos utilizarla para crear combinaciones de variables por ejemplo  $Z_t = 4 + \epsilon_t$      $\epsilon_t \sim N(0, 1)$ . Permite los operadores,

$+$ ,  $-$ ,  $*$ ,  $/$ ,  $\wedge$

(suma, resta, producto, potencia) entre otros.

- Para obtener **información sobre la muestra** pinchar en la etiqueta *Muestra*. En ella encontraremos, entre otras, las siguientes opciones:

Establecer rango  
 Recuperar rango completo  
 Restringir, a partir de un criterio  
 etc.

### Ejemplo 1.1

Vamos a trabajar con el archivo de datos *data4 - 1.gdt* ya que en los temas siguientes va a ser uno de los ejemplos que seguiremos. Está incluido como archivo de muestra en la pestaña *Ramanathan*. Una vez abierto podemos buscar información sobre sus variables tal y como se ha indicado. Siguiendo la ruta indicada encontramos la siguiente *Información del conjunto de datos*

```
DATA4-1: Data on single family homes in University City community
of San Diego, in 1990.
price = sale price in thousands of dollars (Range 199.9 - 505)
sqft = square feet of living area (Range 1065 - 3000)
bedrms = number of bedrooms (Range 3 - 4)
baths = number of bathrooms (Range 1.75 - 3)
```

Donde aparece una somera descripción de los datos disponibles y su fuente y/o origen. En este caso nos dicen que son datos de hogares de la comunidad universitaria de San Diego en 1990, de lo que deducimos que son datos de sección cruzada ya que se refieren a un único año. También aparecen los nombres de las variables y su descripción así como el rango de cada una (la amplitud del intervalo de valores que toma la variable en la muestra) y la fuente de los datos. Los estadísticos principales son los siguientes:

Estadísticos principales, usando las observaciones 1 - 14

Variable	Media	Mediana	Mínimo	Máximo
price	317,493	291,500	199,900	505,000
sqft	1910,93	1835,00	1065,00	3000,00
bedrms	3,64286	4,00000	3,00000	4,00000
baths	2,35714	2,25000	1,75000	3,00000
Variable	Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
price	88,4982	0,278741	0,653457	-0,529833
sqft	577,757	0,302344	0,485258	-0,672125
bedrms	0,497245	0,136499	-0,596285	-1,64444
baths	0,446291	0,189336	0,331609	-1,39015

Donde se nos muestra, para cada variable, su media, mediana, valores mínimo y máximo, desviación típica, coeficiente de variación (C.V.), coeficiente de asimetría y coeficiente de exceso de curtosis.

Los gráficos de las variables *price* y *sqft* son:

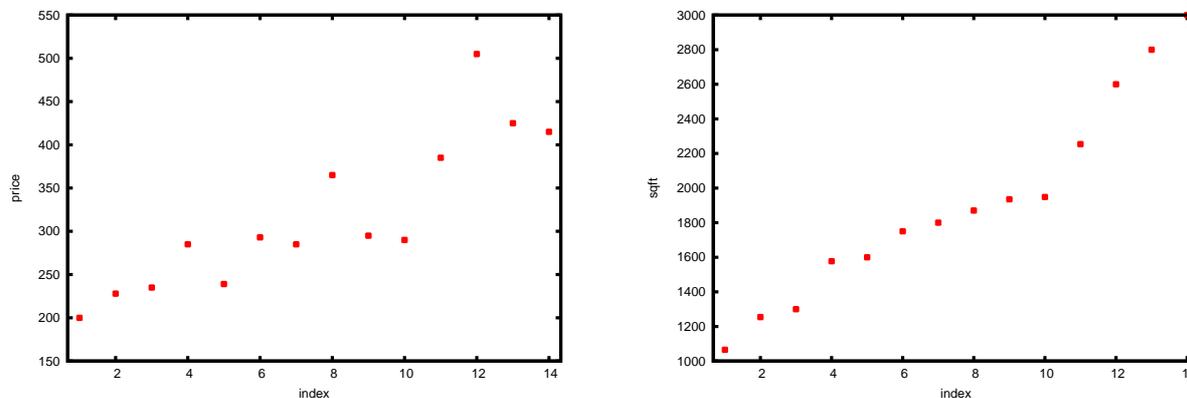


Figura 1.1: Gráficos de las observaciones para las variables *price* y *sqft*

Volviendo a la pantalla de inicio. También estaban disponibles al iniciar el programa las etiquetas *Herramientas* y *Ayuda*. En *Herramientas* disponemos de instrumentos de análisis muy útiles como:

- En *Tablas estadísticas* los valores críticos de las distribuciones Normal Tipificada, t-Student y F-Snedecor entre otras distribuciones.
- Un buscador de valores p.
- Un calculadora de estadísticos de contraste como la igualdad de medias o varianzas.
- La posibilidad de dibujar distribuciones o curvas.
- Hacer contrastes no paramétricos.
- Generar numeros aleatorios.

En *Ayuda* encontramos al *Guía del usuario* y la *Guía de instrucciones*.

## 1.5. Apéndice

La Estadística, la Econometría y muchas otras materias se alimentan de datos. En ocasiones el número de datos u observaciones recogidas es tan grande que no es fácil obtener unos resultados interpretables con claridad. En este tema vamos a dedicar las clases a resumir de forma clara y precisa la información que nos transmiten los datos sin ocultar características importantes. No hay una estrategia correcta para ello en general depende del tipo de datos y del fin del estudio. En este tema vamos a introducir gráficos y tablas que resuman la información de los datos, por ejemplo gráficos de barra, tarta, gráficos de series temporales. Los gráficos, diagramas y tablas pueden mejorar la comunicación de los datos a los clientes u proveedores.

### 1.5.1. La naturaleza de los datos

En la actualidad el individuo tiene a su disposición datos de casi todo aquello que desee conocer. La Administraciones públicas, los entes privados y los investigadores científicos publican y recogen datos. Es muy importante saber resumir e interpretar la información que proporcionan. Pensemos en el lanzamiento de un nuevo producto al mercado. El productor necesitará conocer su nivel de demanda potencial, *todos* sus posibles compradores. A este conjunto de *todos* sus posibles compradores se le denomina *población*. En ocasiones el tamaño de la población puede ser muy grande e incluso infinito por lo que debemos limitarnos a estudiar una parte más reducida de la misma, que denominamos *muestra*.

La población es el conjunto de todos los objetos que interesan al investigador. Al número de observaciones u objetos que conforman la población se le denomina tamaño de la población y se denota por  $N$ . La muestra es un subconjunto observado de valores poblacionales. El número de observaciones que conforman la muestra se denomina tamaño de la muestra y se denota por  $n$ .

El muestreo aleatorio simple o muestreo aleatorio es el procedimiento que se utiliza para seleccionar una muestra de  $n$  observaciones de una población en el que cada miembro de la población se elige estrictamente al azar. Cada miembro de la población se elige con la misma probabilidad y todas las muestras posibles de un tamaño dado,  $n$ , tienen la misma probabilidad de ser seleccionadas.

Tomar una muestra y estudiar sus características es un medio para extraer conclusiones sobre la población. Por ejemplo si estamos interesados en conocer la renta media de las familias de un país, el tamaño de la población es tan grande que solo podríamos tomar una muestra aleatoria de por ejemplo 1000 familias y calcular su renta media. Esta media basada en datos muestrales se llama *estadístico*. Si pudiéramos calcular la renta media de todas las familias la media resultante se llamaría *parámetro*. En este curso estudiaremos cómo se toman decisiones sobre un parámetro basándonos en un estadístico. Trabajaremos en un entorno de incertidumbre ya que el valor exacto del parámetro es desconocido. Un parámetro es una característica específica de la población. Un estadístico es una característica específica de la muestra.

Una vez que tenemos definido un problema y recolectados los datos objeto de interés, estos se analizan utilizando métodos estadísticos. La estadística descriptiva y la estadística inferencial permiten convertir los datos en conocimiento útil para la toma de decisiones. La estadística descriptiva está formada por métodos gráficos y numéricos que permiten resumir y procesar los datos

convirtiéndolos en información. La estadística inferencial permite hacer estimaciones, contraste de hipótesis, predicciones y previsiones.

Una vez que tenemos recolectados los datos debemos transformarlos en información. Para ello podemos llevar a cabo un análisis gráfico de los mismos y también un análisis descriptivo.

### 1.5.2. Clasificación de las variables

Antes de abordar el estudio gráfico y el estudio descriptivo de los datos vamos a hablar de las variables. Una variable es una característica que al ser medida en diferentes individuos es susceptible de adoptar diferentes valores. Las variables pueden clasificarse de varias formas, por ahora lo haremos en dos. Cuando las clasificamos según el tipo y la cantidad de información que contienen podemos distinguir entre variables categóricas y numéricas. Cuando las clasificamos por nivel de medición podemos distinguir entre variables cualitativas y cuantitativas.

Una primera clasificación divide los datos en variables categóricas y variables numéricas. Las **variables categóricas** son aquellas que generan respuestas que pertenecen a categorías o grupos. Por ejemplo ¿tiene coche? ¿utiliza tarjeta de crédito? También son variables categóricas las respuestas a preguntas sobre sexo, estado civil, nivel de educación y también lo son las respuestas con diferentes grados de intensidad como las que van de totalmente de acuerdo a totalmente en desacuerdo.

Una variable numérica es aquella que toma valores. Una **variable numérica discreta** puede tener un número infinito de valores, pero en general proviene de un proceso de recuento, por ejemplo número de aprobados en Matemáticas II o el número de matriculados en una asignatura. Una **variable numérica continua** puede tomar cualquier valor en un intervalo dado de números reales y normalmente proviene de un proceso de medición. Por ejemplo la altura y el peso, la temperatura, el tiempo, la distancia entre dos puntos, etc.

También podemos clasificar los datos por su nivel de medición en variables cualitativas y cuantitativas. **Variables cuantitativas** son aquellas que podemos valorar numéricamente. Por ejemplo, la renta disponible de una familia, el precio de un bien, la renta per cápita. **Variables cualitativas** son aquellas que miden cualidades y que por lo tanto no se miden con un valor numérico y será el investigador el que se lo asigne según un criterio. Por ejemplo, si un individuo es hombre o mujer, si está o no casado, si trabaja en turno de noche o no, si tiene estudios superiores o no. En las variables cualitativas es el investigador el que establece el valor de la variable para cada característica.

Las variables cualitativas pueden tener niveles de medición nominales y ordinales mientras que los datos cuantitativos pueden tener niveles de medición basados en intervalos y en razones. Son datos nominales los que responden a preguntas categóricas, el sexo, el lugar de nacimiento, el estado civil. Por ejemplo, la respuesta a la pregunta sexo del individuo son palabras que describen la categoría: hombre, mujer y se ha de asignarlas un código de manera arbitraria, por ejemplo: hombre = 1, mujer = 0, este valor solo se emplea para clasificar a los individuos, la diferencia entre ellos no tienen ningún valor. Los datos ordinales indican el orden que ocupa el objeto, por ejemplo en una encuesta pueden preguntar sobre el grado de dificultad de un determinado procedimiento y las respuestas pueden ser: alto=1, medio=2, bajo=3. Las respuestas siguen un orden, son ordinales, pero la diferencia entre ellas no tiene significado mensurable.

Los niveles de medición basados en intervalos y razones se refieren a datos en una escala ordenada en los que la diferencia entre las mediciones tiene un significado. Una escala de intervalos indica el orden y la distancia con respecto a un cero arbitrario medidos en intervalos unitarios, por ejemplo el año es un dato de este tipo donde el nivel de referencia es el calendario gregoriano. Otro ejemplo sería la temperatura. Los datos basados en una escala de razones indican el orden y la diferencia con respecto a un cero natural y los cocientes entre dos medidas tienen un significado. Por ejemplo una persona que tiene 90 años tiene el doble de años que una persona de 45 años. El peso, la altura son variables de este tipo.

Con respecto a la unidad de medida los podemos clasificar en:

1. *Datos de serie temporal*: Reflejan la evolución de una variable a lo largo del tiempo, según esto la variable estará ordenada cronológicamente con un orden lógico. Las variables medidas en series temporales se denotan con el subíndice  $t$  y este puede referirse a observaciones temporales mensuales, trimestrales, diarias, cuatrimestrales, anuales, etc. Ejemplo: la demanda doméstica de electricidad (DDE) del segundo trimestre de 1972 al cuarto trimestre de 1993. En este caso la población serían todos los posibles valores de dicha demanda DDE a lo largo del tiempo y la muestra el período que vamos a estudiar, 1972:2-1993:4, el tamaño de la muestra es de 87 observaciones.
2. *Datos de sección cruzada o corte transversal*: Son datos atemporales dado que miden el comportamiento de una variable en diferentes unidades y en el mismo momento del tiempo. Ejemplo: el salario de los trabajadores del sector de educación en las diferentes comunidades autónomas españolas en el año 2011.
3. *Datos de panel*: es la unión de datos de serie temporal y datos de sección cruzada. Están fuera del objetivo del curso.

Hay que notar que una variable puede ser clasificada en varias de las categorías anteriores, una variable cualitativa puede ser categórica nominal u ordinal. Una variable numérica discreta como la edad, variable cuantitativa en su naturaleza puede ser convertida en una variable cualitativa cuando por ejemplo en el estudio no nos interese más que conocer en que tramo de edad se encuentre el individuo y por ejemplo lo clasifiquemos en menores de 25 años, entre 25 y 45 años y mayores de 45 años. Son tres tramos de edad y a cada uno de ellos le asignaremos un valor arbitrario. Además serán datos de serie temporal o de sección cruzada.

Una vez recogidos los datos de interés y asignados los arbitrarios si es necesario podemos proceder a analizarlos gráficamente.

### **1.5.3. Representación gráfica de las variables**

Las variables categóricas suelen representarse utilizando tablas de distribución de frecuencias y gráficos como los de barras, de Pareto o de tarta. Las variables numéricas se describen por histogramas, ojivas y diagramas de tallo y hojas.

La distribución de frecuencias o tabla de frecuencias es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente. En principio, en la tabla de

frecuencias se detalla cada uno de los valores diferentes en el conjunto de datos junto con el número de veces que aparece, es decir, su *frecuencia*. Se puede complementar la frecuencia absoluta con la denominada frecuencia relativa, que indica la frecuencia en porcentaje sobre el total de datos.

La tabla de frecuencias puede representarse gráficamente en un histograma o diagrama de barras. Normalmente en el eje vertical se colocan las frecuencias y en el horizontal los intervalos de valores.

### Ejemplo 1.1

La Tabla 1.1 muestra el número de aprobados en tres asignaturas de la Licenciatura en Economía en los cursos 2009-2010 y 2010-2011. Las clases que muestra la tabla son las respuestas posibles a la variable categórica. La pregunta realizada es ¿Cuántos alumnos han aprobado la asignatura  $XX$  en el curso  $YY$ ?

Asignatura	2009-2010	2010-2011
Microeconomía	60	76
Estadística para economistas	89	98
Econometría	85	105

Tabla 1.1: N° de aprobados en tres asignaturas de la Licenciatura en Economía.

La misma información muestra el gráfico de barras recogido en la Figura 1.2:

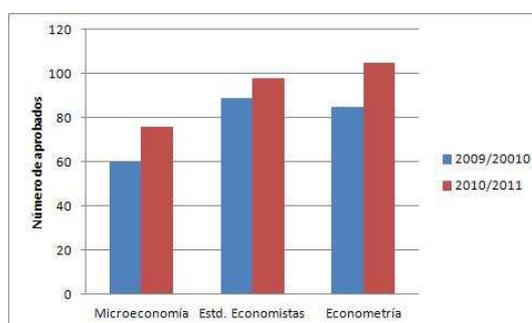


Figura 1.2: N° de aprobados en tres asignaturas de la Licenciatura en Economía.

### Ejemplo 1.2

La Tabla 1.2 muestra el gasto en sanidad,  $GS$ , en billones de dólares en el año 1993 para los 50 estados de Estados Unidos junto con Washington DC:

Con los datos se ha construido la Tabla 1.3 de frecuencias. Se han construido cinco intervalos para los posibles valores de la variable. Para cada uno de ellos se ha calculado su punto medio y el número de observaciones que caen dentro del intervalo y que aparecen en la columna *frecuencia*. La columna *relativa* muestra la frecuencia relativa es decir el

porcentaje sobre el total de observaciones del número de observaciones que caen dentro del intervalo. La columna *acumulada* agrega la frecuencia relativa del intervalo a la del anterior.

Observaciones 1-51, número de cajas = 7,  
media = 15,2649, desv.típ.=17,8877

A partir de la tabla de frecuencias se pueden dibujar los siguientes gráficos de las distribuciones, a la izquierda de frecuencia y a la derecha de frecuencia acumulada:

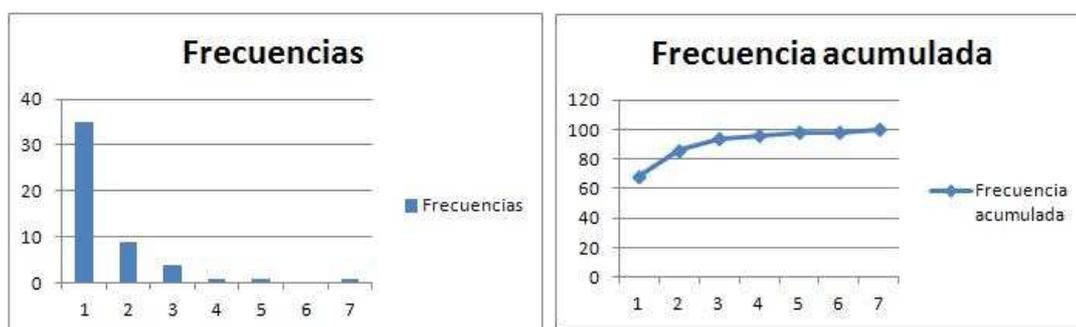


Figura 1.3: Distribución de frecuencias del gasto en sanidad

El gráfico de la izquierda es un histograma cuyos intervalos se corresponden con los de la tabla de distribución de frecuencias. La altura de la barra es proporcional al número de observaciones del intervalo. El gráfico de la derecha es una ojiva, muestra una línea que conecta los puntos que son el porcentaje acumulado de observaciones situadas por debajo del límite superior de cada intervalo en una distribución de frecuencias acumuladas.

En el histograma encontramos información sobre la muestra. Si la muestra está equilibrada, es decir si las observaciones se distribuyen de forma uniforme a ambos lados del punto medio, la forma del histograma será simétrica. En el ejemplo la muestra claramente no está equilibrada, esta sesgada positivamente ya que se extiende hacia la derecha donde los valores son positivos. Si se extiende hacia los valores negativos se dice que está sesgada negativamente.

#### Gráficas para datos de serie temporal

En el caso de disponer de datos de serie temporal además de analizar su distribución de frecuencias nos interesará su evolución temporal.

### Ejemplo 1.3

Se dispone de una muestra de observaciones sobre la demanda doméstica de electricidad para el periodo de 1972:2 a 1993:4 en San Diego en millones de kilowatios/hora. Las observaciones se recogen en la Tabla 1.4:

Estado	GS	Estado	GS	Estado	GS	Estado	GS	Estado	GS
WY	0,998	NH	3,452	MS	6,187	LA	13,014	GA	20,104
VT	1,499	HI	3,485	IA	7,341	MN	14,194	NC	18,241
DC	4,285	ME	3,433	OR	7,999	MD	15,154	NJ	25,741
AK	1,573	NV	3,747	OK	8,041	WI	14,502	MI	27,136
ND	2,021	NE	4,400	CT	12,216	TN	16,203	OH	33,456
DE	2,260	NM	3,878	CO	10,066	MO	15,949	IL	34,747
SD	1,953	WV	5,197	SC	9,029	WA	15,129	PA	41,521
MT	2,103	UT	4,118	KY	10,384	IN	16,401	FL	44,811
RI	3,428	AR	6,111	AZ	10,635	MA	23,421	TX	49,816
ID	2,277	KS	6,903	AL	12,060	VA	16,682	NY	67,033
CA	94,178								

Tabla 1.2: Gasto sanitario por estado

intervalo	punto medio	frecuencia	relativa	acumulada
< 15,530	7,7650	35	68,63 %	68,63 %
15,530 - 31,060	23,295	9	17,65 %	86,27 %
31,060 - 46,590	38,825	4	7,84 %	94,12 %
46,590 - 62,120	54,355	1	1,96 %	96,08 %
62,120 - 77,650	69,885	1	1,96 %	98,04 %
77,650 - 93,180	85,415	0	0,00 %	98,04 %
>= 93,180	100,94	1	1,96 %	100,00 %

Tabla 1.3: Distribución de frecuencias para el gasto en sanidad

t	DDE	t	DDE	t	DDE	t	DDE	t	DDE
1972:2	586,608	1976:4	860,741	1981:2	883,718	1985:3	1062,046	1989:4	1279,626
1972:3	625,797	1977:1	946,673	1981:3	958,449	1985:4	1079,565	1990:1	1485,453
1972:4	704,154	1977:2	793,593	1981:4	980,453	1986:1	1137,534	1990:2	1206,455
1973:1	817,206	1977:3	820,803	1982:1	1062,744	1986:2	992,370	1990:3	1394,055
1973:2	667,642	1977:4	876,286	1982:2	887,136	1986:3	1087,314	1990:4	1336,979
1973:3	661,827	1978:1	993,199	1982:3	933,269	1986:4	1107,644	1991:1	1455,121
1973:4	732,839	1978:2	842,539	1982:4	978,160	1987:1	1263,671	1991:2	1210,595
1974:1	796,876	1978:3	893,369	1983:1	1020,435	1987:2	1054,402	1991:3	1261,056
1974:2	655,335	1978:4	998,721	1983:2	917,487	1987:3	1107,669	1991:4	1407,925
1974:3	692,599	1979:1	1135,091	1983:3	975,387	1987:4	1214,441	1992:1	1446,215
1974:4	768,172	1979:2	901,289	1983:4	998,109	1988:1	1335,147	1992:2	1255,008
1975:1	877,881	1979:3	942,471	1984:1	1015,545	1988:2	1122,214	1992:3	1504,511
1975:2	722,124	1979:4	1024,852	1984:2	919,639	1988:3	1238,305	1992:4	1405,336
1975:3	697,671	1980:1	1068,690	1984:3	1040,188	1988:4	1233,677	1993:1	1512,306
1975:4	804,178	1980:2	914,884	1984:4	1082,732	1989:1	1432,962	1993:2	1241,345
1976:1	880,008	1980:3	930,812	1985:1	1164,663	1989:2	1156,211	1993:3	1405,036
1976:2	763,618	1980:4	969,382	1985:2	945,277	1989:3	1276,944	1993:4	1392,183
1976:3	790,428	1981:1	1028,968						

Tabla 1.4: Demanda doméstica de electricidad

La distribución de frecuencias correspondiente se muestra en la Tabla 1.5.

Observaciones 1-87, número de cajas = 9,  
media = 1035,08, desv.típ.=233,605

intervalo	punto medio	frecuencia	relativa	acumulada
< 644,46	586,61	2	2,30 %	2,30 %
644,46 - 760,18	702,32	8	9,20 %	11,49 %
760,18 - 875,89	818,03	10	11,49 %	22,99 %
875,89 - 991,60	933,74	20	22,99 %	45,98 %
991,60 - 1107,3	1049,5	16	18,39 %	64,37 %
1107,3 - 1223,0	1165,2	10	11,49 %	75,86 %
1223,0 - 1338,7	1280,9	10	11,49 %	87,36 %
1338,7 - 1454,4	1396,6	7	8,05 %	95,40 %
>= 1454,4	1512,3	4	4,60 %	100,00 %

Tabla 1.5: Distribución de frecuencias para la demanda doméstica de electricidad

A continuación se muestra la gráfica de la distribución de frecuencias relativas, a la izquierda, y a la derecha la evolución temporal de la variable. El gráfico de frecuencias muestra una distribución moderadamente asimétrica positiva. El gráfico de serie temporal mostrado a la derecha indica una tendencia creciente en la variable y un patrón de comportamiento anual que se repite.

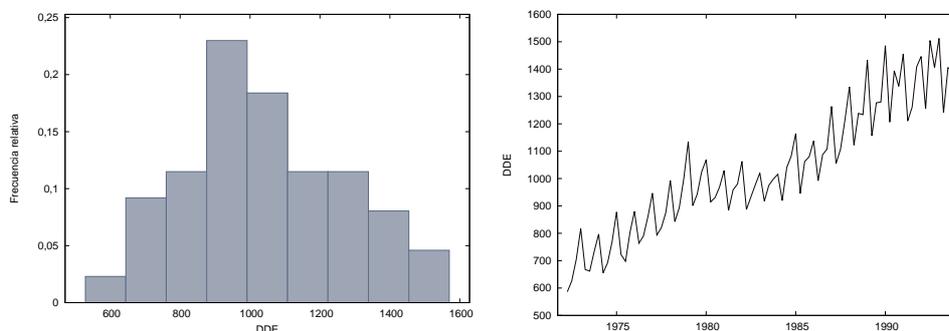


Figura 1.4: Distribución de frecuencias relativas y evolución temporal de la demanda doméstica de electricidad

#### Gráficas para describir relaciones entre variables

En muchas ocasiones los estudios se centran en las relaciones entre dos variables. En este caso la gráfica de la nube de puntos informa tanto del rango como de la relación entre las variables. El gráfico muestra la nube de puntos o pares de observaciones de las variables.

Por ejemplo podemos estudiar la relación entre el salario de un individuo y su antigüedad en la empresa, la dependencia del precio de una vivienda con respecto a la superficie de la misma, la relación entre consumo y renta, etc. En todos estos ejemplos una variable depende de la otra, el

consumo depende de la renta, el salario depende de la antigüedad en el puesto, el precio depende de la superficie de la vivienda. A las variables consumo, precio de la vivienda y salario se les llama variables dependientes porque *depende de* la renta, la superficie de la vivienda y la antigüedad mientras que estas se denominan variables independientes ya que no dependen de otra. La variable independiente (denotada habitualmente por  $X$ ) se muestra en el eje de abscisas mientras que la variable dependiente (denotada habitualmente por  $Y$ ) se muestra en el eje de ordenadas.

### Ejemplo 1.4

Supongamos que disponemos datos sobre consumo per cápita y renta per cápita en dólares de 1987 para Estados Unidos en el periodo de 1970-1991. En el gráfico de la izquierda se muestra la evolución temporal de ambas variables. A la derecha se muestra la nube de puntos de ambas variables. La muestra consta de 22 observaciones luego la nube está constituida por 22 puntos, para cada año un punto recoge el valor de la variable consumo y el valor de la variable renta ( $PPCE_t, PDPI_t$ ).

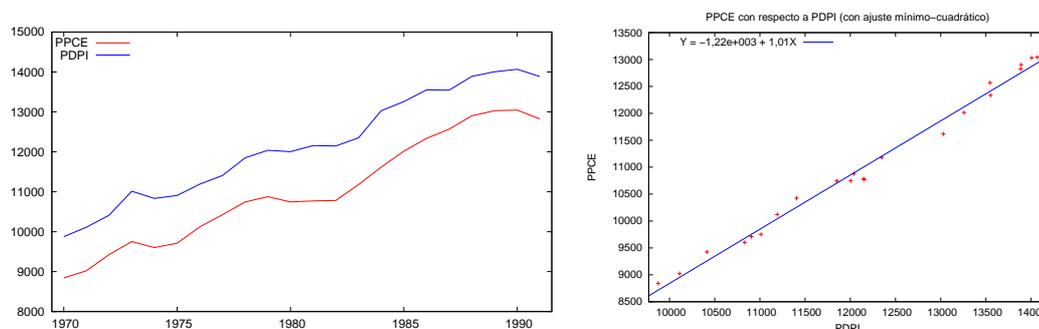


Figura 1.5: Evolución temporal del consumo y renta per cápita junto a la nube de puntos

El gráfico de serie temporal muestra que ambas variables evolucionan de forma pareja y son crecientes con  $t$ . La nube de puntos muestra la misma relación creciente y lineal entre las variables. La recta que aparece en la nube de puntos muestra esta relación lineal. A esta recta se le llama función de regresión muestral y nos ocuparemos de ella en el tema 5.

#### 1.5.4. Descripción numérica de los datos

En la sección anterior se han descrito los datos de una muestra gráficamente. En esta sección vamos a describirlos numéricamente con medidas de tendencia central, medidas de variabilidad y medidas del sentido y del grado de relación entre dos variables.

**Medidas de la tendencia central:** Las medidas de tendencia central dan información numérica sobre una observación “típica” de los datos.

- Media aritmética o media muestral: La media aritmética de un conjunto de datos es la suma de

los valores de los datos dividida por el número de observaciones.

$$\bar{X} = \frac{\sum_1^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- Mediana: es la observación que ocupa el lugar central de un conjunto de observaciones ordenadas en sentido ascendente o descendente. Si el tamaño de muestra  $n$  es impar la mediana es el valor que se encuentra en medio. Si el tamaño muestral es par la mediana es la media de las dos observaciones que se encuentran en el medio. La mediana se encuentra en la posición  $0,50(n + 1)$ .
- Moda, es el valor que aparece con más frecuencia. Puede no existir.

¿Cuál de las tres es la mejor medida para describir la tendencia central de los datos, la media, la mediana o la moda? Depende de los datos. Si los datos son numéricos en general será útil la media, pero si son categóricos será más útil la mediana o la moda. Por ejemplo imaginemos que tenemos cinco individuos dos hombres y tres mujeres y que el investigador asocia el valor 1 a los hombres y el valor cero a las mujeres. La media muestral de la variable “sexo” es  $2/5$  lo que no tiene sentido, sin embargo la moda es el valor 0 que indica que hay más mujeres que hombres en la muestra. Imaginemos que la variable objeto de estudio es la renta. La media se verá incrementada por aquellos valores muy altos y no describirá bien la variable sin embargo la mediana es el nivel de renta por encima del cual están la mitad de los hogares de la muestra.

**Medidas de la variabilidad:** Junto a la media es necesario presentar estadísticos descriptivos que midan la variabilidad o dispersión de los datos u observaciones con respecto a la media.

- Rango: es la diferencia entre los valores mayor y menor de la muestra.
- Varianza muestral: la varianza muestral es la suma de los cuadrados de las diferencias entre cada observación y la media muestral dividida por el tamaño de la muestra menos 1:

$$S^2 = \frac{\sum_1^n (X_i - \bar{X})^2}{n - 1}$$

- Desviación típica muestral: Es la raíz de la varianza muestral:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_1^n (X_i - \bar{X})^2}{n - 1}}$$

- Coeficiente de variación: es una medida de la dispersión relativa que expresa la desviación típica en porcentaje de la media (siempre que esta sea positiva). El coeficiente de variación muestral se define:

$$CV = \frac{S}{\bar{X}} \times 100\% \quad \text{si } \bar{X} > 0$$

### Medidas de las relaciones entre las variables

- Covarianza: es una medida de la relación entre dos variables. Un valor positivo indica una relación lineal directa o creciente y un valor negativo una relación lineal inversa o decreciente. La covarianza muestral se define:

$$Cov(X, Y) = S_{XY} = \frac{\sum_1^n (X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

donde  $X_i$  e  $Y_j$  son los valores observados de las variables  $X$  e  $Y$ ,  $\bar{X}$  e  $\bar{Y}$  son, respectivamente, sus medias muestrales y  $n$  es el tamaño de muestra. El coeficiente de correlación da una medida estandarizada de la relación lineal entre dos variables. Indica el sentido y el grado de la relación.

- Coeficiente de correlación: es el cociente entre la covarianza y el producto de las desviaciones típicas entre las variables:

$$r_{XY} = \frac{Cov(X, Y)}{S_X S_Y}$$

El coeficiente de correlación está comprendido entre  $-1$  y  $1$ ,  $-1 \leq r_{XY} \leq 1$ . Cuanto más cerca se encuentra de  $1$  más cerca se encuentran los datos de puntos de una línea recta ascendente que indica una *relación lineal positiva*. Cuanto más cerca de  $-1$  más cerca se encuentran los datos de puntos de una línea recta descendente que indica una *relación lineal negativa*. Cuando  $r = 0$  no existe ninguna *relación lineal* entre las variables.

## 1.6. Bibliografía del tema

### Referencias bibliográficas básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con gretl:

- [1] Ramanathan, R. (2002), Instructor's Manual to accompany, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), Student Solutions Manual, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

### Referencias Bibliográficas Complementarias:

- [1] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). Análisis de regresión con gretl. OpenCourseWare. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-gretl/Courseisting>).
- [2] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación online de la Facultad de C.C. Económicas y Empresariales.
- [3] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones.

- [4] Esteban, MV (2008). Estadística Actuarial: Regresión Lineal, Sarriko On Line 3/08.Publicación on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.
- [5] Esteban, M.V. (2007). Colección de ejercicios y exámenes. Material docente. Servicio de Publicaciones.
- [6] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.
- [7] Greene, W. (1998), Análisis Económico, Ed. Prentice Hall, 3ª edición.
- [8] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.
- [9] Verbeek, M. (2004). A Guide to Modern Econometrics. Wiley.



## Tema 2

# Modelo de Regresión Lineal General: especificación

En este tema nos ocuparemos de analizar las relaciones entre variables y nuestro objetivo fundamental será explicar el comportamiento de una variable, que llamaremos variable a explicar, mediante un conjunto de variables económicas, que llamaremos explicativas. Modelizaremos la relación entre las variables mediante una ecuación matemática y daremos entrada en la misma a una variable aleatoria que nos permita recoger la aleatoriedad del fenómeno económico. Así, aprenderemos a especificar el Modelo de Regresión Lineal General, poniendo especial cuidado en el tratamiento de las variables explicativas cualitativas.

### Competencias a trabajar en estas sesiones:

- C1. Analizar de forma crítica los elementos básicos del modelo de regresión lineal con el objetivo de comprender la lógica de la modelización econométrica y poder especificar relaciones causales entre las variables.
- C4. Presentar de forma clara y concisa, tanto oralmente como por escrito, las conclusiones obtenidas en una aplicación empírica.

### Al final de este tema deberíais ser capaces de:

1. Explicar y entender el alcance de las hipótesis básicas sobre el comportamiento del modelo de regresión lineal general (C1).
2. Interpretar los parámetros de un modelo (C1).
3. Saber especificar correctamente modelos que incluyan variables cualitativas (C1).
4. Comprender la especificación del modelo de regresión lineal y, en particular, el significado y las implicaciones de los supuestos básicos (C1).
5. Saber incorporar en el modelo de regresión variables cuantitativas y cualitativas (C1).

6. Interpretar los coeficientes del modelo de regresión, incluyendo los de especificaciones no lineales en las variables (C1).
7. Organizar y sistematizar información estadística relevante (C4).
8. Utilizar un software econométrico (Gretl) para el análisis de bases de datos económicos e interpretar sus resultados (C1).

**Bibliografía Recomendada:**

Al final del tema tenéis recogida la bibliografía correspondiente. En particular se os recomienda leer los capítulos correspondientes a la bibliografía básica detallados a continuación:

- Gujarati, D. y Porter, D.C. (2010). En Parte I: Caps. 1, 2, 3, 4, 5, 7, 8 y 9. Parte II: Cap. 10.
- Ramanathan, R. (2002). Caps. 3, 4, 5 y 7.
- Wooldridge, J.M. (2006). Caps. 2, 3, 4 y 7

## 2.1. Especificación del Modelo de Regresión Lineal General (MRLG): supuestos básicos

En Economía, en muchas situaciones, varias variables independientes influyen conjuntamente en una variable dependiente. El modelo de regresión múltiple permite averiguar el efecto simultáneo de varias variables independientes en una variable dependiente. Por ejemplo:

- El salario es una función del nivel de estudios, la experiencia, la edad y el puesto de trabajo.
- El precio de un piso es función, entre otras características, de su superficie, número de habitaciones y baños, localización y la existencia o no de ascensor.
- La cantidad vendida de un bien depende de su precio, del precio de la competencia y del ciclo económico entre otras variables.
- La producción de una empresa depende de los factores de producción, capital y fuerza de trabajo.

La especificación de un modelo consiste en seleccionar las variables independientes que explican a la variable objeto de estudio y determinar la forma funcional del mismo. Vamos a comenzar el análisis de regresión determinando nuestro objetivo y los recursos disponibles para lograrlo.

**Objetivo:** Cuantificar la relación existente entre una variable dependiente a la que denotaremos por  $Y$ , y un conjunto de  $K$  variables independientes,  $X_1, X_2, \dots, X_K$  mediante la especificación de un modelo lineal.

**Recursos disponibles:** Se dispone de una muestra de observaciones de las variables  $Y, X_1, X_2, \dots, X_K$  de tamaño  $N$ , que es el número de observaciones disponibles sobre todas las variables. Se denota:

$$\begin{aligned} Y_i &= \text{observación } i\text{-ésima de } Y \\ X_{ki} &= \text{observación } i\text{-ésima de } X_k \quad \forall k = 1, \dots, K \end{aligned}$$

donde  $X_{ki}$  es una observación de las disponibles en la muestra  $i = 1, 2, \dots, N$ .

### Ejemplo 2.1

El objetivo de este ejemplo es proponer un modelo económico para una cadena de comida rápida de USA. El gerente de dicha cadena ha de tomar decisiones sobre su política de precios y el gasto en publicidad. Para valorar el efecto en sus ventas de diferentes estructuras de precios y diferentes niveles de gasto en publicidad la cadena fija precios y gasto en publicidad diferentes en las distintas ciudades en que está implantada. Uno de sus objetivos es analizar cómo cambian sus ingresos por ventas cuando cambia el nivel de gasto en publicidad. ¿Un incremento en los gastos en publicidad se traduce en un incremento en ventas? Si esto ocurre así, ¿el incremento en las ventas es suficiente para

justificar el incremento en el gasto en publicidad? Su otro objetivo fundamental es fijar una adecuada política o estrategia de precios, ¿una reducción en el precio lleva a un incremento o decrecimiento de los ingresos por ventas? Si la reducción en precios lleva solo a un pequeño incremento en la cantidad vendida, los ingresos por ventas caerán (demanda inelástica en precio) pero si una reducción en el precio conlleva un gran incremento en la cantidad vendida, los ingresos por ventas crecerán (demanda elástica en precio).

Para proponer un modelo económico que describa el comportamiento de las ventas de la cadena vamos a empezar suponiendo que las ventas se relacionan linealmente con el precio del producto y el gasto en publicidad. La ecuación que recoge al modelo económico es:

$$S = \beta_1 + \beta_2 P + \beta_3 A \quad (2.1)$$

Donde  $S$  son las ventas mensuales en una de las ciudades en que está implantada la cadena,  $P$  es el precio del producto en dicha ciudad y  $A$  el gasto mensual en publicidad en la ciudad referida. Se analiza el comportamiento de ciudades con poblaciones comparables ya que obviamente las ventas en grandes ciudades son mayores que las ventas en ciudades pequeñas.

Por otro lado hemos de reflexionar sobre cómo medir la variable  $P$ . Un local de comida rápida ofrece un buen número de productos alternativos: hamburguesas, pizzas, pollo rebozado, aritos, shakes, etc cada uno con su propio precio y no está claro cual es el precio de referencia a elegir. Lo más adecuado es tomar un precio medio de todos los productos. Necesitamos datos sobre ese precio medio y cómo cambia de ciudad en ciudad. Para ello el gerente construye un índice de precios de todos los productos vendidos en el mes, medido en dólares, para cada ciudad donde la cadena está implantada, la variable  $P$ . Las ventas mensuales y el gasto mensual en publicidad en la ciudad se miden en miles de dólares.

**Modelo de Regresión lineal General (MRLG). Modelización** El Modelo de Regresión Lineal General se escribe:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N$$

donde habitualmente  $X_{1i} = 1 \forall i$ , de forma que  $\beta_1$  es un término independiente y entonces,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N.$$

### Elementos del MRLG

- $Y$  es la variable a explicar, variable dependiente o endógena.
- $X_k$   $k = 1, \dots, K$  son las  $K$  variables explicativas, variables independientes o exógenas.

- $\beta_k$   $k = 1, \dots, K$  son los coeficientes de la regresión o parámetros (desconocidos).
- $u$  es la perturbación aleatoria o término de error.
- el subíndice  $i$  denota la observación correspondiente. El subíndice  $i$  se utiliza cuando tenemos observaciones de sección cruzada y el subíndice  $t$  cuando tenemos observaciones de serie temporal.
- $N$  es el tamaño muestral, el número de observaciones disponibles de las variables objeto de estudio. Cuando trabajamos con datos de serie temporal el tamaño muestral se denota por  $T$ .

La perturbación aleatoria  $u_i$  es una variable aleatoria no observable que pretende recoger:

- Variables no incluidas en el modelo.
- Comportamiento aleatorio de los agentes económicos.
- Errores de medida.

**Representación del MRLG en forma matricial** El modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N \quad (2.2)$$

puede escribirse para todas las observaciones disponibles como el siguiente sistema de  $N$  ecuaciones:

$$\begin{cases} Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_K X_{K1} + u_1 & i = 1 \\ Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_K X_{K2} + u_2 & i = 2 \\ \vdots & \vdots \\ Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + u_i & i = i \\ \vdots & \vdots \\ Y_N = \beta_1 + \beta_2 X_{2N} + \beta_3 X_{3N} + \dots + \beta_K X_{KN} + u_N & i = N \end{cases}$$

o bien en forma matricial como

$$\begin{matrix} Y & = & X & \beta & + & u \\ (N \times 1) & & (N \times K) & (K \times 1) & & (N \times 1) \end{matrix}$$

donde

$$\begin{matrix} \begin{matrix} Y \\ (N \times 1) \end{matrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{bmatrix} & \begin{matrix} X \\ (N \times K) \end{matrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{K1} \\ 1 & X_{22} & X_{32} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{2i} & X_{3i} & \cdots & X_{Ki} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{2N} & X_{3N} & \cdots & X_{KN} \end{bmatrix} & \begin{matrix} \beta \\ (K \times 1) \end{matrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} & \begin{matrix} u \\ (N \times 1) \end{matrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix} \end{matrix}$$

### 2.1.1. Hipótesis básicas.

El modelo debe completarse con la especificación de las propiedades estocásticas de la variable de interés  $Y$ . A partir de las propiedades de  $Y$  es posible conocer las propiedades de los distintos métodos de estimación, elegir el mejor estimador en el modelo, realizar contrastes, etc. Las condiciones bajo las cuales vamos a trabajar en un principio se denominan **hipótesis básicas**. Bajo estas hipótesis estimaremos y analizaremos el modelo para, finalmente, predecir  $Y$ . En una segunda etapa, podemos considerar otras situaciones, relajando algunas de estas hipótesis, analizando si los procedimientos de estimación y contraste anteriores siguen siendo válidos. Las hipótesis básicas se refieren a los distintos elementos de la regresión.

#### 1. Hipótesis sobre la perturbación aleatoria

- La perturbación  $u_i$  tiene media cero para todo  $i$ ,  $E(u_i) = 0 \quad \forall i$ . La perturbación mide las diferencias con respecto a un promedio,  $u_i = Y_i - E(Y_i)$  y a priori no tenemos razones para suponer que todas las desviaciones están por encima o por debajo de ese promedio, por ello parece lógico pensar que en media las desviaciones son cero. Para la perturbación en  $i$  lo escribimos como  $E(u_i) = 0 \quad \forall i$ , cuando miramos al modelo en forma matricial escribimos esta hipótesis como  $E(u) = \vec{0}$ :

$$E(u) = E \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_N) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}$$

- $E(u_i^2) = \sigma_u^2 \quad \forall i$  es decir la varianza de la perturbación es desconocida e igual a  $\sigma^2$  para todas las observaciones. Estamos suponiendo igual dispersión o variabilidad. A esta hipótesis se le conoce con el nombre de *Homocedasticidad*:

$$V(u_i) = E(u_i - E(u_i))^2 = E(u_i^2) = \sigma^2$$

$$E(u_1^2) = E(u_2^2) = E(u_3^2) = \dots = E(u_N^2) = \sigma^2$$

El caso contrario, cuando la dispersión varía a lo largo de la muestra se denomina *Heterocedasticidad*:  $E(u_i^2) = \sigma_i^2$ . La Figura 2.1 ilustra ambas situaciones:

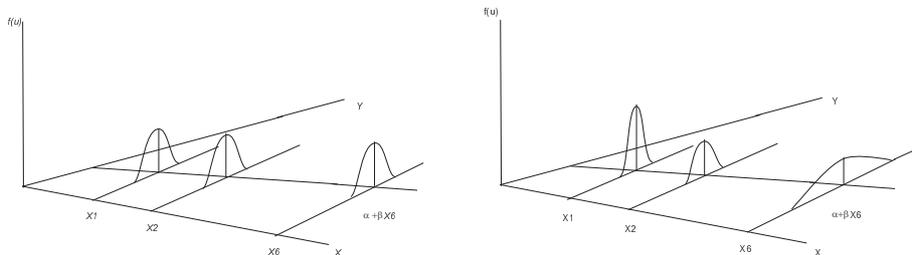


Figura 2.1: Perturbaciones homocedásticas versus heterocedásticas

Hay que notar que generalmente  $\sigma^2$  será desconocida y por tanto en el modelo tendremos que estimar  $(K + 1)$  incógnitas, los  $k$ -coeficientes poblacionales desconocidos más la varianza poblacional de la perturbación  $\sigma^2$ .

- $E(u_i u_j) = 0 \quad \forall i, j \quad i \neq j$ . La covarianza entre perturbaciones de distintas observaciones es cero.

$$Cov(u_i, u_j) = E(u_i - E(u_i))(u_j - E(u_j)) = E(u_i u_j) = 0 \quad \forall i, j \quad i \neq j$$

$$E(u_1 u_2) = E(u_1 u_3) = E(u_4 u_{20}) = 0$$

A esta hipótesis también se la llama hipótesis de *No Autocorrelación*. Uniendo la hipótesis de homocedasticidad y la hipótesis de no autocorrelación podemos describir la matriz de varianzas y covarianzas de la perturbación.

$$E(uu') = \sigma^2 I_N$$

$$\begin{aligned} E(uu') &= E \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} [ u'_1 \quad u'_2 \quad \dots \quad u'_N ] = E \begin{bmatrix} u_1 u'_1 & u_1 u'_2 & \dots & u_1 u'_N \\ u_2 u'_1 & u_2 u'_2 & \dots & u_2 u'_N \\ \vdots & \vdots & \ddots & \vdots \\ u_N u'_1 & u_N u'_2 & \dots & u_N u'_N \end{bmatrix} = \\ &= E \begin{bmatrix} u_1^2 & u_1 u'_2 & \dots & u_1 u'_N \\ u_2 u'_1 & u_2^2 & \dots & u_2 u'_N \\ \vdots & \vdots & \ddots & \vdots \\ u_N u'_1 & u_N u'_2 & \dots & u_N^2 \end{bmatrix} = \begin{bmatrix} E(u_1^2) & E(u_1 u'_2) & \dots & E(u_1 u'_N) \\ E(u_2 u'_1) & E(u_2^2) & \dots & E(u_2 u'_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_N u'_1) & E(u_N u'_2) & \dots & E(u_N^2) \end{bmatrix} = \\ &= \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I_N \end{aligned}$$

A la hipótesis que reconoce que las varianzas de la perturbación no son constantes en el tiempo o las observaciones se le conoce como hipótesis de *Heterocedasticidad*. A la hipótesis que reconoce que las covarianzas entre perturbaciones de distinto momento del tiempo, o entre distintas observaciones, son distintas de cero se le conoce con el nombre de *Autocorrelación*.

- Las perturbaciones siguen una distribución normal. Si definimos la perturbación aleatoria como la suma de errores independientes entre sí, a través del Teorema Central del

Límite podremos suponer una distribución normal y escribir esta hipótesis junto con las anteriores como:

$$u \sim NID(0, \sigma^2 I_N)$$

donde estamos escribiendo la distribución del vector de perturbaciones  $u$  y decimos que las perturbaciones siguen una distribución normal, idéntica e independientemente distribuidas, de media cero y varianza constante igual a  $\sigma^2$ . Son independientes dado que su covarianza es cero y dado que todas tienen igual varianza y covarianza su distribución es idéntica, por ello para una perturbación en  $i$  escribimos su distribución como  $u_i \sim N(0, \sigma^2)$ .

Estas propiedades pueden también escribirse conjuntamente como

$$u_i \sim NID(0, \sigma_u^2) \quad \forall i = 1, \dots, N$$

ó en forma matricial,

$$\begin{matrix} u \\ (N \times 1) \end{matrix} \sim N \left( \begin{matrix} 0_N \\ (N \times 1) \end{matrix}, \begin{matrix} \sigma_u^2 I_N \\ (N \times N) \end{matrix} \right)$$

## 2. Hipótesis sobre las variables exógenas $X$ .

- Las variables explicativas son variables no aleatorias (no estocásticas o fijas). Esto quiere decir que cuando digamos que cambiamos la muestra los valores de las variables exógenas no cambian y sólo cambian los valores de la variable endógena  $Y_i$ . Como consecuencia de que las variables exógenas sean fijas tendremos que son incorrelacionadas con las perturbaciones:

$$E(X'u) = X'E(u) = 0$$

- La matriz  $X$  es de rango completo e igual a  $K$  con  $K < N$ ,  $rg(X) = K$ , es decir no hay ninguna combinación lineal exacta entre las columnas de  $X$ , son todas linealmente independientes con lo que el rango de la matriz es igual al número de coeficientes desconocido ya que en  $X$  tenemos una columna por parámetro. A esta hipótesis se le conoce con el nombre de *No Multicolinealidad*. El que además exijamos que  $K < N$  es porque necesitamos tener más observaciones que coeficientes a estimar en el modelo.

## 3. Hipótesis sobre la forma funcional.

- Linealidad en los coeficientes.
- Modelo correctamente especificado. Todas las variables  $X_1, X_2, \dots, X_K$  explican  $Y$  y no hay ninguna otra de fuera del modelo que explique a  $Y$ .

## 4. Los coeficientes permanecen constantes a lo largo de toda la muestra.

## 2.2. Función de Regresión Poblacional. Interpretación de los coeficientes.

Dados los supuestos básicos del MRLG,

$$\begin{aligned} E(Y_i) &= E(\beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i) \\ &= \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \underbrace{E(u_i)}_{=0} \\ &= \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki}. \end{aligned}$$

A  $E(Y_i)$  se la denomina **Función de Regresión Poblacional** (FRP) y sus coeficientes, pueden interpretarse como:

- $\beta_1 = E(Y_i | X_{2i} = \dots = X_{Ki} = 0)$ . Valor medio o esperado de  $Y_i$  cuando las variables explicativas son todas cero.
- $\beta_k = \frac{\partial E(Y_i)}{\partial X_{ki}} = \frac{\Delta E(Y_i)}{\Delta X_{ki}} \quad \forall k = 2, \dots, K$ . Incremento (o decremento) en el valor esperado de  $Y_i$  cuando la variable explicativa  $X_k$  se incrementa en una unidad, *manteniéndose constantes el resto de las variables*. Un aumento unitario en la variable explicativa  $X_k$  conlleva un aumento medio de  $\beta_k$  unidades en la variable endógena, *ceteris paribus*.

### Ejemplo 2.2

En el ejemplo de la cadena de comida rápida habíamos propuesto un modelo económico que recogía una relación lineal entre las ventas mensuales, el precio medio mensual de todos los productos y el gasto en publicidad mensual para una ciudad dada en la que está enclavada la cadena:

$$S = \beta_1 + \beta_2 P + \beta_3 A \quad (2.3)$$

Además se dispone de las observaciones de dichas variables en un mes concreto para un conjunto de 75 ciudades. Si añadimos el término de perturbación obtenemos el siguiente modelo econométrico:

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + u_i \quad i = 1, \dots, 75 \quad (2.4)$$

$\beta_1, \beta_2$  y  $\beta_3$  son los parámetros desconocidos cuyo valor queremos estimar.

- El coeficiente  $\beta_1$  es el valor esperado de las ventas cuando el precio y el gasto en publicidad es cero  $\beta_1 = E(S_i | P_i = A_i = 0)$ .

- El coeficiente  $\beta_2 = \frac{\partial E(S_i)}{\partial P_i}$ , mide el cambio esperado en las ventas cuando el precio medio cambia en una unidad permaneciendo el gasto en publicidad constante. Luego, considerando dos ciudades con el mismo gasto en publicidad, para aquella ciudad que tenga un precio medio una unidad más caro se espera que sus ventas cambien en media en  $\beta_2$  miles de dólares.
- El coeficiente  $\beta_3 = \frac{\partial E(S_i)}{\partial A_i}$ , mide el cambio esperado en las ventas cuando el gasto en publicidad cambia en una unidad permaneciendo el precio medio constante. Luego, considerando dos ciudades con el mismo precio medio, para aquella ciudad que tenga un un gasto en publicidad una unidad más alto se espera que cambie en media sus ventas en  $\beta_3$  miles de dólares.

### Ejemplo 2.3

Se propone la siguiente especificación de la función de consumo agregada para estudiar la relación en Estados Unidos en el periodo 1960-2005 entre el consumo personal, GCP, y el ingreso, PIB, ambos en miles de millones de dólares:

$$GCP_t = \beta_1 + \beta_2 PIB_t + u_t$$

$\beta_2$  recoge el incremento esperado en el consumo personal por unidad de incremento en el  $PIB$ . Además tiene interpretación económica ya que es la propensión marginal a consumir que según la teoría keynesiana esta limitada entre 0 y 1.  $\beta_1$  es el valor esperado del consumo cuando el valor del  $PIB$  es cero.

### Ejemplo 2.4

Se dispone de una base de datos para 51 estados de E.E.U.U. sobre el gasto agregado en transporte urbano ( $EXPTRAV$ ) y la renta disponible agregada ( $INCOME$ ) correspondientes al año 1993<sup>1</sup>. Las variables que se consideran son:

$EXPTRAV$  = Gasto agregado en transporte urbano, en billones de dólares, (Rango 0,708 - 42,48).

$INCOME$  = Renta disponible agregada, en billones de dólares, (Rango 9,3 - 683,5).

Un modelo para analizar si la renta disponible agregada explica el gasto agregado en transporte urbano es el siguiente<sup>2</sup>:

$$EXPTRAV_i = \beta_1 + \beta_2 INCOME_i + u_i \quad i = 1, \dots, 51 \quad (2.5)$$

El parámetro  $\beta_1$  recoge el valor esperado del gasto en transporte cuando la renta es cero,  $\beta_1 = E(EXPTRAV | INCOME = 0)$ . La pendiente  $\beta_2$  recoge el incremento en el valor

<sup>1</sup>Fuente: Statistical Abstract of U.S. (1995), recogida en Ramanathan, Ramu (2002) *Introductory econometrics with applications*. Fichero de datos data8-2.gdt.

<sup>2</sup>Son datos de sección cruzada luego utilizamos el subíndice  $i = 1, \dots, N$ .

esperado del gasto en transporte cuando la renta se incrementa en una unidad, es este caso cuando se incrementa en un billón de dólares,  $\beta_2 = \frac{\partial E(EXPTRAV)}{\partial INCOME}$ . Esperaríamos signo positivo.

### Ejemplo 2.5

Estamos interesados en explicar el precio de una vivienda, en miles de dólares (PRICE), mediante las variables explicativas: el tamaño de la casa o el número de pies cuadrados del área habitable (SQFT), el número de habitaciones (BEDRMS) y el número de baños (BATHS). Formulamos el modelo de regresión lineal múltiple:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, 2, \dots, N \quad (2.6)$$

Interpretación de los coeficientes:

- El coeficiente  $\beta_1 = E(PRICE_i | SQFT_i = BEDRMS_i = BATHS_i = 0)$  es el valor medio esperado de aquellas viviendas que no tienen ningún pie cuadrado de área habitable, ni habitaciones ni baños.
- El coeficiente  $\beta_2 = \frac{\partial E(PRICE_i)}{\partial SQFT_i}$ , mide el incremento en el valor esperado del precio de una vivienda cuando su superficie se incrementa en un pie cuadrado, manteniéndose el resto de variables constante. Luego, considerando dos casas con el mismo número de habitaciones y de baños, para aquella casa que tenga un pie cuadrado más de área habitable se espera que cambie en media su precio de venta en  $\beta_2$  miles de dólares.
- El coeficiente  $\beta_3 = \frac{\partial E(PRICE_i)}{\partial BEDRMS_i}$ , mide el incremento en el valor esperado del precio de una vivienda cuando el número de habitaciones de la misma se incrementa en una unidad, manteniéndose el resto de variables constante. Considerando dos casas con el mismo número de pies cuadrados de área habitable y número de baños, para aquella casa que tenga una habitación más se espera que cambie en media su precio de venta en  $\beta_3$  miles de dólares.
- El coeficiente  $\beta_4 = \frac{\partial E(PRICE_i)}{\partial BATHS_i}$ , mide el incremento en el valor esperado del precio de una vivienda cuando el número de habitaciones de la misma se incrementa en una unidad, manteniéndose el resto de variables constante. Considerando dos casas con el mismo número de pies cuadrados de área habitable y número de habitaciones, para aquella casa que tenga un baño más se espera que cambie en media su precio de venta en  $\beta_4$  miles de dólares.

### Ejemplo 2.6

Se especifica la siguiente función de salarios en el año 2002:

$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \quad i = 1, 2, \dots, N$$

donde  $W_i$  es el salario anual del individuo  $i$  y  $S_{2i}$  es una variable ficticia que se define:

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

La interpretación de los coeficientes de regresión del modelo es la siguiente:

- $\beta_1 = E(W_i|S_{2i} = 0)$  luego es el salario esperado cuando el individuo es hombre. Esperaríamos signo positivo.
- $E(W_i|S_{2i} = 1) = \beta_1 + \beta_2$  es el salario esperado de una mujer. Luego  $\beta_2$  es el incremento o decremento en el salario esperado para un individuo por el hecho de ser mujer. Por tanto  $\beta_2$  recoge el efecto diferencial en el salario esperado entre hombres y mujeres. Si es cierto que existe discriminación salarial por sexo esperaríamos que tuviera signo negativo. De la misma forma si no existiera discriminación salarial por sexo, es decir si hombres y mujeres tuvieran el mismo salario, su valor sería cero.

### Ejemplo 2.7

Se especifica la siguiente función de salarios en el año 2002:

$$W_i = \beta_1 + \beta_2 S_{2i} + \beta_3 X_i + u_i \quad i = 1, 2, \dots, N$$

donde  $W_i$  es el salario anual del individuo  $i$ ,  $X_i$  son los años de experiencia del individuo  $i$  y  $S_{2i}$  es una variable ficticia que se define:

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

La interpretación de los coeficientes de regresión del modelo es la siguiente:

- $\beta_1 = E(W_i|S_{2i} = X_i = 0)$  luego es el salario esperado cuando el individuo es hombre y no tiene experiencia. Esperaríamos signo positivo.
- $E(W_i|S_{2i} = 1, X_i = 0) = \beta_1 + \beta_2$  luego  $\beta_2$  es el incremento o decremento en el salario esperado para un individuo cuando no tiene experiencia por el hecho de ser mujer. Por tanto  $\beta_2$  recoge el efecto diferencial en el salario esperado entre hombres y mujeres con igual experiencia laboral. Si es cierto que existe discriminación salarial por sexo esperaríamos que tuviera signo negativo. De la misma forma, si no existiera discriminación salarial por sexo su valor sería cero.
- $\beta_3 = \frac{\partial E(W_i)}{\partial X_i}$  es el incremento en el salario esperado del individuo  $i$  cuando la experiencia se incrementa en un año. Es independiente del sexo del individuo  $i$  luego es el mismo para hombres y mujeres. Esperaríamos signo positivo, a mayor experiencia mayor remuneración.

### Ejemplo 2.8

Se especifica la siguiente función de ventas de una empresa para el período de Enero de 1978 a Diciembre de 2002:

$$V_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \quad t = 1, 2, \dots, T$$

donde  $V_t$  son las ventas de la empresa en el momento  $t$  y las variables  $D_{jt}$  son variables ficticias que se definen:

$$D_{jt} = \begin{cases} 1 & \text{si la observación } t \text{ pertenece al trimestre } j \quad j = 2, 3, 4 \\ 0 & \text{en caso contrario} \end{cases}$$

La interpretación de los coeficientes de regresión del modelo es la siguiente:

- $E(V_t | D_{2t} = D_{3t} = D_{4t} = 0) = \beta_1$  es el valor esperado de las ventas en el primer trimestre.
- $E(V_t | D_{2t} = 1; D_{3t} = D_{4t} = 0) = \beta_1 + \beta_2$  es el valor esperado de las ventas en el segundo trimestre. Luego  $\beta_2$  es el diferencial entre las ventas esperadas en el segundo trimestre y el primer trimestre.
- $E(V_t | D_{3t} = 1; D_{2t} = D_{4t} = 0) = \beta_1 + \beta_3$  es el valor esperado de las ventas en el tercer trimestre. Luego  $\beta_3$  es el diferencial entre las ventas esperadas en el tercer trimestre y el primer trimestre.
- $E(V_t | D_{2t} = D_{3t} = 0; D_{4t} = 1) = \beta_1 + \beta_4$  es el valor esperado de las ventas en el cuarto trimestre. Luego  $\beta_4$  es el diferencial entre las ventas esperadas en el cuarto trimestre y el primer trimestre.

### Ejemplo 2.9

En este ejemplo vamos a analizar los determinantes de la oferta laboral de las mujeres casadas<sup>3</sup>. Para ello vamos a utilizar el archivo *mroz87.gdt* incluido en el programa *gretl*, en la carpeta de archivos de muestra “Gretl”. En él se dispone de observaciones de las siguientes variables, entre otras<sup>4</sup>:

- *LFP*: Variable ficticia que toma valor 1 si la mujer ha trabajado en 1975 y cero en otro caso
- *HOURS*: Número de horas trabajadas por la esposa en 1975, (*WHRS*).
- *KL6*: Número de hijos menores de seis años, en la familia.
- *K618*: Número de hijos entre seis y dieciocho años, en la familia.
- *AGE*: Edad de la esposa, (*WA*).

<sup>3</sup>Fichero *mroz87.gdt*, disponible en *gretl* pestaña Gretl. Fuente: “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions”, *Econometrica* 55, 765-799.

<sup>4</sup>Algunas de las variables han sido renombradas para que el ejercicio resulte más cómodo. Entre paréntesis aparece el nombre original utilizado en el fichero para que las podáis reconocer fácilmente.

- *EDUC*: Años de educación recibidos por la esposa, (*WE*).
- *WAGE*: Salario de la esposa en el momento de la encuesta, 1976, (*RPWG*).
- *FAMINC*: Renta familiar en dólares de 1975.
- *EXPER*: Años de experiencia en la actualidad, (*AX*).

Se trata de una muestra de sección cruzada con 428 observaciones de mujeres trabajadoras donde el término trabajadoras implica que tienen un salario monetario.

Con la muestra anterior creamos las siguientes variables:

$$\begin{aligned} l\_WAGE_i &= \ln(WAGE)_i, \\ sq\_EXPER_i &= EXPER_i^2, \\ NWIFEINC_i &= FAMINC_i - (WAGE_i \times HOURS_i) \end{aligned}$$

Dado nuestro objetivo de estudiar los determinantes de la oferta laboral de la población femenina casada, la variable a estudiar es *HOURS*. Como determinantes de la misma podemos pensar en incluir en el modelo el salario en logaritmos, *l\_WAGE*, los años de educación recibidos, *EDUC*, la edad, *AGE*, el número de hijos de la familia, *KL6* y *K618*, y la variable *NWIFEINC*, que de alguna manera mide la importancia de las rentas familiares que no dependen de los ingresos de la esposa. Así, el modelo a estimar sería:

$$\begin{aligned} HOURS_i &= \beta_1 + \beta_2 l\_WAGE_i + \\ &\beta_3 EDUC_i + \beta_4 AGE_i + \beta_5 KL6_i + \beta_6 K618_i + \beta_7 NWIFEINC_i + u_i \end{aligned} \quad (2.7)$$

A priori esperaríamos que los coeficientes de las variables *l\_WAGE*, *EDUC* y *K618* fuesen positivos, *ceteris paribus*. Es de esperar que si se tiene un sueldo alto la estimulación a seguir trabajando sea mayor. También es lógico pensar que cuando la preparación (estudios) es mayor, se tenga un empleo mejor y mejor remunerado, luego es más probable que la mujer trabaje fuera de casa. Familias con mayor número de hijos necesitan de una mayor renta, por lo que es probable que la esposa decida trabajar. Por otro lado, cuando hay niños pequeños podemos pensar que la esposa se retire del mercado de trabajo para cuidarlos, por lo que esperaríamos un signo negativo en el coeficiente que acompaña a *KL6* manteniendo el resto de factores constante. A mayor edad, también es posible que no se quiera trabajar fuera de casa, por lo que, esperamos signo negativo para la variable *AGE* *ceteris paribus*. Finalmente, manteniendo el resto de factores constantes también esperaríamos signo negativo para el coeficiente de la variable *NWIFEINC* ya que, si la renta disponible que no depende de la remuneración de la mujer es alta, es probable que eso desincentive a trabajar fuera de casa a la esposa. Es importante tener en cuenta la fecha de la recogida de datos, 1975, cuando todavía la participación en el mercado de trabajo de la mujer no estaba consolidada.

**Algunas consideraciones sobre la linealidad en parámetros** Cuando decimos que el MRLG es un modelo lineal queremos decir que  $Y$  o alguna transformación de  $Y$  es lineal en las  $X$  o en alguna transformación lineal de las  $X$ . Hay dos tipos de linealidad, linealidad en variables y linealidad en parámetros. Nosotros estamos interesados en la linealidad en parámetros: es decir las derivadas con respecto a los coeficientes desconocidos son una función lineal sólo de las  $X$ .

$$\beta_k = \frac{\partial E(Y_i)}{\partial X_{ki}} \quad k = 1, 2, \dots, K$$

El modelo lineal más sencillo es el Modelo de Regresión Lineal Simple donde la variable endógena  $Y$  queda explicada por una única variable exógena  $X$

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad i = 1, 2, \dots, N$$

De igual forma es lineal el Modelo de Regresión Lineal General donde la variable endógena  $Y$  se explica con un conjunto de  $k$ -variables explicativas  $(X_{1i}, X_{2i}, \dots, X_{Ki})$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N$$

Dado que estamos interesados sólo en la linealidad en parámetros también serán considerados lineales los siguientes modelos:

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i \quad \longrightarrow \quad Y_i = \beta_1 + \beta_2 Z_i + u_i \quad \text{con} \quad Z_i = 1/X_i$$

$$Y_i = \beta_1 + \beta_2 X_i^2 + u_i \quad \longrightarrow \quad Y_i = \beta_1 + \beta_2 W_i + u_i \quad \text{con} \quad W_i = X_i^2$$

que son lineales en parámetros según lo dicho anteriormente aunque no lo sean en variables. Ahora bien, existen otras relaciones que aunque en principio no son lineales pueden transformarse en lineales y por tanto son perfectamente estimables en nuestros términos. Por ejemplo:

1. Sea el siguiente modelo:

$$X_i = AB^{Y_i} u_i$$

podemos transformar el modelo en lineal en parámetros tomado logaritmos y obtener:

$$Y_i = \beta_1 + \beta_2 \text{Ln} X_i + u_i \quad (2.8)$$

donde  $\beta_2 = (\text{Ln} B)^{-1}$  y  $\beta_1 = (\frac{\text{Ln} A}{\text{Ln} B})$  a esta transformación se le llama *semilogarítmica*.

2. Sea el modelo:

$$Y_i = AX_i^B u_i \quad \longrightarrow \quad \text{Ln} Y_i = \beta_1 + \beta_2 \text{Ln} X_i + u_i \quad (2.9)$$

donde  $\beta_1 = \text{Ln} A$ , a esta transformación se le llama *doblemente logarítmica*.

**Ejemplo 2.10**

Un ejemplo específico de esta última transformación es la función Cobb-Douglas de la teoría de producción. La función de producción Cobb-Douglas, en su forma estocástica, se expresa como:

$$Q_t = A L_t^{\beta_2} K_t^{\beta_3} e^{u_t}$$

De la ecuación anterior se deduce que la relación entre la producción y los factores capital y trabajo es claramente no lineal. Sin embargo, podemos transformar el modelo tomando logaritmos y obtener la siguiente relación lineal en los parámetros  $\beta_1, \beta_2$  y  $\beta_3$ :

$$Q_t = A L_t^{\beta_2} K_t^{\beta_3} e^{u_t} \longrightarrow \ln Q_t = \beta_1 + \beta_2 \ln L_t + \beta_3 \ln K_t + u_t \quad (2.10)$$

siendo  $\beta_1 = \ln A$ . Una ventaja de este tipo de modelos como el recogido en la ecuación (2.10), en los que **todas** las variables están medidas en logaritmos, es que los parámetros de pendiente además de recibir la interpretación habitual pueden interpretarse en términos de elasticidades:

$$\beta_2 = \frac{\partial E(\ln Q_t)}{\partial \ln L_t} = \frac{\partial E(Q_t)}{\partial L_t} \frac{L_t}{Q_t}$$

$$\beta_3 = \frac{\partial E(\ln Q_t)}{\partial \ln K_t} = \frac{\partial E(Q_t)}{\partial K_t} \frac{K_t}{Q_t}$$

Es decir  $\beta_k$   $k = 2, 3$ , miden el cambio porcentual o elasticidad (parcial) generado en la variable endógena como consecuencia de un cambio porcentual (un 1 %) en la variable exógena correspondiente, ceteris paribus. En el ejemplo anterior  $\beta_2$  y  $\beta_3$  representan las elasticidades de la función de producción con respecto a los factores de producción trabajo y capital respectivamente.

Por otro lado la suma ( $\beta_2 + \beta_3$ ) da información sobre los rendimientos a escala, es decir, la respuesta de la producción a un cambio proporcional en los factores de producción. Si la suma es 1 existen rendimientos constantes a escala, al duplicar los factores de producción se duplica la producción. Si la suma es menor que 1 existen rendimientos decrecientes a escala, al duplicar los factores de producción ésta crece menos del doble. Si la suma es mayor que 1 existen rendimientos crecientes a escala, al duplicar los factores de producción ésta crece más del doble.

Es importante notar que para la ecuación (2.8) la interpretación de los parámetros como elasticidades no es posible ya que al no estar la variable  $Y_i$  en logaritmos:

$$\beta = \frac{\partial E(Y_i)}{\partial \ln X_i} = \frac{\partial E(Y_i)}{\partial X_i} X_i$$

**2.2.1. Forma funcional**

La elección de la forma funcional que recoge la relación existente entre la variable dependiente y las variables explicativas es un aspecto de la especificación de un modelo muy importante en el análisis económico. De hecho, la teoría económica no siempre propone relaciones lineales entre variables de

interés. Es el caso, por ejemplo, de la función de consumo de un bien que aumenta con la renta pero no de forma indefinida ni a ritmo constante sino, en general, a una tasa decreciente, o de las funciones de costes marginales que suelen tener forma de  $U$ , veáse la Figura 5.1.

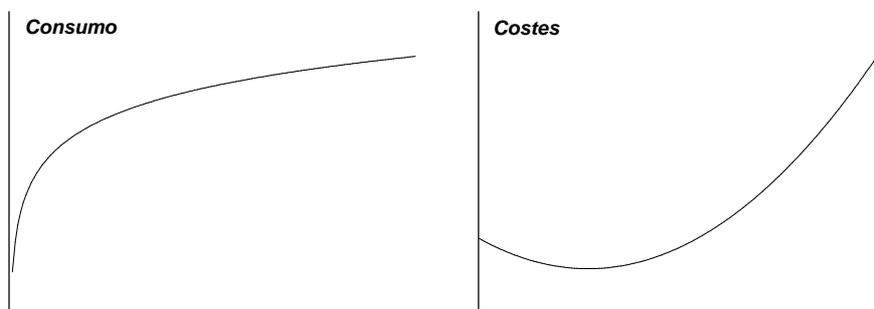


Figura 2.2: Relaciones económicas no lineales

Es necesario hacer énfasis en el hecho de que el supuesto de *linealidad* del modelo de regresión no implica una relación lineal entre las variables sino un modelo en el que los parámetros entran de forma lineal. Por “lineal en los parámetros” se entiende que los parámetros no se multiplican entre sí, no están elevados a potencias, etc. Sin embargo tanto regresando como regresores, sí se pueden transformar para obtener al final un modelo de regresión lineal que satisfaga los supuestos clásicos. Este hecho hace que el modelo de regresión lineal sea bastante flexible y se pueda utilizar para modelar relaciones entre variables económicas no lineales. Así, tanto la función de consumo como la función de costes marginales de la Figura 5.1 se pueden modelizar utilizando formas funcionales sencillas no lineales en las variables. En el caso de la función de consumo, el supuesto de rendimientos decrecientes se puede representar mediante modelos logarítmicos o semilogarítmicos del tipo:

$$\ln C = \alpha + \beta \ln R + u \quad (2.11)$$

$$C = \alpha + \beta \ln R + u \quad (2.12)$$

y las funciones de costes totales se pueden representar mediante funciones polinómicas:

$$CM = \beta_1 + \beta_2 Q + \beta_3 Q^2 + u \quad (2.13)$$

Los modelos (5.1), (5.2) y (5.3) cumplen el supuesto de linealidad porque son lineales en los parámetros y se pueden analizar dentro del marco del MRLG. Ahora bien, como no son modelos lineales en las variables, el efecto marginal del regresor sobre la variable dependiente no va a ser constante. Por ejemplo, en el modelo (5.3), el efecto marginal de un incremento unitario de la producción sobre los costes marginales viene dado por:

$$\frac{\partial E(CT)}{\partial Q} = \beta_2 + 2\beta_3 Q$$

Este resultado implica que la pendiente de la función de costes marginales no es constante sino que es una función lineal de  $Q$  que involucra a los parámetros  $\beta_2$  y  $\beta_3$ .

Otra forma de modelar relaciones no lineales entre las variables explicativas y el regresando es incluir términos de interacción, es decir, el producto de varios regresores del modelo. Consideremos, por ejemplo, el siguiente modelo:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 \times X_3) + u$$

Este modelo es lineal en los parámetros, por lo que cumple el supuesto de linealidad. El efecto marginal de  $X_2$  sobre  $Y$  es:

$$\frac{\partial E(Y)}{\partial X_2} = \beta_2 + \beta_4 X_3$$

de forma que el incremento esperado en  $Y$  ante un incremento unitario en  $X_2$  no es constante sino que depende del valor de  $X_3$ .

Los modelos que no cumplen el supuesto de *linealidad* se pueden clasificar en dos grupos. En el primer grupo se encuentran los modelos que no son lineales en los parámetros pero que se pueden linealizar mediante alguna transformación. En este grupo entra por ejemplo la función de producción Cobb-Douglas que no es lineal ni en las variables ni en los parámetros, pero tomando logaritmos se obtiene una función que no es lineal en las variables pero sí es lineal en los parámetros. El segundo grupo lo forman los modelos que no son lineales en los parámetros y que no se pueden linealizar mediante ninguna transformación, por ejemplo,

$$Y = \beta_1 + X_1^{\beta_2\beta_3} + X_2^{\beta_2} + u$$

Este tipo de modelos se estima por mínimos cuadrados no lineales.

### 2.3. Utilización de variables explicativas cualitativas

A lo largo del curso se han especificado mayoritariamente modelos con variables de naturaleza cuantitativa, es decir, aquéllas que toman valores numéricos. Sin embargo, las variables también pueden ser cualitativas, es decir, pueden tomar valores no numéricos como categorías, clases o atributos. Por ejemplo, son variables cualitativas el género de las personas, el estado civil, la raza, el pertenecer a diferentes zonas geográficas, momentos históricos, estaciones del año, etc. De esta forma, el salario de los trabajadores puede depender del género de los mismos; la tasa de criminalidad puede venir determinada por la zona geográfica de residencia de los individuos; el PIB de los países puede estar influenciado por determinados acontecimientos históricos como las guerras; las ventas de un determinado producto pueden ser significativamente distintas en función de la época del año, etc. En esta sección, aunque seguimos manteniendo que la variable dependiente es cuantitativa, vamos a considerar que ésta puede venir explicada por variables cualitativas y/o cuantitativas y veremos como trabajar con ellas incluyéndolas como regresores en el MRLG.

Dado que las categorías de las variables no son directamente cuantificables, las vamos a cuantificar construyendo unas variables artificiales llamadas ficticias, binarias o dummies, que son numéricas. Estas variables toman arbitrariamente el valor 1 si la categoría está presente en el individuo y 0 en caso contrario<sup>5</sup>.

<sup>5</sup>Las variables ficticias pueden tomar dos valores cualesquiera, sin embargo, la interpretación de los coeficientes es más sencilla si se consideran los valores 0 y 1.

$$D_i = \begin{cases} 1 & \text{si la categoría está presente} \\ 0 & \text{en caso contrario} \end{cases}$$

Por ejemplo si queremos estudiar la dependencia del salario ( $W_i$ ) con respecto al sexo del individuo definiremos dos variables ficticias:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases}$$

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

la variable sexo tiene dos categorías o estados de la naturaleza: hombre y mujer, para recogerlos utilizamos dos variables ficticias que dividen la muestra en dos clases hombres y mujeres, y asignamos un valor arbitrario a cada clase<sup>6</sup>.

En este tema ya hemos trabajado con ellas, el Ejemplo 2.6 especificamos la función de salario en función del regresor cualitativo sexo e interpretamos sus parámetros. En el Ejemplo 2.7 además se añadió un regresor cuantitativo, la experiencia y se interpretaron los parámetros. Si se retoman dichos ejercicios se puede ver que trabajar con variables cualitativas o con variables cuantitativas a la hora de interpretar los coeficientes de la regresión y estimarlos es indiferente sin embargo hay que tener en cuenta algunas reglas a la hora de especificar el modelo. A conocer éstas vamos a dedicar las secciones siguientes.

### 2.3.1. Modelo que recoge sólo efectos cualitativos: comparando medias. Sólo un conjunto de variables ficticias.

Supongamos que tenemos datos de salarios de hombres y mujeres,  $W_i$  y creemos que, en media, existen diferencias salariales entre estos dos grupos. Para contrastar que esto es cierto podemos recoger el efecto cualitativo sexo sobre el salario utilizando las variables ficticias:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases} \quad S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

y podemos especificar el siguiente modelo como ya se hizo en el Ejemplo 2.6:

$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M \quad u_i \sim NID(0, \sigma^2) \quad (2.14)$$

---

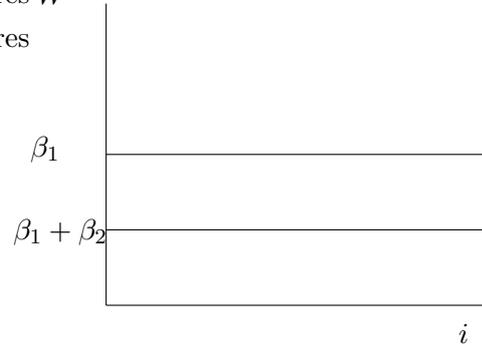
<sup>6</sup>Elegir los valores (0,1) es muy cómodo pero podríamos elegir otros, por ejemplo:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases} \quad S_{2i} = \begin{cases} 2 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

Hay que notar que el modelo (2.14) da lugar a dos ecuaciones:

$$\begin{aligned} W_i &= \beta_1 + u_i & i = 1, \dots, N_H & \quad \text{para los hombres } W \\ W_i &= \beta_1 + \beta_2 + u_i & i = 1, \dots, N_M & \quad \text{para las mujeres} \end{aligned}$$

$\beta_1$  es el salario esperado cuando el individuo es hombre,  $\beta_1 + \beta_2$  es el salario esperado de una mujer y  $\beta_2$  recoge el efecto diferencial en el salario esperado entre hombres y mujeres. Si no existiera discriminación salarial por sexo, es decir si hombres y mujeres tuvieran el mismo salario, su valor sería cero. En el gráfico podemos observar estos efectos donde se supone que  $\beta_2$  es negativo por razones didácticas.



- Alternativa de especificación del modelo (2.14):

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M \quad (2.15)$$

de donde suponiendo  $u_i \sim NID(0, \sigma^2)$

$\alpha_1 = E(W_i | S_{1i} = 1; S_{2i} = 0)$  es el salario esperado de un hombre

$\alpha_2 = E(W_i | S_{1i} = 0; S_{2i} = 1)$  es el salario esperado de una mujer

por tanto estos coeficientes recogen el salario medio dentro del grupo.

En este caso el modelo (2.15) da lugar a dos ecuaciones:

$$\begin{aligned} W_i &= \alpha_1 + u_i & i = 1, \dots, N_H & \quad \text{para los hombres} \\ W_i &= \alpha_2 + u_i & i = 1, \dots, N_M & \quad \text{para las mujeres} \end{aligned}$$

La relación entre los parámetros del modelo (2.14) y los del modelo (2.15) es la siguiente:

$$\beta_1 = \alpha_1 \quad \beta_1 + \beta_2 = \alpha_2 \quad \text{luego} \quad \beta_2 = \alpha_2 - \alpha_1$$

### Ejercicio 2.1

Interpreta los coeficientes de la siguiente regresión:

$$W_i = \beta_1 S_{1i} + \beta_2 + u_i \quad i = 1, \dots, N_H + N_M \quad u_i \sim NID(0, \sigma^2)$$

donde  $W_i$  es el salario del individuo  $i$  y

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases} \quad S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

¿Qué diferencia hay entre ésta especificación y la especificación del modelo (2.14)?

### 2.3.2. Dos o más conjuntos de variables ficticias

Supongamos que pensamos que en el nivel de salarios influye además del sexo el nivel de educación. Para recoger estos efectos podemos definir dos conjuntos de variables ficticias, sexo y educación, la primera con dos categorías o estados de la naturaleza y la segunda con tres, y recoger cada categoría o estado de la naturaleza con un variable ficticia. Así, definimos:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases} \quad E_{1i} = \begin{cases} 1 & \text{si } i \text{ tiene hasta estudios primarios} \\ 0 & \text{en caso contrario} \end{cases}$$

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases} \quad E_{2i} = \begin{cases} 1 & \text{si } i \text{ tiene hasta estudios secundarios} \\ 0 & \text{en caso contrario} \end{cases}$$

$$E_{3i} = \begin{cases} 1 & \text{si } i \text{ tiene hasta estudios universitarios} \\ 0 & \text{en caso contrario} \end{cases}$$

siendo  $E_{ij}$  sucesos excluyentes. La especificación correspondiente es:

$$W_i = \mu + \alpha_2 S_{2i} + \beta_2 E_{2i} + \beta_3 E_{3i} + u_i \quad i = 1, \dots, N_H + N_M \quad (2.16)$$

donde hemos excluido una categoría de cada factor cualitativo. Podemos obtener el salario esperado de los diferentes individuos de la muestra:

$$E(W_i/S_{2i} = E_{2i} = E_{3i} = 0) = \mu, \text{ salario esperado de un hombre con estudios primarios.}$$

$$E(W_i/E_{2i} = 1; S_{2i} = E_{3i} = 0) = \mu + \beta_2, \text{ salario esperado de un hombre con estudios secundarios.}$$

$$E(W_i/E_{3i} = 1; S_{2i} = E_{2i} = 0) = \mu + \beta_3, \text{ salario esperado de un hombre con estudios universitarios.}$$

$$E(W_i/S_{2i} = 1; E_{2i} = E_{3i} = 0) = \mu + \alpha_2, \text{ salario esperado de una mujer con estudios primarios}$$

$$E(W_i/S_{2i} = E_{2i} = 1; E_{3i} = 0) = \mu + \alpha_2 + \beta_2, \text{ salario esperado de una mujer con estudios secundarios.}$$

$$E(W_i/S_{2i} = E_{3i} = 1; E_{2i} = 0) = \mu + \alpha_2 + \beta_3, \text{ salario esperado de una mujer con estudios universitarios.}$$

Esta información podemos resumirla en la siguiente tabla:

$E(W_i)$	$E_{1i}$	$E_{2i}$	$E_{3i}$
$S_{1i}$	$\mu$	$\mu + \beta_2$	$\mu + \beta_3$
$S_{2i}$	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_3$

y podemos interpretar los parámetros como sigue:

- $\mu$  Base de comparación.
- $\alpha_2$  Efecto diferencial en el salario medio debido al factor sexo. Por tanto es el diferencial en el salario medio entre hombres y mujeres independientemente de su nivel de educación.
- $\beta_2$  Efecto diferencial en el salario medio debido a tener un nivel de estudios secundarios. Por tanto es el diferencial en el salario medio, para hombres y mujeres, entre tener un nivel de estudios primarios y tener secundaria.
- $\beta_3$  Efecto diferencial en el salario medio debido a tener un nivel de estudios universitarios. Por tanto es el diferencial en el salario medio, para hombres y mujeres, entre tener un nivel de estudios primarios y tener estudios universitarios.

La matriz de regresores del modelo sería:

$$X = \begin{bmatrix} i_{N_1} & 0 & 0 & 0 \\ i_{N_2} & 0 & i_{N_2} & 0 \\ i_{N_3} & 0 & 0 & i_{N_3} \\ i_{N_4} & i_{N_4} & 0 & 0 \\ i_{N_5} & i_{N_5} & i_{N_5} & 0 \\ i_{N_6} & i_{N_6} & 0 & i_{N_6} \end{bmatrix}$$

donde  $i_{N_j}$  es un vector de unos de tamaño el número de individuos que cumplen las condiciones, por ejemplo  $i_{N_6}$  es un vector de unos de tamaño el número de mujeres con estudios universitarios. Cuando existen dos o más conjuntos de variables ficticias lo que no debemos hacer es incluir todas las variables ficticias y un término independiente. En el caso anterior tenemos dos conjuntos con dos y tres estados de la naturaleza respectivamente, si proponemos la especificación:

$$W_i = \mu^* + \alpha_1^* S_{1i} + \alpha_2^* S_{2i} + \beta_1^* E_{1i} + \beta_2^* E_{2i} + \beta_3^* E_{3i} + u_i \quad i = 1, \dots, N_H + N_M \quad (2.17)$$

el determinante  $|X'X| = 0$ , no se cumplirían todas las hipótesis básicas y no podríamos estimar separadamente ninguno de los coeficientes. La matriz de regresores del modelo (2.17) es:

$$X = \begin{bmatrix} i_{N_1} & i_{N_1} & 0 & i_{N_1} & 0 & 0 \\ i_{N_2} & i_{N_2} & 0 & 0 & i_{N_2} & 0 \\ i_{N_3} & i_{N_3} & 0 & 0 & 0 & i_{N_3} \\ i_{N_4} & 0 & i_{N_4} & i_{N_4} & 0 & 0 \\ i_{N_5} & 0 & i_{N_5} & 0 & i_{N_5} & 0 \\ i_{N_6} & 0 & i_{N_6} & 0 & 0 & i_{N_6} \end{bmatrix} \Rightarrow rg(X) < K$$

### 2.3.3. Inclusión de variables cuantitativas

En cualquiera de los modelos anteriores puede incluirse una-s variable-s cuantitativas, por ejemplo si creemos que el salario depende no solo de sexo sino también del número de horas trabajadas, variable que denotamos como  $X_i$  propondremos:

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + \beta X_i + u_i \quad i = 1, \dots, N_H + N_M \quad (2.18)$$

Donde el coeficiente  $\beta$  se interpreta de la forma habitual,  $\beta = \frac{\partial E(W_i)}{\partial X_i}$ . En forma matricial el modelo sería:

$$\begin{bmatrix} W_H \\ W_M \end{bmatrix} = \begin{bmatrix} i_H & 0 & X_H \\ 0 & i_M & X_M \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \begin{bmatrix} u_H \\ u_M \end{bmatrix} \Rightarrow Y = X\beta + u$$

La especificación alternativa correspondiente sería:

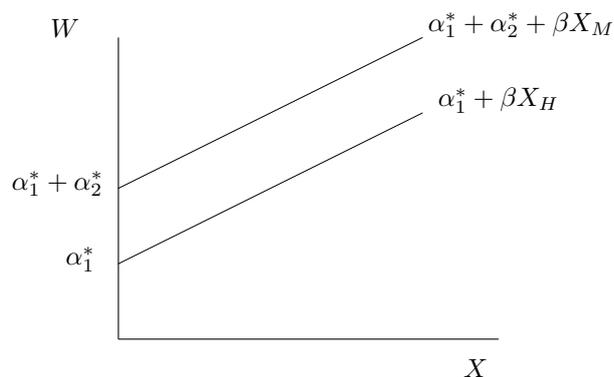
$$W_i = \alpha_1^* + \alpha_2^* S_{2i} + \beta X_i + u_i \quad (2.19)$$

$$i = 1, \dots, N_H + N_M$$

Donde el coeficiente  $\beta$  se interpreta de la forma habitual. En forma matricial el modelo sería:

$$\begin{bmatrix} W_H \\ W_M \end{bmatrix} = \begin{bmatrix} i_H & 0 & X_H \\ i_M & i_M & X_M \end{bmatrix} \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \\ \beta \end{bmatrix} + \begin{bmatrix} u_H \\ u_M \end{bmatrix}$$

$$\Rightarrow Y = X\beta + u$$



### 2.3.4. Comportamiento estacional

Las variables ficticias permiten recoger fácilmente comportamientos estacionales, como se hizo en el Ejemplo 2.8. Por ejemplo, que las ventas de una empresa sean sistemáticamente superiores en alguno de los trimestres del año y que ese comportamiento se repita sistemáticamente año tras año es un clásico patrón de comportamiento sistemático estacional. Este comportamiento se produce en datos de series temporales de periodo inferior al anual y puede ser estudiado fácilmente mediante variables ficticias.

Por ejemplo para recoger el comportamiento estacional de una variable  $Y_t$  muestreada trimestralmente podemos proponer el modelo:

$$Y_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \quad t = 1, 2, \dots, T$$

donde  $t$  es el tiempo y las variables  $D_{jt}$  son variables ficticias estacionales que se definen:

$$D_{jt} = \begin{cases} 1 & \text{si la observación } t \text{ pertenece al trimestre } j \quad j = 2, 3, 4 \\ 0 & \text{en caso contrario} \end{cases}$$

La especificación alternativa sería:

$$Y_t = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \quad t = 1, 2, \dots, T$$

### 2.3.5. Efectos de interacción

**Entre factores cualitativos y cuantitativos** En las ecuaciones (2.18) y (2.19) se recogen cambios en ordenada pero no en pendiente, sin embargo podemos pensar que el número de horas trabajadas cambia según el sexo del individuo con lo cual debemos recoger cambios en pendiente. Este efecto podemos analizarlo asociando las variables ficticias a la variable cuantitativa. Así proponemos el siguiente modelo:

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + \beta_1 (S_{1i} \times X_i) + \beta_2 (S_{2i} \times X_i) + u_i \quad i = 1, \dots, N_H + N_M \quad (2.20)$$

$$E(W_i/S_{1i} = 1; S_{2i} = 0) = \alpha_1 + \beta_1 X_i$$

$$E(W_i/S_{1i} = 0; S_{2i} = 1) = \alpha_2 + \beta_2 X_i$$

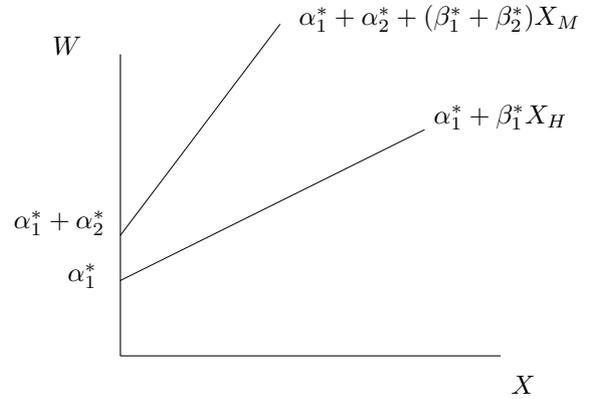
donde  $\beta_1$  y  $\beta_2$  recogen el incremento en el salario medio ante un aumento unitario en las horas trabajadas, para los hombres y para las mujeres respectivamente.

Una especificación alternativa sería:

$$W_i = \alpha_1^* + \alpha_2^* S_{2i} + \beta_1^* X_i + \beta_2^* (S_{2i} \times X_i) + u_i$$

$$i = 1, \dots, N_H + N_M \quad (2.21)$$

siendo  $\alpha_2^*$  el incremento salarial en media por el hecho de ser mujer y  $\beta_2^*$  el incremento en el salario medio de una mujer con respecto a un hombre ante un aumento de una hora en el número de horas trabajado.



**Entre factores cualitativos** En el modelo (2.16) se supone que el efecto de cada factor es constante para todos los niveles de los demás factores. Sin embargo si suponemos que el efecto diferencial del sexo variase con el nivel de educación existiría un efecto interacción entre las variables ficticias sexo y educación, que podemos recoger así:

$$W_i = \mu + \alpha_2 S_{2i} + \beta_2 E_{2i} + \beta_3 E_{3i} + \gamma_2 (S_{2i} \times E_{2i}) + \gamma_3 (S_{2i} \times E_{3i}) + u_i \quad i = 1, \dots, N_H + N_M \quad (2.22)$$

donde la tabla que resume el comportamiento de la recta de regresión poblacional sería:

$E(W_i)$	$E_{1i}$	$E_{2i}$	$E_{3i}$
$S_{1i}$	$\mu$	$\mu + \beta_2$	$\mu + \beta_3$
$S_{2i}$	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2 + \gamma_2$	$\mu + \alpha_2 + \beta_3 + \gamma_3$

y podemos interpretar los parámetros como sigue:

- $\mu$  base de comparación.
- $\beta_2$  Efecto diferencial en el salario medio debido a tener un nivel de estudios secundarios, con respecto a tener estudios primarios, para los hombres.
- $\beta_3$  Efecto diferencial en el salario medio debido a tener un nivel de estudios universitarios, con respecto a tener estudios primarios, para los hombres.
- $\alpha_2$  Efecto diferencial en el salario medio entre los hombres y las mujeres para un nivel de educación primaria.
- $\alpha_2 + \gamma_2$  Efecto diferencial en el salario medio, entre hombres y mujeres, para un nivel de educación secundaria.
- $\alpha_2 + \gamma_3$  Efecto diferencial en el salario medio, entre hombres y mujeres, para un nivel de educación universitaria.
- $\beta_2 + \gamma_2$  Efecto diferencial en el salario medio debido a tener un nivel de estudios secundarios, con respecto a tener estudios primarios, para las mujeres.
- $\beta_3 + \gamma_3$  Efecto diferencial en el salario medio debido a tener un nivel de estudios universitarios, con respecto a tener estudios primarios, para las mujeres.

## 2.4. Bibliografía del tema

### Referencias bibliográficas básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con *gretl*:

- [1] Ramanathan, R. (2002), *Instructor's Manual to accompany*, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), *Student Solutions Manual*, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

### Referencias Bibliográficas Complementarias:

- [1] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Análisis de regresión con gretl*. Open Course Ware. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Courseisting>).
- [2] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación on-line de la Facultad de C.C. Económicas y Empresariales.
- [3] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones.
- [4] Esteban, MV (2008). *Estadística Actuarial: Regresión Lineal*, Sarriko On Line 3/08. Publicación on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.
- [5] Esteban, M.V. (2007). *Colección de ejercicios y exámenes*. Material docente. Servicio de Publicaciones.
- [6] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). *Ejercicios de Econometría*. Editorial McGraw-Hill.
- [7] Greene, W. (1998), *Análisis Econométrico*, Ed. Prentice Hall, 3ª edición.
- [8] Ramanathan, R. (2002), *Introductory Econometrics with applications*, Ed. South-Western, 5th. edition.
- [9] Verbeek, M. (2004). *A Guide to Modern Econometrics*. Wiley.



## Tema 3

# Modelo de Regresión Lineal General: estimación

En este tema nos ocuparemos de estimar el Modelo de Regresión Lineal General. El método de estimación que desarrollaremos son los Mínimos Cuadrados Ordinarios, MCO, que bajo ciertas hipótesis de comportamiento sobre los distintos elementos del modelo nos proporcionará estimadores con buenas propiedades, lineales, insesgados y de mínima varianza. Propiedades que buscaremos no solo para el estimador mínimo cuadrático sino también para otros estimadores alternativos.

Para finalizar el tema veremos como realizar análisis de regresión mediante el software *gretl*.

### Competencias a trabajar en estas sesiones:

- C2. Aplicar la metodología econométrica básica para estimar y validar relaciones económicas en base a la información estadística disponible sobre las variables y utilizando los instrumentos informáticos apropiados.
- C3. Interpretar razonadamente los resultados obtenidos en la estimación y validación del modelo econométrico con el objetivo de elaborar informes económicos.
- C4. Presentar de forma clara y concisa, tanto oralmente como por escrito, las conclusiones obtenidas en una aplicación empírica.

### Al final de este tema deberíais ser capaces de:

- 1. Aplicar el estimador de Mínimos Cuadrados Ordinarios, MCO (C2).
- 2. Distinguir entre la perturbación y el residuo u error de estimación. Conocer las distribuciones respectivas (C2).
- 3. Conocer y saber demostrar las propiedades del estimador de MCO (C2 y C3).
- 4. Organizar y sistematizar información estadística relevante (C3).

5. Utilizar un software econométrico (Gretl) para el análisis de bases de datos económicos e interpretar sus resultados (C2 , C3 y C4).

**Bibliografía Recomendada:**

Al final del tema tenéis recogida la bibliografía correspondiente. En particular se os recomienda leer los capítulos correspondientes a la bibliografía básica detallados a continuación:

- Gujarati, D. y Porter, D.C. (2010). En Parte I: Caps. 1, 2, 3, 4, 5, 7, 8 y 9. Parte II: Cap. 10.
- Ramanathan, R. (2002). Caps. 3, 4, 5 y 7.
- Wooldridge, J.M. (2006). Caps. 2, 3, 4 y 7

### 3.1. Estimación por Mínimos Cuadrados Ordinarios (MCO)

- Nuestro **objetivo** es estimar los parámetros desconocidos  $\beta_k$ ,  $k = 1, \dots, K$  de

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N$$

$$Y = X\beta + u \quad \text{en forma matricial.}$$

A los parámetros estimados los denotamos  $\hat{\beta}_k$  y la estimación del modelo es

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_K X_{Ki} \quad i = 1, 2, \dots, N$$

$$\hat{Y} = X\hat{\beta} \quad \text{en forma matricial,}$$

a la cual denominamos **Función de Regresión Muestral** (FRM).

- **Elementos adicionales**

- La perturbación del modelo recoge todo aquello que no ha sido explicado por la parte sistemática del modelo y se obtiene como la diferencia entre la variable a explicar y la recta de regresión poblacional:

$$u_i = Y_i - E(Y_i) \quad i = 1, 2, \dots, N$$

$$u = Y - X\beta \quad \text{en forma matricial.}$$

- El residuo mide el error cometido al estimar la variable endógena y se define como la diferencia entre la variable a explicar y la recta de regresión muestral<sup>1</sup>:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki} \quad i = 1, 2, \dots, N$$

$$\hat{u} = Y - \hat{Y} = Y - X\hat{\beta} \quad \text{en forma matricial.}$$

Este error proviene de dos fuentes: la primera, por el hecho de no poder obtener los valores de la perturbación ( $u_i$ ) y la segunda se debe a que la estimación de los coeficientes desconocidos  $\beta_k$   $k = 1, \dots, K$  introduce un error adicional. Es importante, por tanto, diferenciar y no confundir el residuo con la perturbación.

- **Representación gráfica:** Cuando  $K = 2$  el modelo se escribe

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

y se denomina Modelo de Regresión Lineal Simple (MRLS). Entonces, la relación entre la FRM, FRP, residuos y perturbaciones puede visualizarse en la Figura 3.1.

En el Gráfico 3.1 la función de regresión poblacional está trazada en color negro así como los coeficientes poblacionales, la ordenada ( $\beta_1$ ) y la pendiente ( $\beta_2$ ). Podemos ver que el valor  $Y_i$  se

<sup>1</sup>Los residuos son a la FRM lo que las perturbaciones a la FRP. Sin embargo, no son buenos estimadores de las mismas (al menos, no si la muestra es pequeña) porque no tienen las mismas propiedades, se puede ver en el Anexo 3 incluido en el Apéndice del Tema.

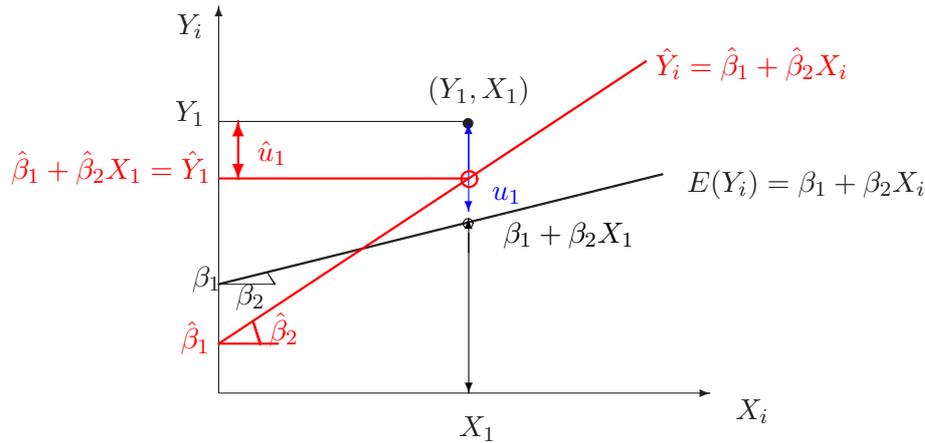


Figura 3.1: Función de regresión poblacional y función de regresión muestral

obtiene como la suma del valor que toma la parte sistemática  $\beta_1 + \beta_2 X_i$  (situada sobre la FRP) y del valor que toma la perturbación  $u_i$ , esto es,  $Y_i = \beta_1 + \beta_2 X_i + u_i$ .

La función de regresión muestral y los coeficientes estimados ( $\hat{\beta}_1$  y  $\hat{\beta}_2$ ) están representados en color rojo. La diferencia entre la FRP y la FRM se debe a los errores que se cometen en la estimación de los coeficientes de la regresión ( $\hat{\beta}_1 \neq \beta_1, \hat{\beta}_2 \neq \beta_2$ ). Basándonos en la FRM podemos obtener el valor del punto  $Y_i$  como la suma del valor estimado de la parte sistemática  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$  (situado sobre la FRM) y del valor que toma el residuo  $\hat{u}_i$ , esto es,  $Y_i = \hat{Y}_i + \hat{u}_i$ .

### 3.1.1. Método de estimación de Mínimos Cuadrados Ordinarios (MCO)

El objetivo es obtener un vector de **estimadores** de los parámetros ó coeficientes  $\beta_1, \dots, \beta_K$ , para el cual buscamos  $K$  estimadores respectivamente. Un estimador es una fórmula que, aplicada a los datos, da como resultado un valor numérico. Al valor numérico calculado a partir de un estimador y de una muestra dada, lo denominamos **estimación**. A ambos los denotamos  $\hat{\beta}$ , el cual puede ser, consecuentemente, un vector de estimadores o de estimaciones, según nos estemos refiriendo a la fórmula o a los valores numéricos resultantes. Pero no son lo mismo.

Necesitamos un criterio de estimación, que dados el modelo y la muestra nos permita obtener la función de regresión muestral. Proponemos el criterio de estimación mínimo cuadrático ordinario por su sencillez y buenas propiedades, que demostraremos más adelante.

El criterio de estimación Mínimo Cuadrático Ordinario (MCO) se expresa:

$$\min_{\hat{\beta}} \sum_{i=1}^N \hat{u}_i^2 = \min_{\hat{\beta}} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

A continuación buscaremos la expresión de los estimadores mínimo cuadráticos obtenidos como resultado de aplicar el criterio MCO al modelo de regresión lineal general.

• **Estimador MCO del MRLG**

Criterio:

$$\begin{aligned} \min_{\hat{\beta}_1, \dots, \hat{\beta}_K} \sum_{i=1}^N \hat{u}_i^2 &= \min_{\hat{\beta}_1, \dots, \hat{\beta}_K} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \\ \min_{\hat{\beta}_1, \dots, \hat{\beta}_K} \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki})^2 & \end{aligned} \quad (3.1)$$

Las  $K$  Condiciones de Primer Orden (C.P.O.) de mínimo son

$$\begin{aligned} \frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\beta}_1} &= 0 \\ \frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\beta}_2} &= 0 \\ \frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\beta}_3} &= 0 \\ &\vdots \\ \frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\beta}_K} &= 0 \end{aligned}$$

de donde se obtienen las ecuaciones normales:

$$\begin{aligned} -2 \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki}) &= 0 \\ -2 \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki}) X_{2i} &= 0 \\ &\vdots \\ -2 \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki}) X_{Ki} &= 0 \end{aligned}$$

que pueden escribirse como:

$$\begin{aligned} \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \dots + \hat{\beta}_K \sum X_{Ki} \\ \sum X_{2i} Y_i &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \dots + \hat{\beta}_K \sum X_{2i} X_{Ki} \\ &\vdots \\ \sum X_{Ki} Y_i &= \hat{\beta}_1 \sum X_{Ki} + \hat{\beta}_2 \sum X_{Ki} X_{2i} + \dots + \hat{\beta}_K \sum X_{Ki}^2 \end{aligned}$$

En forma matricial,  $\sum_{i=1}^N \hat{u}_i^2 = \hat{u}'\hat{u}$  donde  $\hat{u}$  es un vector  $N \times 1$  y el criterio puede escribirse  $(1 \times 1)$

$$\min_{\hat{\beta}} \hat{u}'\hat{u} = \min_{\hat{\beta}} (Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

Las  $K$  Condiciones de Primer Orden (C.P.O.) de mínimo son

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = 0 \Rightarrow -2X'(Y - X\hat{\beta}) = 0.$$

Despejando, obtenemos las **ecuaciones normales** en forma matricial:

$$X'Y = X'X\hat{\beta}_{MCO}. \quad (3.2)$$

de donde el **estimador MCO** (en forma matricial) es:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y \quad (3.3)$$

en el que  $X'X$  es una matriz de orden  $K \times K$ ,  $X'Y$  un vector de orden  $K \times 1$  y  $\hat{\beta}$  un vector de orden  $K \times 1$ , tales que

$$X'X = \begin{matrix} (K \times K) \\ \left[ \begin{array}{cccccc} N & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{Ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & \cdots & \sum X_{2i}X_{Ki} \\ \sum X_{3i} & \sum X_{3i}X_{2i} & \sum X_{3i}^2 & \cdots & \sum X_{3i}X_{Ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{Ki} & \sum X_{Ki}X_{2i} & \sum X_{Ki}X_{3i} & \cdots & \sum X_{Ki}^2 \end{array} \right] \end{matrix}$$

$$X'Y = \begin{matrix} (K \times 1) \\ \left[ \begin{array}{c} \sum Y_i \\ \sum X_{2i}Y_i \\ \sum X_{3i}Y_i \\ \vdots \\ \sum X_{Ki}Y_i \end{array} \right] \end{matrix} \quad \hat{\beta} = \begin{matrix} (K \times 1) \\ \left[ \begin{array}{c} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_K \end{array} \right] \end{matrix}.$$

El estimador MCO cumple también las condiciones de segundo orden de mínimo, con lo cual es, efectivamente, la solución al problema de minimización de la suma de los residuos al cuadrado.

### Ejercicio 3.1

Sea el modelo de regresión lineal simple donde se regresa  $Y_t$  sobre  $X_t$ , incluyendo un término independiente.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1, \dots, T$$

Sin utilizar notación matricial:

1. Escribe el sistema de ecuaciones correspondiente al modelo propuesto.
2. Escribe la función objetivo correspondiente a la estimación por MCO de los parámetros desconocidos. Deriva las condiciones de primer orden.

3. Obtén las ecuaciones normales correspondientes al modelo.
4. Obtén la expresión de  $\hat{\beta}_1$  y  $\hat{\beta}_2$ .

Utilizando notación matricial:

1. Escribe la expresión matricial del modelo.
2. Escribe la función objetivo correspondiente a la estimación por MCO de los parámetros desconocidos. Deriva las condiciones de primer orden.
3. Obtén las ecuaciones normales correspondientes al modelo.
4. Obtén la expresión del estimador del vector de parámetros desconocidos  $\hat{\beta}$ .

### Ejercicio 3.2

Sea el siguiente modelo de regresión lineal simple donde se regresa  $Y_t$  sobre  $X_t$ .

$$Y_t = \beta X_t + u_t \quad t = 1, \dots, T$$

Sin utilizar notación matricial:

1. Escribe el sistema de ecuaciones correspondiente al modelo propuesto.
2. Escribe la función objetivo correspondiente a la estimación por MCO del parámetro desconocido. Deriva la condición de primer orden.
3. Obtén la ecuación normal del modelo.
4. Obtén la expresión de  $\hat{\beta}$ .

Utilizando matrices escribe la expresión matricial del modelo y obtén la expresión de  $\hat{\beta}$ .

### Algunas equivalencias de notación

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N \quad \Leftrightarrow \quad Y = X\beta + u$$

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} \quad i = 1, 2, \dots, N \quad \Leftrightarrow \quad E(Y) = X\beta$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_K X_{Ki} \quad i = 1, 2, \dots, N \quad \Leftrightarrow \quad \hat{Y} = X\hat{\beta}$$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_K X_{Ki} + \hat{u}_i \quad i = 1, 2, \dots, N \quad \Leftrightarrow \quad Y = X\hat{\beta} + \hat{u}$$

$$\hat{u}_i = Y_i - \hat{Y}_i \quad i = 1, 2, \dots, N \quad \Leftrightarrow \quad \hat{u} = Y - \hat{Y}$$

### Interpretación de los coeficientes estimados por MCO

- $\hat{\beta}_1 = \widehat{E}(Y_i | X_{ki} = 0, \forall k = 2, \dots, K)$ . Valor esperado *estimado* de  $Y_i$  cuando las variables explicativas son todas cero.
- $\hat{\beta}_k = \frac{\partial \widehat{E}(Y_i)}{\partial X_{ki}} = \frac{\Delta \widehat{E}(Y_i)}{\Delta X_{ki}} \quad \forall k = 2, \dots, K$ . Incremento esperado *estimado* (ó *decremento esperado estimado*) en  $Y_i$  cuando la variable  $X_k$  se incrementa en una unidad, manteniéndose constantes el resto de las variables explicativas.

### Ejemplo 3.1

Vamos a retomar ahora el Ejemplo 2.2 donde se analizaban los determinantes de las ventas de una cadena de comida rápida. Se dispone de una base de datos para 75 ciudades en las que está enclavadas una cadena de comida rápida sobre sus ventas, precio y gasto en publicidad. Son datos de sección cruzada y la descripción de las variables disponibles es<sup>2</sup>:

S = Ingresos mensuales por ventas en miles de dólares (Rango 62,400 - 91,200)

P = Índice de precios de todos los productos vendidos en un mes (Rango 4,83 - 6,49)

A = Gasto en publicidad (Rango 0,5 - 3,1)

Para analizar si el precio y el gasto en publicidad son factores que explican o no el ingreso por ventas se especifica el siguiente modelo:

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + u_i \quad i = 1, \dots, 75 \quad (3.4)$$

Para estimar el modelo se utilizan las observaciones disponibles en el fichero *andy.gdt* y que son las siguientes<sup>3</sup>:

Las estimaciones obtenidas resultan de aplicar el criterio MCO  $\hat{\beta} = (X'X)^{-1}X'Y$ :

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 75 & \sum P_i & \sum A_i \\ \sum P_i & \sum P_i^2 & \sum P_i A_i \\ \sum A_i & \sum P_i A_i & \sum A_i^2 \end{bmatrix}^{-1} \times \begin{bmatrix} \sum S_i \\ \sum S_i P_i \\ \sum S_i A_i \end{bmatrix} =$$

$$\begin{bmatrix} 75,0000 & 426,5400 & 138,3000 \\ 426,5400 & 2445,7074 & 787,3810 \\ 138,3000 & 787,3810 & 306,2100 \end{bmatrix}^{-1} \begin{bmatrix} 5803,1000 \\ 32847,6770 \\ 10789,6000 \end{bmatrix} = \begin{bmatrix} 118,914 \\ -7,90785 \\ 1,86258 \end{bmatrix}$$

- **La función de regresión muestral obtenida es:**

$$\widehat{S}_i = 118,914 - 7,90785P_i + 1,86258A_i$$

<sup>2</sup>Fuente: Ramanathan, Ramu (2002) *Introductory econometrics with applications*. Carpeta PoE, conjunto de datos *andy.gdt*

<sup>3</sup>Puedes acceder a estos datos ejecutando *gretl* → *En Archivo* → *Abrir datos* → *Archivo de muestra* → *Elige PoE*, el fichero *andy.gdt*.

• **Interpretación de los signos obtenidos:**

Para la variable  $P$  el signo es negativo lo que indica que la demanda es elástica. Estimamos que permaneciendo el gasto en publicidad constante un incremento de un dólar en el precio lleva a una caída en los ingresos mensuales de 7908\$. O lo que es lo mismo una reducción de un dólar en el precio se estima que produce un incremento de las ventas de 7908\$. En este caso una estrategia de reducción de precios a través de ofertas especiales sería exitosa en incrementar los ingresos por ventas.

Sin embargo la magnitud del cambio en precios es muy importante. Un cambio de 1\$ en el precio es relativamente un cambio grande. La media muestral del precio es 5,99 y su desviación típica es 0,52. Un cambio en precio de un 10% es más realista y en este caso el cambio estimado en los ingresos por ventas es de 791\$.

El signo del coeficiente estimado para el gasto en publicidad es positivo. Estimamos que manteniéndose el precio constante, un incremento en el gasto en publicidad de 1000\$ lleva a un incremento en los ingresos por ventas de 1863\$. Esta información puede ser utilizada para analizar si un incremento en el gasto en publicidad incrementa el beneficio teniendo en cuenta el coste de producir una hamburguesa más.

El término independiente implica que cuando ambos precio y gasto en publicidad es cero los ingresos por ventas son 118,914\$. Esto no es posible, a precio cero ingresos por ventas cero. En muchos casos el término independiente no es interpretable.

Estadísticos principales, usando las observaciones 1 - 75

Variable	Media	Mediana	Mínimo	Máximo
sales	77,3747	76,5000	62,4000	91,2000
price	5,68720	5,69000	4,83000	6,49000
advert	1,84400	1,80000	0,500000	3,10000
Variable	Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
sales	6,48854	0,0838587	-0,0106308	-0,744672
price	0,518432	0,0911577	0,0618457	-1,33284
advert	0,831677	0,451018	0,0370873	-1,29511

En el tema siguiente veremos cómo realizar contraste de hipótesis y en el Tema 4 veremos cómo hacer predicción. Sin embargo es fácil ver que para un precio de  $P_i = 5,5$  y un gasto en publicidad de  $A_i = 1,2$  el valor predicho de las ventas es:

$$\hat{S}_i = 118,91 - 7,908P_i + 1,863A_i = 118,91 - 7,9079 \times 5,5 + 1,863 \times 1,2 = 77,656\$$$

**Ejemplo 3.2**

Supongamos que se dispone de datos para estimar la relación en Estados Unidos para el periodo 1960-2005 entre el consumo personal, GCP, y el ingreso, PIB, propuesta en

el Ejemplo 2.3 y que la regresión estimada es la siguiente:

$$\widehat{GCP}_t = -299,5913 + 0,721PIB_t$$

La propensión marginal a consumir es 0,72 lo que indica que cuando el ingreso real se incrementa en un dólar el consumo personal aumenta en 72 centavos. La ordenada es  $-299,5913$  lo que indica que si el ingreso es cero el nivel promedio del consumo es negativo e igual a 299,59 dólares. No tiene interpretación económica.

Si las unidades de ambas variables fuese billones de \$: por cada billón de dólares de incremento en el PIB el consumo se incrementaría en 0,721 billones, Luego por cada 100 billones de incremento en PIB el consumo se incrementa en 72,1 billones de dólares. Cuando el PIB es cero el consumo es negativo e igual a 299,5913 billones de dólares.

### Ejemplo 3.3

Vamos a retomar ahora el Ejemplo 2.5 donde se analizaban los determinantes del precio de la vivienda. Se dispone de una base de datos sobre el precio de una vivienda y distintas características de la misma para 14 viviendas vendidas en la comunidad universitaria de San Diego en 1980. Son datos de sección cruzada y la descripción de las variables disponibles es<sup>4</sup>:

PRICE = precio de venta de la vivienda en miles de dólares (Rango 199,9 - 505)

SQFT = pies cuadrados de área habitable (Rango 1065 - 3000)

BEDRMS= número de dormitorios (Rango 3 - 4)

BATHS = número de baños (Rango 1,74 - 3)

Para analizar si el tamaño, el número de habitaciones y el número de baños son factores que explican o no el precio de la vivienda se especifica el siguiente modelo:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, \dots, 14 \quad (3.5)$$

Para estimar el modelo se utilizan las observaciones disponibles en el fichero data4-1.gdt y que son las siguientes<sup>5</sup>:

Las estimaciones obtenidas resultan de aplicar el criterio MCO  $\hat{\beta} = (X'X)^{-1}X'Y$ :

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 14 & \sum SQFT_i & \sum BEDRMS_i & \sum BATHS_i \\ \sum SQFT_i & \sum SQFT_i^2 & \sum SQFT_i BEDRMS_i & \sum SQFT_i BATHS_i \\ \sum BEDRMS_i & \sum BEDRMS_i SQFT_i & \sum BEDRMS_i^2 & \sum BEDRMS_i BATHS_i \\ \sum BATHS_i & \sum BATHS_i SQFT_i & \sum BATHS_i BEDRMS_i & \sum BATHS_i^2 \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} \sum PRICE_i \\ \sum SQFT_i PRICE_i \\ \sum BEDRMS_i PRICE_i \\ \sum BATHS_i PRICE_i \end{bmatrix} = \begin{bmatrix} 14 & 26753 & 51 & 33 \\ 26753 & 55462515 & 99193 & 65699,75 \\ 51 & 99193 & 189 & 121,75 \\ 33 & 65699,75 & 121,75 & 80,375 \end{bmatrix}^{-1} \begin{bmatrix} 4444,9 \\ 9095985,5 \\ 16372,7 \\ 10821,075 \end{bmatrix} = \begin{bmatrix} 129,062 \\ 0,1548 \\ -21,5875 \\ -12,1928 \end{bmatrix}$$

<sup>4</sup>Fuente: Ramanathan, Ramu (2002) *Introductory econometrics with applications*. Conjunto de datos data4-1.gdt

<sup>5</sup>Puedes acceder a estos datos ejecutando `gretl → En Archivo → Abrir datos → Archivo de muestra → Elige Ramanathan, el fichero data4-1.gdt.`

Obsv.	<i>PRICE</i>	<i>SQFT</i>	<i>BEDRMS</i>	<i>BATHS</i>
1	199,9	1065	3	1,75
2	228,0	1254	3	2,00
3	235,0	1300	3	2,00
4	285,0	1577	4	2,50
5	239,0	1600	3	2,00
6	293,0	1750	4	2,00
7	285,0	1800	4	2,75
8	365,0	1870	4	2,00
9	295,0	1935	4	2,50
10	290,0	1948	4	2,00
11	385,0	2254	4	3,00
12	505,0	2600	3	2,50
13	425,0	2800	4	3,00
14	415,0	3000	4	3,00

Tabla 3.1: Datos de características de viviendas. Fichero 4-1.gdt.

- **La función de regresión muestral obtenida es:**

$$\widehat{PRICE}_i = 129,062 + 0,1548 SQFT_i - 21,5875 BEDRMS_i - 12,1928 BATHS_i$$

- **Interpretación de los signos obtenidos:**

Los signos obtenidos son los adecuados. Para la variable *SQFT* el signo es positivo ya que manteniendo el resto de variables constantes lógicamente si aumenta el área habitable aumentará el precio del piso. Si manteniendo el resto de variables constante la superficie habitada aumenta en un pie cuadrado el precio medio estimado de una vivienda aumentará en 154,8 dólares. También son adecuados los signos para *BEDRMS* y *BATHS* ya que en ambos casos se mantiene constante la superficie habitable por lo que se aumenta el número de habitaciones (o baños) a costa de una menor superficie de éstas, lo cual es lógico que se valore negativamente por el comprador medio. Así, si se aumenta el número de habitaciones, manteniendo constante el número de baños y la superficie de la vivienda, el precio medio se estima disminuirá en 21.588 dólares. Manteniéndose constante la superficie habitable y el número de habitaciones el hecho de tener un baño más redundante en habitaciones más pequeñas por lo que se estima que el precio medio se reducirá en 12.193 dólares.

Mediante las estimaciones obtenidas podemos estimar el incremento medio en el precio de la vivienda ante cambios en las variables explicativas. Por ejemplo, si mantenemos el número de baños, tenemos una habitación más y aumenta el área habitable en 500 pies cuadrados, el cambio en el precio medio estimado de una vivienda será de 55,812 dólares:

$$\begin{aligned} \Delta \widehat{PRICE}_i &= 0,1548 \Delta SQFT_i - 21,588 \Delta BEDRMS_i - 12,192 \Delta BATHS_i = \\ &= (0,1548 \times 500) - 21,588 \times 1 - 12,192 \times 0 = 77,4000 - 21,588 = 55,812 \end{aligned}$$

### 3.1.2. Propiedades de la Función de Regresión Muestral, FRM

1. Los residuos son ortogonales a las variables explicativas:  $X'\hat{u} = 0$  ( $\hat{u}'X = 0$ ).

$$X'\hat{u} = X'(Y - \hat{Y}) = X'(Y - X\hat{\beta}) = 0$$

por las ecuaciones normales.

2. Los residuos son ortogonales a las estimaciones de la variable endógena:  $\hat{Y}'\hat{u} = 0$  ( $\hat{u}'\hat{Y} = 0$ ).

$$\hat{Y}'\hat{u} = (X\hat{\beta})'\hat{u} = \hat{\beta}' \underbrace{X'\hat{u}}_{=0} = 0$$

Si el modelo tiene término independiente, es decir, si  $X_{1i} = 1$ , entonces la primera fila de  $X'\hat{u}$  es igual a  $\sum \hat{u}_i$  y tenemos que

3. La suma de los residuos es cero:  $\sum_{i=1}^N \hat{u}_i = 0$ .

$$X'\hat{u} = 0 \Leftrightarrow \begin{bmatrix} \sum_1^N \hat{u}_i \\ \sum_1^N X_{2i}\hat{u}_i \\ \sum_1^N X_{3i}\hat{u}_i \\ \vdots \\ \sum_1^N X_{Ki}\hat{u}_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \sum_{i=1}^N \hat{u}_i = 0$$

4. La media muestral de  $Y$  es igual a la media muestral de las estimaciones de  $Y$ :  $\bar{Y} = \bar{\hat{Y}}$ .

$$\begin{aligned} \hat{u}_i = Y_i - \hat{Y}_i &\Leftrightarrow Y_i = \hat{Y}_i + \hat{u}_i \\ \sum Y_i &= \sum \hat{Y}_i + \underbrace{\sum \hat{u}_i}_{=0} \\ \frac{1}{N} \sum Y_i &= \frac{1}{N} \sum \hat{Y}_i \Rightarrow \bar{Y} = \bar{\hat{Y}} \end{aligned}$$

5. La FRM pasa por el vector de medias:  $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2\bar{X}_2 + \dots + \hat{\beta}_K\bar{X}_K$ .

$$\begin{aligned} \sum_{i=1}^N \hat{u}_i = 0 &\Leftrightarrow \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_K X_{Ki}) = 0 \\ \sum Y_i - N\hat{\beta}_1 - \hat{\beta}_2 \sum X_{2i} - \dots - \hat{\beta}_K \sum X_{Ki} &= 0 \\ \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \dots + \hat{\beta}_K \sum X_{Ki} \\ \frac{1}{N} \sum Y_i &= \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{N} \sum X_{2i} + \dots + \hat{\beta}_K \frac{1}{N} \sum X_{Ki} \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \dots + \hat{\beta}_K \bar{X}_K \end{aligned}$$

**Nota:** Las propiedades 1 y 2 se cumplen siempre, mientras que las 3, 4 y 5 se cumplen **sólo** si el modelo tiene un término independiente.

### 3.1.3. Medidas de bondad del ajuste

Definimos la variación de la variable  $Y$  como la distancia de los valores observados de la variable a su media muestral. La suma de esas variaciones al cuadrado es la variación que se quiere explicar con la variación de las variables explicativas. Se le denota como  $SCT$  y se lee Suma de Cuadrados Total. Lógicamente, el ajuste realizado será mejor cuanto mayor sea la proporción explicada de esa variación.

$$SCT = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - N\bar{Y}^2 = Y'Y - N\bar{Y}^2$$

Cuando el modelo tenga término independiente podremos dividir la variación total en dos partes, variación explicada y variación sin explicar.

Dado que  $Y = \hat{Y} + \hat{u}$ , tenemos:

$$\begin{aligned} Y'Y &= (\hat{Y} + \hat{u})'(\hat{Y} + \hat{u}) = \\ &= \hat{Y}'\hat{Y} + \underbrace{\hat{Y}'\hat{u}}_{=0} + \underbrace{\hat{u}'\hat{Y}}_{=0} + \hat{u}'\hat{u} = \hat{Y}'\hat{Y} + \hat{u}'\hat{u} \end{aligned}$$

Restando en ambos lados  $N\bar{Y}^2$ ,

$$Y'Y - N\bar{Y}^2 = \hat{Y}'\hat{Y} - N\bar{Y}^2 + \hat{u}'\hat{u}$$

Si el modelo tiene término independiente,  $\bar{Y} = \bar{\hat{Y}}$  de donde,

$$\begin{aligned} Y'Y - N\bar{Y}^2 &= \hat{Y}'\hat{Y} - N\bar{\hat{Y}}^2 + \hat{u}'\hat{u} \\ \sum Y_i^2 - N\bar{Y}^2 &= \sum \hat{Y}_i^2 - N\bar{\hat{Y}}^2 + \sum \hat{u}_i^2 \\ \underbrace{\sum (Y_i - \bar{Y})^2}_{SCT} &= \underbrace{\sum (\hat{Y}_i - \bar{\hat{Y}})^2}_{SCE} + \underbrace{\sum \hat{u}_i^2}_{SCR} \\ SCT &= SCE + SCR \end{aligned}$$

siendo<sup>6</sup>:

SCT: Suma de Cuadrados Total, mide la variación total.

SCE: Suma de Cuadrados Explicada, mide la variación explicada.

SCR: Suma de Cuadrados Residual, mide la variación sin explicar.

$$\begin{aligned} SCT &= \sum (Y_i - \bar{Y})^2 = Y'Y - N\bar{Y}^2 \\ SCE &= \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \hat{Y}'\hat{Y} - N\bar{\hat{Y}}^2 \\ SCR &= \sum \hat{u}_i^2 = Y'Y - \hat{Y}'\hat{Y} = Y'Y - \hat{\beta}'X'Y \end{aligned}$$

<sup>6</sup>En el Anexo 1 incluido en el Apéndice del tema aparecen distintas expresiones de la SCT, SCE y SCR que pueden resultar útiles.

**Coefficiente de determinación,  $R^2$** 

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

- Si existe término independiente en el modelo el  $R^2$  estará entre los valores 0 y 1. Por la misma razón si no existe término independiente el  $R^2$  no tiene sentido.
- El coeficiente de determinación mide la bondad del ajuste o lo que es lo mismo la variabilidad de la variable endógena explicada con la variabilidad de las variables exógenas. Es un porcentaje.
- A mayor  $R^2$  mejor ajuste. Podemos tener la tentación de mejorar el ajuste incluyendo variables exógenas y este proceder es un error. El problema que presenta el coeficiente de determinación es que aumenta o se mantiene constante con la inclusión de nuevas variables explicativas en el modelo, aunque éstas no contribuyan a explicar la variable endógena. Debido a este problema, se define otra medida de bondad de ajuste, el coeficiente de determinación corregido,  $\bar{R}^2$ .

**Coefficiente de determinación corregido,  $\bar{R}^2$** 

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{\frac{SCR}{(N-K)}}{\frac{SCT}{(N-1)}} = 1 - \frac{(N-1) SCR}{(N-K) SCT} \\ &= 1 - \frac{(N-1)}{(N-K)}(1 - R^2) \end{aligned}$$

- Cualquiera que sea el número de variables incluidas en un modelo la SCT será constante y por tanto si incluimos una nueva variable la SCR será menor y la SCE será mayor.
- Dado que  $\bar{R}^2$  se define como una ponderación del  $R^2$  por los grados de libertad tendrá en cuenta estos últimos.
- Este coeficiente, penaliza la inclusión de nuevas variables explicativas. Si la nueva variable incluida explica a la variable endógena compensando la pérdida de grados de libertad, es decir compensando el hecho de estimar un coeficiente más, el  $\bar{R}^2$  aumenta. Sin embargo si la nueva variable incluida no explica a la variable endógena compensando la pérdida de grados de libertad el  $\bar{R}^2$  disminuye.
- Si  $K = 1$ ,  $\bar{R}^2 = R^2$ .
- Si  $K > 1$ ,  $\bar{R}^2 \leq R^2$ .

El  $R^2$  y el  $\bar{R}^2$  son sólo dos estadísticos y no deben ser utilizados para comparar la especificación de modelos entre sí, sólo los contrastes de hipótesis que se verán más adelante son la herramienta adecuada.

Existen otros criterios de selección de modelos: el criterio de información de Akaike (AIC) o los criterios Bayesiano de Schwarz (BIC) y de Hannan-Quinn (HQC). Estos criterios se calculan en

función de la suma de cuadrados residual y de algún factor que penalice por la pérdida de grados de libertad. Un modelo más complejo, con más variables explicativas, reducirá la suma de cuadrados residual pero aumentará el factor de penalización. Utilizando estos criterios se escogería aquel modelo con un menor valor de AIC, BIC o HQC. Normalmente no suelen dar la misma elección, siendo el criterio AIC el que elige un modelo con mayor número de parámetros. El cálculo de estos criterios es algo complejo sin embargo el programa *gretl* los muestra automáticamente en el output de regresión. Únicamente los veremos con dicho programa.

**Coefficientes de correlación** El coeficiente de correlación lineal simple mide el grado de asociación lineal entre dos variables. Para  $X$  e  $Y$  se define

$$r_{xy} = \frac{\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N}}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{N}} \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N}}} = \frac{\sum X_i Y_i - N \bar{X} \bar{Y}}{\sqrt{\sum X_i^2 - N \bar{X}^2} \sqrt{\sum Y_i^2 - N \bar{Y}^2}}$$

El coeficiente de correlación simple toma valores entre -1 y 1 y su interpretación podéis recordarla revisando el Tema 1. En el MRLG tendremos una matriz de coeficientes de correlación habitualmente denotada por  $R$ :

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1K} \\ r_{21} & r_{22} & \dots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{K1} & r_{K2} & \dots & r_{KK} \end{bmatrix}$$

La matriz de correlación  $R$  se define como aquella matriz cuyos elementos son el coeficiente de correlación simple entre dos variables  $i$  y  $j$ , tal que:

- $r_{1k}$  representa la correlación entre  $Y$  y  $X_k$   $k = 1, 2, \dots, K$
- $r_{kk} = 1$ , los elementos de la diagonal principal son todos unos. Muestran la correlación de una variable consigo misma.
- $r_{kh}$ , muestran la correlación de la variable exógena  $k$  con la variable exógena  $h$ .
- Además es una matriz simétrica.

En el modelo lineal general la correlación entre  $Y$  y  $X_2$  no está adecuadamente recogida por el coeficiente de correlación simple ya que parte de la variación de  $Y$  será debida al resto de variables exógenas. Será necesario descontar este efecto tanto de  $Y$  como de  $X_2$ . Por ejemplo, en el modelo

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

para estudiar la influencia de  $X_2$  en  $Y$  utilizaremos el coeficiente de correlación parcial entre  $Y$  y  $X_2$  que mide la correlación que queda entre estas dos variables después de eliminar el efecto de  $X_3$  sobre  $Y$  y sobre  $X_2$ .

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

**Ejemplo 3.4**

Con los resultados de la regresión del modelo (2.4) y los datos del fichero *andy.gdt* calculamos el coeficiente de determinación y el coeficiente de determinación corregido:

$$\begin{aligned} SCT &= Y'Y - N\bar{Y}^2 = 452128,4100 - 75 \times 77,375^2 = 3111,6131 \\ SCR &= Y'Y - \hat{\beta}X'Y = 452128,4100 - 450409,4671 = 1718,9429 \\ R^2 &= 1 - \frac{SCR}{SCT} = 1 - \frac{1718,943}{3111,6131} = 0,448258 \\ \bar{R}^2 &= 1 - \frac{(N-1)}{(N-K)}(1 - R^2) = 1 - \frac{75-1}{75-3}(1 - 0,448258) = 0,432932 \end{aligned}$$

La correlación entre  $P$  y  $A$  es:  $corr(P, A) = 0,0263$  un valor muy bajo y positivo. En términos de matriz de correlación:

$$R = \begin{bmatrix} 1,0 & 0,0263 \\ & 1,0 \end{bmatrix}$$

**Ejemplo 3.5**

Con los datos de la Tabla 3.1 y los resultados de la estimación del modelo (3.5) calculamos el coeficiente de determinación y el coeficiente de determinación corregido:

$$\begin{aligned} SCT &= Y'Y - N\bar{Y}^2 = 1512980 - 14 \times 317,493^2 = 101754,7293 \\ SCR &= Y'Y - \hat{\beta}X'Y = 1512980 - 1496279,9 = 16700,1 \\ R^2 &= 1 - \frac{SCR}{SCT} = 1 - \frac{16700,1}{101754,7293} = 0,835976 \\ \bar{R}^2 &= 1 - \frac{(N-1)}{(N-K)}(1 - R^2) = 1 - \frac{14-1}{14-4}(1 - 0,835976) = 0,786769 \end{aligned}$$

Luego el 83,59% de la variabilidad en el precio de la vivienda queda explicada por la variabilidad del tamaño de la vivienda, el número de dormitorios y el número de baños. Es un ajuste bastante alto. El  $\bar{R}^2$  se interpreta de igual manera.

También podemos calcular la matriz de correlaciones entre  $SQFT$ ,  $BEDRMS$  y  $BATHS$ :

$$R = \begin{bmatrix} 1,0 & 0,4647 & 0,7873 \\ & 1,0 & 0,5323 \\ & & 1,0 \end{bmatrix}$$

Luego las variables exógenas están correlacionadas positivamente entre sí. El coeficiente más alto es el coeficiente de correlación simple entre  $SQFT$  y  $BATHS$ .

## 3.2. Propiedades de los estimadores MCO

El método de MCO es sólo uno de los posibles métodos de estimación, la pregunta es ¿cómo podemos elegir entre estimadores? obviamente en base a sus propiedades sobre su comportamiento en muestras repetidas. Estas propiedades son insesgadez, varianza pequeña y error cuadrático medio.

**Insesgadez** Un estimador es insesgado si su valor esperado coincide con el verdadero valor del parámetro. Sea  $\hat{\theta}$  un estimador del parámetro  $\theta$ , será insesgado si  $E(\hat{\theta}) = \theta$ .

**Varianza mínima** Desearemos que la varianza de un estimador sea lo más pequeña posible ya que cuanto menor sea la varianza muestral mayor es la precisión del estimador.

Si estamos comparando dos estimadores insesgados elegiremos aquel que tenga la menor varianza. Pero si estamos comparando dos estimadores sesgados o un estimador sesgado y uno insesgado este criterio no nos sirve y debemos introducir uno nuevo, el concepto de error cuadrático medio.

**Error cuadrático Medio (ECM)**  $ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + Sesgo(\hat{\theta})^2$  donde  $Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$ . En base a este criterio elegimos el estimador con menor ECM.

### 3.2.1. Propiedades de los estimadores MCO

Sea el modelo de regresión lineal general

$$Y = X\beta + u \quad u \sim NID(0, \sigma^2 I_N)$$

donde se cumplen todas las hipótesis básicas. El estimador MCO de los coeficientes

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

tiene las siguientes propiedades:

- Es lineal en las perturbaciones.
- Es insesgado.
- Tiene varianza mínima entre todos los estimadores lineales e insesgados

Demostración:

- **Linealidad.** Como las variables explicativas son no aleatorias, la única variable aleatoria de su expresión es la perturbación, luego el estimador MCO es una combinación lineal de las perturbaciones.

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y = \\ &= (X'X)^{-1}X'(X\beta + u) = \\ &= \beta + (X'X)^{-1}X'u\end{aligned}$$

- **Insesgadez.** Dado que:

$$E(u) = 0$$

y la matriz X es no aleatoria,

$\hat{\beta}_{MCO}$  es insesgado, es decir, su valor esperado es igual al vector de coeficientes del modelo.

$$\begin{aligned} E(\hat{\beta}) &= E(\beta + (X'X)^{-1}X'u) = \\ &= E(\beta) + (X'X)^{-1}X' \underbrace{E(u)}_{=0} = \beta \end{aligned}$$

- **Varianza mínima.** Dado que:

$$E(u) = 0$$

$$E(uu') = \sigma^2 I_N$$

y la matriz X es no aleatoria,

$$\begin{aligned} V(\hat{\beta}) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] = \\ &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \\ &= E\left[\left[(X'X)^{-1}X'u\right]\left[(X'X)^{-1}X'u\right]'\right] = \\ &= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] = \\ &= (X'X)^{-1}X'E[uu']X(X'X)^{-1} = \\ &= (X'X)^{-1}X'\sigma^2I_NX(X'X)^{-1} = \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Esta matriz de varianzas y covarianzas es mínima y nos lo garantiza el Teorema de Gauss-Markov.

$$\begin{aligned} V(\hat{\beta}) &= \begin{bmatrix} V(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & Cov(\hat{\beta}_1, \hat{\beta}_3) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_K) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & V(\hat{\beta}_2) & Cov(\hat{\beta}_2, \hat{\beta}_3) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_K) \\ Cov(\hat{\beta}_3, \hat{\beta}_1) & Cov(\hat{\beta}_3, \hat{\beta}_2) & V(\hat{\beta}_3) & \cdots & Cov(\hat{\beta}_3, \hat{\beta}_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_K, \hat{\beta}_1) & Cov(\hat{\beta}_K, \hat{\beta}_2) & Cov(\hat{\beta}_K, \hat{\beta}_3) & \cdots & V(\hat{\beta}_K) \end{bmatrix} = \\ &= \sigma^2 \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1K} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2K} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{K1} & a_{K2} & a_{K3} & \cdots & a_{KK} \end{bmatrix} = \sigma^2(X'X)^{-1} \end{aligned}$$

donde  $a_{kk}$  es el elemento  $(k, k)$  de  $(X'X)^{-1}$ . Como toda matriz de varianzas y covarianzas, es simétrica.

**Teorema de Gauss-Markov:** Dados los supuestos básicos del modelo de regresión lineal general, “dentro de la clase de estimadores lineales e insesgados,  $\hat{\beta}$  es el estimador eficiente, es decir,  $\hat{\beta}$  tiene mínima varianza”.

### 3.2.2. Estimación de la varianza de las perturbaciones

En la matriz de varianzas y covarianzas del estimador MCO aparece la varianza de las perturbaciones, lo habitual es que sea desconocida y haya de ser estimada. Habitualmente se utiliza el siguiente estimador **insesgado**<sup>7</sup> de  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K} = \frac{SCR}{N-K} = \frac{\sum \hat{u}_i^2}{N-K} \quad \text{y} \quad E(\hat{\sigma}^2) = \sigma^2$$

Por tanto podremos utilizarlo como el estimador apropiado de la varianza de la perturbación. Para trabajar con él es útil escribirlo en términos de las variables observables mediante las matrices  $Y$ ,  $X$ , así:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K} = \frac{Y'Y - \hat{\beta}'X'Y}{N-K} = \frac{Y'Y - \hat{\beta}X'X\hat{\beta}}{N-K}$$

Bajo las hipótesis básicas, **un estimador insesgado de la matriz de varianzas y covarianzas**, de  $\hat{\beta}_{MCO}$  es

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

#### Ejemplo 3.6

Con los datos disponibles en el fichero andy.gdt y los resultados de la estimación del modelo (2.4) se calcula la siguiente matriz de varianzas y covarianzas estimada:

$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}'X'Y}{N-K} = \frac{452128,4100 - 450409,4671}{75-3} = \frac{1718,943}{72} = 23,8742$$

$$\begin{aligned} \hat{V}(\hat{\beta}_{MCO}) &= 23,8742 \times \begin{bmatrix} 75,0000 & 426,5400 & 138,3000 \\ 426,5400 & 2445,7074 & 787,3810 \\ 138,3000 & 787,3810 & 306,2100 \end{bmatrix}^{-1} = \\ &= \begin{bmatrix} 40,34330 & -6,79506 & -0,74842 \\ -6,79506 & 1,20120 & -0,01974 \\ -0,74842 & -0,01974 & 0,46675 \end{bmatrix} \end{aligned}$$

<sup>7</sup>En el Anexo 2 incluido en el Apéndice del Tema aparece demostrada la insesgades de  $\hat{\sigma}^2$ .

**Ejemplo 3.7**

Con los datos de la Tabla 3.1 y los resultados de la estimación del modelo (3.5) se calcula la siguiente matriz de varianzas y covarianzas estimada:

$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}'X'Y}{N - K} = \frac{1513039,0100 - 1496338,9414}{14 - 4} = 1670,0069$$

$$\hat{V}(\hat{\beta}_{MCO}) = 1670,0069 \times \begin{bmatrix} 14 & 26753 & 51 & 33 \\ 26753 & 55462515 & 99193 & 65699,75 \\ 51 & 99193 & 189 & 121,75 \\ 33 & 65699,75 & 121,75 & 80,375 \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} 7797,47 & 0,670891 & -1677,13 & -1209,37 \\ 0,670891 & 0,00102019 & -0,0754606 & -0,995066 \\ -1677,13 & -0,0754606 & 730,585 & -356,4 \\ -1209,37 & -0,995066 & -356,4 & 1870,56 \end{bmatrix}$$

**Ejemplo 3.8**

Vamos a retomar el Ejemplo 2.6 utilizado para ilustrar la especificación de un modelo que recoge sólo efectos cualitativos, es decir tenemos un único conjunto de variables ficticias. Estamos comparando medias.

Suponíamos que disponíamos de datos de salarios de hombres y mujeres,  $W_i$  y creemos que, en media, existen diferencias salariales entre estos dos grupos. Para contrastar que esto es cierto podemos recoger el efecto cualitativo sexo sobre el salario utilizando las variables ficticias:

$$S_{1i} = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{en caso contrario} \end{cases} \quad S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

y podemos especificar el siguiente modelo como ya se hizo en el Ejemplo 2.6:

$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M \quad u_i \sim NID(0, \sigma^2)$$

Recordemos que  $\beta_1$  es el salario esperado cuando el individuo es hombre,  $\beta_1 + \beta_2$  es el salario esperado de una mujer y  $\beta_2$  recoge el efecto diferencial en el salario esperado entre hombres y mujeres. Si no existiera discriminación salarial por sexo, es decir si hombres y mujeres tuvieran el mismo salario, su valor sería cero.

- Estimación del modelo anterior:

$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M$$

$$\begin{bmatrix} W_H \\ W_M \end{bmatrix} = \begin{bmatrix} i_H & 0 \\ i_M & i_M \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_H \\ u_M \end{bmatrix} \Rightarrow Y = X\beta + u$$

Notación utilizada:  $N_H$  es el número de individuos varones y  $N_M$  el número de mujeres.  $W_H, W_M$  son vectores columna que recogen los salarios de hombres y mujeres, por tanto de orden  $N_H \times 1$  y  $N_M \times 1$ , respectivamente.  $i_H, i_M$  son vectores de unos de tamaño  $N_H \times 1$  y  $N_M \times 1$  respectivamente.

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \left[ \begin{bmatrix} i'_H & i'_M \\ 0 & i'_M \end{bmatrix} \begin{bmatrix} i_H & 0 \\ i_M & i_M \end{bmatrix} \right]^{-1} \begin{bmatrix} i'_H & i'_M \\ 0 & i'_M \end{bmatrix} \begin{bmatrix} W_H \\ W_M \end{bmatrix} = \\ &= \begin{bmatrix} N_H + N_M & N_M \\ N_M & N_M \end{bmatrix}^{-1} \begin{bmatrix} \sum W_H + \sum W_M \\ \sum W_M \end{bmatrix} = \begin{bmatrix} \bar{W}_H \\ \bar{W}_M - \bar{W}_H \end{bmatrix} \end{aligned}$$

que sería el equivalente a estimar cada ecuación por separado, en las dos ecuaciones a las que da lugar el modelo (2.14):

$$\begin{aligned} W_i &= \beta_1 + u_i \quad i = 1, \dots, N_H && \text{para los hombres} \\ W_i &= \beta_1 + \beta_2 + u_i \quad i = 1, \dots, N_M && \text{para las mujeres} \end{aligned}$$

- Alternativa de especificación :

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M$$

de donde suponiendo  $u_i \sim NID(0, \sigma^2)$

$\alpha_1 = E(W_i | S_{1i} = 1; S_{2i} = 0)$  es el salario esperado de un hombre

$\alpha_2 = E(W_i | S_{1i} = 0; S_{2i} = 1)$  es el salario esperado de una mujer

por tanto estos coeficientes recogen el salario medio dentro del grupo.

- Estimación del modelo alternativo:

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + u_i \quad i = 1, \dots, N_H + N_M$$

$$\begin{bmatrix} W_H \\ W_M \end{bmatrix} = \begin{bmatrix} i_H & 0 \\ 0 & i_M \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} u_H \\ u_M \end{bmatrix} \Rightarrow Y = X\beta + u$$

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

$$\begin{aligned} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} &= \left[ \begin{bmatrix} i'_H & 0 \\ 0 & i'_M \end{bmatrix} \begin{bmatrix} i_H & 0 \\ 0 & i_M \end{bmatrix} \right]^{-1} \begin{bmatrix} i'_H & 0 \\ 0 & i'_M \end{bmatrix} \begin{bmatrix} W_H \\ W_M \end{bmatrix} = \\ &= \begin{bmatrix} N_H & 0 \\ 0 & N_M \end{bmatrix}^{-1} \begin{bmatrix} \sum W_H \\ \sum W_M \end{bmatrix} = \begin{bmatrix} \sum W_H / N_H \\ \sum W_M / N_M \end{bmatrix} = \begin{bmatrix} \bar{W}_H \\ \bar{W}_M \end{bmatrix} \end{aligned}$$

$$\hat{W}_i = \hat{\alpha}_1 S_{1i} + \hat{\alpha}_2 S_{2i} = \bar{W}_H S_{1i} + \bar{W}_M S_{2i}$$

Los mismos resultados se obtendrían si hubiésemos estimado las ecuaciones por separado en las dos ecuaciones a que da lugar la especificación alternativa:

$$W_i = \alpha_1 + u_i \quad i = 1, \dots, N_H \quad \text{y} \quad W_i = \alpha_2 + u_i \quad i = 1, \dots, N_H$$

### Ejercicio 3.3

Supongamos que en la regresión mínimo cuadrática ordinaria de Y sobre X el coeficiente estimado de X es 1.2. Di si son ciertas o falsas las siguientes afirmaciones, razonando las respuestas:

1. Tomando distintas muestras para la variable Y pueden obtenerse otras estimaciones.
2. La distribución de estas estimaciones debe estar centrada en torno al verdadero valor 1.2.

### Ejercicio 3.4

Obtener la esperanza matemática y varianza del siguiente estimador del parámetro de pendiente en el MLS cuando se cumplen todas las hipótesis básicas:

$$\tilde{\beta} = \hat{\beta}_{MCO} + \frac{c}{N}$$

donde c es una constante arbitraria.

### Ejercicio 3.5

Se quiere estimar un modelo que explique el ahorro generado,  $S$ , en función del tipo de interés,  $r$ , de la forma:

$$S_t = \alpha + \beta r_t + u_t$$

1. Para estimar con mayor precisión  $\beta$ , si pudieras elegir la muestra en diferentes períodos, ¿la elegirías durante un período de tiempo en el cual los tipos de interés fueran fluctuantes o durante un período en el cual los tipos de interés fueran relativamente constantes?
2. ¿Qué ocurriría con los estimadores MCO de los coeficientes de la regresión si el ahorro ha fluctuado muy poco en torno a un valor constante durante el período muestral?

### Ejercicio 3.6

Supongamos que la relación entre las variables  $X$  e  $Y$  es la siguiente  $Y_t = \beta X_t + u_t$  donde  $X$  es una variable no estocástica y  $u_t \sim NID(0, \sigma^2)$ . Se definen dos estimadores para el parámetro desconocido  $\beta$ :

$$\beta^* = \frac{\sum Y_t}{\sum X_t} \quad \hat{\beta} = \frac{\sum X_t Y_t}{\sum X_t^2}$$

1. Demuestra que ambos estimadores son insesgados.
2. ¿Cuál de los dos estimadores elegirías? ¿Por qué?

### Ejercicio 3.7

Para el vector de perturbaciones  $u$  de orden  $(T \times 1)$ , indica el significado de cada una de las siguientes expresiones y establece su relación con  $\sigma^2$ :

1.  $\frac{1}{T} u' u$
2.  $\frac{1}{T} \hat{u}' \hat{u}$
3.  $E(\hat{u}' \hat{u})$
4.  $E(\hat{u} \hat{u}')$

### Ejercicio 3.8

Se han calculado las siguientes cantidades en una muestra de tamaño  $T = 100$ :

$$\begin{aligned} \sum X_t &= 11,34 & \sum X_t^2 &= 12,16 & \sum Y_t &= 20,72 \\ \sum Y_t^2 &= 84,96 & \sum X_t Y_t &= 22,13 \end{aligned}$$

Estimar las siguientes regresiones:

1.  $Y_t = \beta_1 + \beta_2 X_t + u_t$
2.  $X_t = \alpha_1 + \alpha_2 Y_t + \epsilon_t$
3. ¿Son los estimadores de  $\beta_1$  y  $\beta_2$  iguales a los de  $\alpha_1$  y  $\alpha_2$ ? ¿En qué caso coincidirán?

### Ejercicio 3.9

Supón que el verdadero modelo de regresión es:

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t \quad (3.6)$$

pero nosotros añadimos la variable  $X_3$  y estimamos el siguiente modelo:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (3.7)$$

Discute si el coeficiente de determinación  $R^2$  y el corregido  $\bar{R}^2$  serán mayores en el modelo (3.6) que en el (3.7).

**Ejercicio 3.10**

Comenta con claridad las siguientes afirmaciones:

1. El estimador insesgado  $\hat{\beta}$  con varianza  $V_1$  proporciona siempre mejores estimaciones que el estimador sesgado  $\beta^*$  con varianza  $V_2$  donde  $V_1 > V_2$ .
2. En el modelo de regresión lineal general la matriz de covarianzas estimada del estimador MCO de  $\hat{\beta}$  es la misma para cualquier muestra que tomemos.
3. Las ecuaciones normales del modelo de regresión lineal general implican que el vector de residuos MCO es ortogonal al vector de valores estimados  $\hat{Y}$ .
4. Si  $Y_t$  y  $X_{2t}$  son variables que no están correlacionadas, es decir,  $\sum(Y_t - \bar{Y})(X_{2t} - \bar{X}_2) = 0$ , entonces la estimación MCO del coeficiente  $\beta_2$  en el modelo:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + u_t \quad t = 1, 2, \dots, T$$

es cero.

5. En el MRLG, el supuesto de normalidad de las perturbaciones es necesario para poder estimar los parámetros del modelo.
6. Bajo los supuestos clásicos del modelo de regresión lineal general, el estimador MCO de  $\beta$  sigue una distribución normal multivariante centrada en el verdadero valor de los parámetros  $\beta$ .
7. Dada una muestra de  $N$  observaciones, la ecuación de regresión poblacional pasa por el punto de medias  $(\bar{Y}, \bar{X}_1, \dots, \bar{X}_k)$ .
8. En un modelo de regresión lineal simple donde  $u_t \sim iid(0, \sigma^2)$ , si se pudiera elegir entre dos muestras de valores para la variable  $X$ , una de ellas con  $\sum(X_t - \bar{X})^2 = 100$  y la otra con  $\sum(X_t - \bar{X})^2 = 1000$ , se elegiría la segunda muestra porque la precisión al estimar los parámetros sería mayor.
9. Los residuos mínimo-cuadrático ordinarios de una regresión de  $Y$  sobre la variable explicativa  $X$  son ortogonales a  $X$ , pero no lo son a  $Y$ .
10. En un modelo de regresión lineal simple no existe ninguna relación entre el estimador MCO de la pendiente de la regresión  $\hat{\beta}$  y el coeficiente de correlación simple entre la variable dependiente  $Y$  y la variable independiente  $X$ ,  $r_{XY}$ .
11. Si existe la siguiente relación entre las variables  $X$ ,  $Y$ ,  $Z$ :

$$r_{XY}^2 = 0,8 \quad r_{XZ}^2 = 0,001 \quad r_{YZ}^2 = 0,001$$

podemos concluir que el coeficiente  $r_{XY}^2$  es una mala medida de la correlación entre las variables  $X$  e  $Y$  y se debería utilizar  $r_{XZ}^2$ .

12. La inexistencia de normalidad en las perturbaciones aleatorias no afecta a los contrastes de hipótesis.

### Ejercicio 3.11

En un estudio sobre el número de siniestros automovilísticos declarados se han obtenido los siguientes resultados estimando por MCO:

$$\hat{Y}_i = 40,5 + 10,5D_{1i} + 0,4X_i \quad (3.8)$$

Y ahora proponemos el modelo alternativo:

$$Y_i = \alpha D_{2i} + \alpha^* D_{1i} + \beta X_i + u_i \quad (3.9)$$

donde

$$D_{1i} = \begin{cases} 1 & \text{si usa el automóvil para trabajar} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{2i} = \begin{cases} 1 & \text{si no usa el automóvil para trabajar} \\ 0 & \text{en caso contrario} \end{cases}$$

Explica y halla los valores numéricos de  $\alpha$ ,  $\alpha^*$  y  $\beta$  en (3.9) con la información facilitada por (3.8).

### Ejercicio 3.12

Plantea el modelo de determinación del salario del trabajador de la empresa A en función de sus años de experiencia, categoría profesional (empleado, técnico, directivo) y sexo (masculino, femenino).

1. Escribe la función de salarios de una técnica.
2. Interpreta los coeficientes del modelo.
3. Suponiendo que la empresa tiene 40 trabajadores, los cuales tienen la siguiente distribución por categorías 5 técnicas y 2 directivos, 4 directivos, 2 técnicos y el resto son empleados. Escribe la matriz de regresores.

### Ejercicio 3.13

Se han utilizado datos trimestrales para estimar la función:

$$C_t = \alpha + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \quad (3.10)$$

donde:  $D_{it} = \begin{cases} 1 & \text{si } t \text{ pertenece al } i\text{-ésimo trimestre} \\ 0 & \text{en caso contrario} \end{cases}$

1. Indica qué parámetros son estimables y cuáles no, razonando la respuesta. ¿Y si se impone la restricción  $\sum \beta_i = 0$ ?
2. Un investigador impone la restricción  $\alpha = 0$  y otro impone la restricción  $\beta_4 = 0$ . ¿Serán las estimaciones de  $\beta_i$  diferentes en ambos casos? ¿por qué?

### 3.3. Estimación del MRLG con gretl: principales resultados

#### • Estimación por Mínimos Cuadrados Ordinarios, MCO:

Una vez abierto el fichero de datos con el que vamos a trabajar, vamos a

*Modelo* → *Mínimos Cuadrados Ordinarios*

Aparecerá una ventana para especificar la parte sistemática del modelo donde debemos:

Seleccionar la variable dependiente pinchando a la izquierda sobre ella y a continuación pinchar en la derecha → *la flecha azul*

Seleccionar las variables independientes pinchando a la izquierda sobre ella-s y a continuación pinchar en la derecha → *la flecha verde*

Para obtener los resultados de la estimación MCO pinchar en *Aceptar*. No pinchar en la indicación *Desviaciones Típicas Robustas*.

En esta ventana aparecerán los resultados básicos de la estimación del modelo. Los podemos guardar como texto plano de la manera habitual o como icono con *Archivo* → *Guardar como icono*.

Los resultados que *gretl* nos devuelve muestran entre otros estadísticos la estimación de los parámetros de la recta de ajuste, sus desviaciones típicas y estadísticos de significatividad individual.

#### Ejemplo 3.9

Vamos a utilizar como ejemplo la estimación realizada en el Ejemplo 2.13 sobre el precio de la vivienda y los contrastes desarrollados en el Ejemplo 3.2:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS + u_i \quad i = 1, \dots, 14$$

Los resultados de la estimación MCO mostrados por *gretl* son los siguientes:

Modelo 1: MCO, usando las observaciones 1–14

Variable dependiente: price

	Coefficiente	Desv. Típica	Estadístico <i>t</i>	Valor p
const	129,062	88,3033	1,4616	0,1746
sqft	0,154800	0,0319404	4,8465	0,0007
bedrms	-21,5875	27,0293	-0,7987	0,4430
baths	-12,1928	43,2500	-0,2819	0,7838
Media de la vble. dep.		317,4929	D.T. de la vble. dep.	88,49816
Suma de cuad. residuos		16700,07	D.T. de la regresión	40,86572
$R^2$		0,835976	$R^2$ corregido	0,786769
$F(3, 10)$		16,98894	Valor p (de $F$ )	0,000299
Log-verosimilitud		-69,45391	Criterio de Akaike	146,9078
Criterio de Schwarz		149,4641	Hannan–Quinn	146,6712

En la columna con encabezamiento *Coefficiente* aparece la estimación del coeficiente que acompaña a la correspondiente variable. A continuación aparece su *Desviación Típica* y

el estadístico  $t$  de significatividad individual para el contraste  $H_0 : \beta_k = 0$  así como su correspondiente *valor p*.

A continuación aparecen estadísticos de interés como pueden ser la media de la variable dependiente,  $R^2$  o  $\bar{R}^2$  entre otros. La fila:  $F(3, 10) = 16,98894$ ; Valor  $p$  (de  $F$ ) = 0,000299 se corresponde con el valor muestral del estadístico  $F$  para el contraste de significatividad conjunto y su correspondiente *valor-p*. A continuación aparecen los estadísticos de Akaike, Schwarz y Hannan-Quinn para la selección de modelos.

En la pestaña *Contrastes* que aparece en la pantalla de resultados de la regresión podemos *Omitir u añadir variables, sumar los coeficientes y contrastar combinaciones lineales o restricciones lineales* además podremos realizar contrastes sobre los residuos. De todo ello nos ocuparemos en los temas siguientes.

En la pantalla de resultados de la estimación aparecen en la barra de menú otros estadísticos o resultados que pueden ser de interés, por ejemplo:

- Podemos hacer gráficos de interés: En la opción *Gráficos* podemos hacer gráficos que nos ayudan a interpretar los resultados de la estimación, por ejemplo  
*Gráficos* → *Gráfico de la variable estimada y observada*  
*Gráficos* → *Gráfico de residuos* → *contra alguna de las variables explicativas del modelo*
- En la pestaña *Guardar* podemos guardar variables como los residuos, los residuos al cuadrado, la suma de cuadrados residual y el coeficiente de determinación entre otros.
- En la pestaña *Análisis* nos muestra las estimaciones de la variable endógena, los intervalos de confianza de los coeficientes y la matriz de varianzas y covarianzas entre otros resultados. Para ver y guardar los valores de  $\hat{Y}$ ,  $\hat{u}$  y otros resultados de utilidad:
  - Ver los valores: Pinchar en *Análisis* → *Mostrar variable* y seleccionar *observada, estimada o residuos* según nuestro interés.
  - Guardar los valores: Pinchar en *Guardar* → *seleccionar la variable de interés*.

*Gretl* utiliza por defecto la denominación *yhat*, *uhat* para designar a la variable endógena estimada y a los residuos, respectivamente y en la descripción de la variable indicará por ejemplo para *uhat*: *residuos del modelo 1*, donde el valor 1 indica que corresponde con el primer modelo estimado, esto resulta muy útil pues en general trabajaremos con varios modelos a la vez y hay que distinguir claramente las variables de cada uno.

### Ejemplo 3.10

La matriz de varianzas y covarianzas de los coeficientes estimados es:

Matriz de covarianzas de los coeficientes

const	sqft	bedrms	baths	
7797,5	0,67089	-1677,1	-1209,4	const
	0,0010202	-0,075461	-0,99507	sqft
		730,58	-356,40	bedrms
			1870,6	baths

Los intervalos de confianza de los coeficientes son:

$$t(10, 0, 025) = 2, 228$$

Variable	Coefficiente	Intervalo de confianza 95 %	
const	129,062	-67,6903	325,814
sqft	0,154800	0,0836321	0,225968
bedrms	-21,5875	-81,8126	38,6376
baths	-12,1928	-108,560	84,1742

### 3.3.1. Tratamiento de las variables ficticias en gretl.

*Gretl* permite trabajar tanto con variables ficticias cuantitativas como cualitativas y su tratamiento no difiere, solo debemos de ocuparnos de especificar correctamente el modelo. En el caso de que la variable ficticia no esté construida *gretl* permite hacerlo. En la pantalla inicial en *Añadir* podemos añadir *Variables ficticias periódicas* que se ajustarán lógicamente a la periodicidad muestral del conjunto de datos, *Variables ficticias para las variables discretas seleccionadas* donde por ejemplo si tenemos una variable que toma valores 1, 2 y 3 podremos construir tres variables ficticias tal como

$$D_1 = \begin{cases} 1 & \text{si la variable toma valor 1} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{si la variable toma valor 2} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{si la variable toma valor 3} \\ 0 & \text{en caso contrario} \end{cases}$$

Por supuesto también podremos introducirlas con el editor tal y como se aprendió en el Tema 1.

Veamos un ejemplo aplicado. Abrimos el fichero de datos data7-3 de Ramanathan, que contiene datos para 14 viviendas sobre el precio de venta de la vivienda (PRICE), pies cuadrados habitables (SQFT), número de habitaciones (BEDRMS) y número de baños (BATHS), y una variable ficticia que toma el valor 1 si la vivienda tiene piscina y 0 en caso contrario (POOL), una variable ficticia que toma el valor 1 si la vivienda tiene sala de estar y 0 en caso contrario (FAMROOM) y una variable ficticia que toma el valor 1 si la vivienda tiene chimenea y 0 en caso contrario (FIREPL). Seleccionamos las variables PRICE y POOL y observamos los valores de estas dos variables:

Obs	price	pool
-----	-------	------

### Técnicas de previsión de demanda

1	199,9	1
2	228,0	0
3	235,0	1
4	285,0	0
5	239,0	0
6	293,0	0
7	285,0	0
8	365,0	1
9	295,0	0
10	290,0	0
11	385,0	1
12	505,0	1
13	425,0	0
14	415,0	0

Por ejemplo, la primera vivienda de la muestra tiene un precio de 199.900 dólares y tiene piscina (ya que la variable POOL toma el valor 1), mientras que la segunda no tiene piscina (la variable POOL toma el valor 0) y su precio de venta es de 228.000 dólares, etc.

Con los datos anteriores podemos obtener fácilmente que el precio medio de la vivienda es 317.493 dólares:

Estadísticos principales, usando las observaciones 1 - 14  
para la variable price (14 observaciones válidas)

Media	Mediana	Mínimo	Máximo
317,49	291,50	199,90	505,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
88,498	0,27874	0,65346	-0,52983

Sin embargo, también es posible obtener el precio medio para las viviendas que tienen piscina, por un lado, y para las que no la tienen, por otro. Para ello, en primer, lugar se selecciona el precio para aquellas viviendas con piscina. Seleccionamos la variable PRICE, pinchamos en *Muestra* → *Definir a partir de v. ficticia...*, seleccionamos la variable POOL y aceptamos.

De esta forma hemos seleccionado el precio para aquellas viviendas que tienen piscina<sup>8</sup>. A continuación, se obtienen los estadísticos principales:

Estadísticos principales, usando las observaciones 1 - 5  
para la variable price (5 observaciones válidas)

---

<sup>8</sup>Para restablecer el tamaño muestral inicial pinchar en *Muestra* → *Recuperar el rango completo*.

Media	Mediana	Mínimo	Máximo
337,98	365,00	199,90	505,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
122,99	0,36390	0,15896	-1,2798

Para seleccionar el precio de las viviendas que no tienen piscina, pinchamos en *Muestra*  $\rightarrow$  *Restringir a partir de criterio*, introducimos la condición  $POOL = 0$  y aceptamos. Los estadísticos principales son los siguientes:

Estadísticos principales, usando las observaciones 1 - 9  
para la variable price (9 observaciones válidas)

Media	Mediana	Mínimo	Máximo
306,11	290,00	228,00	425,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
68,959	0,225275	0,87575	-0,52255

Por tanto, el precio medio de las viviendas con piscina es de 337.980 dólares frente a los 306.111 de las viviendas sin piscina. Dado el modelo una vivienda con piscina es en promedio 31.869 dólares más cara que la que no tiene piscina. Notar que no se están teniendo en cuenta otros factores que pueden afectar al precio de la vivienda (número de pies cuadrados habitables, número de habitaciones, etc.).

El sencillo análisis anterior podemos realizarlo mediante un análisis de regresión. Podemos especificar un modelo econométrico utilizando la variable ficticia  $POOL$  como regresor, estimarlo, hacer inferencia e ir incorporando otras características que pueden afectar a los precios de las viviendas. Para comenzar, consideramos el siguiente modelo:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + u_i \quad i = 1, \dots, 14 \quad (3.11)$$

donde

- $\alpha_1$ : precio medio de una vivienda sin piscina.
- $\alpha_1 + \alpha_2$ : precio medio de una vivienda con piscina.
- $\alpha_2$ : diferencia en el precio medio de una vivienda con piscina con respecto a una que no la tiene.

Los resultados de estimar el modelo por Mínimos Cuadrados Ordinarios utilizando *gretl* obtenemos que las estimaciones de los coeficientes son las siguientes:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14  
Variable dependiente: price

Variable	Coefficiente	Desv. típica	Estadístico $t$	valor p
const	306,111	30,2077	10,1335	0,0000
pool	31,8689	50,5471	0,6305	0,5402
Media de la var. dependiente			317,493	
D.T. de la variable dependiente			88,4982	
Suma de cuadrados de los residuos			98550,5	
Desviación típica de los residuos ( $\hat{\sigma}$ )			90,6231	
$R^2$			0,0320632	
$\bar{R}^2$ corregido			-0,0485982	
Grados de libertad			12	
Log-verosimilitud			-81,880	
Criterio de información de Akaike			167,760	
Criterio de información Bayesiano de Schwarz			169,038	

$$\widehat{PRICE}_i = 306,111 + 31,869 POOL_i \quad i = 1, \dots, 14$$

(10,13)      (0,63)

Para contrastar en el modelo (4.25) si hay diferencias significativas en el precio medio de la vivienda entre aquéllas que tienen piscina y las que no, la hipótesis de contraste es  $H_0 : \alpha_2 = 0$ . Este contraste se puede realizar utilizando el estadístico  $t$  habitual cuyo p-valor es 0,5405, por lo que no se rechaza la hipótesis nula para un nivel de significación del 5%, es decir, el precio medio de la vivienda no es significativamente diferente por el hecho de tener piscina. Alternativamente, se puede realizar el contraste utilizando el estadístico  $F$  basado en las sumas de cuadrados de los residuos siendo en este caso el modelo (4.25) el modelo no restringido mientras que el modelo restringido es  $PRICE_i = \alpha_1 + u_i \quad i = 1, \dots, 14$ .

Supongamos que ampliamos el modelo (4.25) incorporando regresores que podrían explicar el precio de la vivienda como: el hecho de que la vivienda tenga sala de estar o no, el hecho que tenga chimenea o no, su superficie, el número de habitaciones y el número de baños. Las dos primeras son variables ficticias que pueden definirse así:

$$FIREPL_i = \begin{cases} 1 & \text{si la vivienda } i\text{-ésima tiene chimenea} \\ 0 & \text{en caso contrario} \end{cases}$$

$$FAMROOM_i = \begin{cases} 1 & \text{si la vivienda } i\text{-ésima tiene sala de estar} \\ 0 & \text{en caso contrario} \end{cases}$$

Mientras que la superficie, el número de baños y el número de habitaciones se definen como en los temas anteriores:

$SQFT_i$  tamaño de la vivienda  $i$ -ésima en pies cuadrados  
 $BEDRMS$  número de habitaciones de la vivienda  $i$ -ésima  
 $BATHS$  número de cuartos de baño de la vivienda  $i$ -ésima

Con todas ellas podemos especificar el siguiente modelo para explicar el precio de la vivienda:

$$PRICE_i = \gamma_1 + \gamma_2 POOL_i + \gamma_3 FAMROOM_i + \gamma_4 FIREPL_i + \beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \dots, 14 \quad (3.12)$$

Donde lo primero a notar es que en el modelo (4.26), afectando a la ordenada, conviven tres conjuntos de variables ficticias con dos categorías cada una, el hecho de tener o no piscina, el hecho de tener o no chimenea y el hecho de tener o no sala de estar, de las cuales sólo se incluye una de cada conjunto y se mantiene el término independiente. Esta forma de definir el modelo es muy cómoda ya que sigue manteniendo los resultados de los modelos con término independiente y permite una fácil interpretación de los coeficientes que acompañan a las variables ficticias. Así,  $\gamma_i$   $i = 2, 3, 4$  recogen el diferencial en el valor esperado de una vivienda por el hecho de poseer la característica correspondiente manteniéndose constante el resto de variables. El resultado de la estimación es:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14  
Variable dependiente: price

Variable	Coefficiente	Desv. típica	Estadístico $t$	valor p
const	39,0571	89,5397	0,4362	0,6758
pool	53,1958	22,0635	2,4110	0,0467
famroom	-21,344	42,8734	-0,4979	0,6338
firepl	26,1880	53,8454	0,4864	0,6416
sqft	0,146551	0,0301014	4,8686	0,0018
bedrms	-7,0455	28,7363	-0,2452	0,8134
baths	-0,263691	41,4547	-0,0064	0,9951
Media de la var. dependiente			317,493	
D.T. de la variable dependiente			88,4982	
Suma de cuadrados de los residuos			9010,24	
Desviación típica de los residuos ( $\hat{\sigma}$ )			35,8773	
$R^2$			0,911504	
$\bar{R}^2$ corregido			0,835650	
$F(6, 7)$			12,0166	
valor p para $F()$			0,00221290	
Log-verosimilitud			-65,134	
Criterio de información de Akaike			144,269	
Criterio de información Bayesiano de Schwarz			148,743	

### La interpretación de los coeficientes estimados es la siguiente:

- $\hat{\gamma}_1 = 39,057$ : el precio medio estimado de las viviendas sin piscina, baños, habitaciones, sala de estar ni chimenea y con 0 pies cuadrados habitables es de 39.057 dólares.
- $\hat{\gamma}_2 = 53,1958$ : la diferencia estimada en el precio medio de las viviendas con piscina con respecto a las que no la tienen, siendo iguales en el resto de características (pies cuadrados habitables, habitaciones, baños, sala de estar y chimenea) es de 53.196 dólares.

- $\hat{\gamma}_3 = -21,34$ : el precio medio estimado de una vivienda con sala de estar es 21.340 dólares inferior al de una sin sala de estar, siendo idénticas en el resto de características. Esto se debe a que, al mantener constante el número de pies cuadrados de la vivienda y el número de habitaciones y baños, incluir una sala de estar hará que el resto de habitaciones o baños sean de menor tamaño.
- $\hat{\gamma}_4 = 26,188$ : el precio medio estimado de una vivienda con chimenea es 26.188 dólares más caro que el de una sin chimenea, siendo idénticas en el resto de características.
- $\hat{\beta}_1 = 0,147$ : el precio medio estimado de una vivienda se incrementa en 147.000 dólares al aumentar en 1 pie cuadrado habitable su superficie, permaneciendo constantes el número de baños y habitaciones.
- $\hat{\beta}_2 = -7,046$ : el precio medio estimado de una vivienda disminuye en 7.046 dólares al aumentar en 1 el número de habitaciones, permaneciendo constantes el número de baños y los pies cuadrados habitaciones. Esto se debe a que las habitaciones serán de menor tamaño.
- $\hat{\beta}_3 = -0,264$ : el precio medio estimado de una vivienda disminuye en 264 dólares al aumentar en 1 el número de baños, permaneciendo constantes el número de habitaciones y los pies cuadrados habitables. De nuevo, las habitaciones serán de menor tamaño.

## 3.4. Apéndice

### 3.4.1. Anexo 1. Distintas expresiones de SCT, SCE y SCR

$$\begin{aligned} SCT &= \sum (Y_i - \bar{Y})^2 = \\ &= \sum Y_i^2 - N\bar{Y}^2 = Y'Y - N\bar{Y}^2 \end{aligned}$$

$$\begin{aligned} SCE &= \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \\ &= \sum \hat{Y}_i^2 - N\bar{\hat{Y}}^2 = \\ &= \hat{Y}'\hat{Y} - N\bar{\hat{Y}}^2 = \\ &= (X\hat{\beta})'(X\hat{\beta}) - N\bar{\hat{Y}}^2 = \\ &= \hat{\beta}' \underbrace{X'X\hat{\beta}}_{=X'Y} - N\bar{\hat{Y}}^2 = \\ &= \hat{\beta}'X'Y - N\bar{\hat{Y}}^2 \end{aligned}$$

$$\begin{aligned}
SCR &= \sum \hat{u}_i^2 = \hat{u}'\hat{u} = \\
&= (Y - \hat{Y})'(Y - \hat{Y}) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = \\
&= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} = \\
&= Y'Y - \hat{\beta}'X'X\hat{\beta} = \\
&= Y'Y - \hat{\beta}'X'Y
\end{aligned}$$

### 3.4.2. Anexo 2. Demostración de la insesgadez de $\hat{\sigma}^2$

#### 1. Propiedades de los residuos MCO

Los residuos MCO se pueden escribir en función de las perturbaciones:

$$\begin{aligned}
\hat{u} &= Y - \hat{Y} = \\
&= Y - X\hat{\beta} = \\
&= Y - X(X'X)^{-1}X'Y = \\
&= [I_N - X(X'X)^{-1}X']Y = \\
&= [I_N - X(X'X)^{-1}X'](X\beta + u) = \\
&= X\beta - X(X'X)^{-1}X'X\beta + [I_N - X(X'X)^{-1}X']u = \\
&= \underbrace{[I_N - X(X'X)^{-1}X']}_{=M}u = Mu
\end{aligned}$$

Las características de la matriz M son:

- Es simétrica:  $M = M'$

$$\begin{aligned}
M' &= [I_N - X(X'X)^{-1}X']' \\
&= I_N' - (X(X'X)^{-1}X')' \\
&= I_N - (X')'[(X'X)^{-1}]'X' \\
&= I_N - X(X'X)^{-1}X' \\
&= M
\end{aligned}$$

- Es idempotente:  $MM = M$

$$\begin{aligned}
MM &= [I_N - X(X'X)^{-1}X'] [I_N - X(X'X)^{-1}X'] = \\
&= I_N - X(X'X)^{-1}X' - X(X'X)^{-1}X' + \\
&\quad X \underbrace{(X'X)^{-1}X'X}_{=I_K} (X'X)^{-1}X' = \\
&= I_N - X(X'X)^{-1}X' = M
\end{aligned}$$

- $\text{rg}(M) = \text{tr}(M) = N - K$

$$\begin{aligned} \text{tr}(M) &= \text{tr}[I_N - X(X'X)^{-1}X'] \\ &= \text{tr}(I_N) - \text{tr}[X(X'X)^{-1}X'] \\ &= \text{tr}(I_N) - \text{tr}[(X'X)^{-1}X'X] \\ &= \text{tr}(I_N) - \text{tr}(I_K) \\ &= N - K \end{aligned}$$

- M es ortogonal a X:  $MX = 0$

$$\begin{aligned} MX &= [I_N - X(X'X)^{-1}X']X \\ &= X - X(X'X)^{-1}X'X = 0 \end{aligned}$$

## 2. Demostración: $E(\hat{\sigma}^2) = \sigma^2$

Como

$$\begin{aligned} E(\hat{u}'\hat{u}) &= E(u'Mu) = \\ &= E(\text{tr}(u'Mu)) = E(\text{tr}(Mu u')) = \\ &= \text{tr}(E(Mu u')) = \text{tr}(ME(uu')) = \\ &= \text{tr}(M\sigma^2 I_N) = \\ &= \sigma^2 \text{tr}(M) = \sigma^2(N - K) \end{aligned}$$

se puede demostrar fácilmente que  $\hat{\sigma}^2$  es un estimador insesgado:

$$E(\hat{\sigma}^2) = E\left(\frac{\hat{u}'\hat{u}}{N - K}\right) = \frac{E(\hat{u}'\hat{u})}{N - K} = \frac{\sigma^2(N - K)}{N - K} = \sigma^2$$

Y por tanto podremos utilizarlo como el estimador apropiado de la varianza de la perturbación.

### 3.5. Bibliografía del tema

#### Referencias bibliográficas básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con *gretl*:

- [1] Ramanathan, R. (2002), *Instructor's Manual to accompany*, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), *Student Solutions Manual*, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

#### Referencias Bibliográficas Complementarias:

- [1] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Análisis de regresión con gretl*. Open Course Ware. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-gretl/Courseisting>).
- [2] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación on-line de la Facultad de C.C. Económicas y Empresariales.
- [3] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones.
- [4] Esteban, MV (2008). *Estadística Actuarial: Regresión Lineal*, Sarriko On Line 3/08. Publicación on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.
- [5] Esteban, M.V. (2007). *Colección de ejercicios y exámenes*. Material docente. Servicio de Publicaciones.
- [6] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). *Ejercicios de Econometría*. Editorial McGraw-Hill.
- [7] Greene, W. (1998), *Análisis Econométrico*, Ed. Prentice Hall, 3ª edición.
- [8] Ramanathan, R. (2002), *Introductory Econometrics with applications*, Ed. South-Western, 5th. edition.
- [9] Verbeek, M. (2004). *A Guide to Modern Econometrics*. Wiley.

## Tema 4

# Modelo de Regresión Lineal General: Inferencia

En este tema desarrollaremos el contraste de restricciones lineales que recojan hipótesis relevantes desde el punto de vista económico dentro del Modelo de Regresión Lineal General. Aprenderemos a contrastar si las variables son relevantes individual y conjuntamente para explicar el comportamiento de la variable objetivo y a hacer contraste de combinaciones lineales, entre otros contrastes de interés.

Veremos que consecuencias tiene en las propiedades de los estimadores y en la inferencia la omisión de variables relevantes y la inclusión de variables irrelevantes. También analizaremos que problemas nos crea la existencia de combinaciones lineales exactas y/o aproximadas entre las variables a incluir como explicativas en el modelo. Una vez el modelo esté correctamente especificado para realizar inferencia podremos utilizarlo para predecir.

Para finalizar el tema veremos como realizar análisis de regresión y contraste de hipótesis mediante el software *gretl*.

### Competencias a trabajar en estas sesiones:

- C2. Aplicar la metodología econométrica básica para estimar y validar relaciones económicas en base a la información estadística disponible sobre las variables y utilizando los instrumentos informáticos apropiados.
- C3. Interpretar razonadamente los resultados obtenidos en la estimación y validación del modelo econométrico con el objetivo de elaborar informes económicos.
- C4. Presentar de forma clara y concisa, tanto oralmente como por escrito, las conclusiones obtenidas en una aplicación empírica.

### Al final de este tema deberíais ser capaces de:

- 1. Saber derivar la distribución del estimador de MCO (C2).
- 2. Saber contrastar la significatividad individual de las variables explicativas (C2 y C3).

3. Saber contrastar la significatividad conjunta de las variables explicativas (C2 y C3).
4. Saber contrastar restricciones lineales de parámetros (C2 y C3).
5. Saber contrastar restricciones múltiples (C2 y C3).
6. Predecir por punto y por intervalo el valor de la variable endógena dados los valores de las variables exógenas en el periodo de predicción (C2 y C3).
7. Utilizar un software econométrico (Gretl) para realizar contraste de hipótesis relevantes para la relación económica de las variables e interpretar sus resultados (C2 , C3 y C4).

### **Bibliografía Recomendada:**

Al final del tema tenéis recogida la bibliografía correspondiente. En particular se os recomienda leer los capítulos correspondientes a la bibliografía básica detallados a continuación:

- Gujarati, D. y Porter, D.C. (2010). En Parte I: Caps. 1, 2, 3, 4, 5, 7, 8 y 9. Parte II: Cap. 10.
- Ramanathan, R. (2002). Caps. 3, 4, 5 y 7.
- Wooldridge, J.M. (2006). Caps. 2, 3, 4 y 7

## 4.1. Distribución del estimador MCO. Estimación por intervalo

### 4.1.1. Distribución del estimador de MCO bajo Normalidad

Si  $Y = X\beta + u$ , donde  $u \sim N(0, \sigma^2 I_N)$ , el estimador MCO, dado que es lineal en las perturbaciones, también seguirá una distribución Normal Multivariante, con vector de medias  $E(\hat{\beta}) = \beta$  y matriz de varianzas y covarianzas  $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ . Es decir,

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Para el  $k$ -ésimo coeficiente,

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk})$$

donde  $a_{kk}$  es el elemento  $(k, k)$  de la matriz  $(X'X)^{-1}$ .

### 4.1.2. Estimación por intervalo

Para el  $k$ -ésimo coeficiente,

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk})$$

Una vez estimada la varianza de la perturbación con el estimador insesgado  $\hat{\sigma}^2$  se puede demostrar que:

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{a_{kk}}} \sim t_{(N-K)}$$

donde  $t_{(N-K)}$  denota la distribución t-Student con  $(N - K)$  grados de libertad, y  $\hat{\sigma} \sqrt{a_{kk}}$  es la desviación estimada del coeficiente estimado. (Notación  $\hat{\sigma} \sqrt{a_{kk}} = \hat{\sigma}_{\hat{\beta}_k}$ ).

El intervalo de confianza asociado es:

$$Pr \left[ \hat{\beta}_k - t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} < \beta_k < \hat{\beta}_k + t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} \right] = 1 - \alpha$$

Con lo que podemos escribir el intervalo de confianza del  $(1 - \alpha)$  por ciento para un coeficiente cualquiera  $\beta_k$  como:

$$IC(\beta_k)_{1-\alpha} = \left( \hat{\beta}_k \pm t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} \right)$$

Este es un *estimador por intervalo* porque en los extremos inferior y superior del intervalo aparecen  $\hat{\beta}_k$  y  $\hat{\sigma}_{\hat{\beta}_k}$ , que son estimadores. Este intervalo es aleatorio, porque para cada muestra se obtiene un valor numérico distinto de  $\hat{\beta}_k$  y  $\hat{\sigma}_{\hat{\beta}_k}$ . Cuando usamos una muestra para obtener las estimaciones, tendremos [un número  $\leq \beta_k \leq$  otro número] y se denomina *estimación por intervalo* de  $\beta_k$  ó intervalo de confianza  $(1 - \alpha)$  para  $\beta_k$ . Un intervalo de confianza nos dice que, con probabilidad  $(1 - \alpha)$  se estima que el parámetro  $\beta_k$  estará dentro de ese rango de valores.

Las propiedades de la variable aleatoria  $IC(\beta_k)$  se basan en la noción del muestreo repetido: si obtuviéramos infinitas muestras de tamaño  $N$  de una misma población, y para cada una de ellas

construyésemos el intervalo, entonces  $(1 - \alpha) \times 100\%$  de todos los intervalos construidos contendrían el verdadero valor (desconocido) de  $\beta_k$ .

¿Para qué sirven las estimaciones por intervalo? La respuesta es que nos dan una información muy valiosa sobre la precisión de las estimaciones por punto, esto es, nos dicen hasta qué punto nos podemos fiar de ellas. Si un intervalo de confianza es ancho (debido a una  $\widehat{V}(\hat{\beta}_k)$  grande) nos está diciendo que no hay mucha información en la muestra sobre  $\beta_k$ . Además, como veremos más adelante, los intervalos sirven para realizar contraste de hipótesis.

### Ejemplo 4.1

Para los valores estimados de la ecuación (2.4) obtenemos los siguientes intervalos de estimación:

- Para la variable precio,  $P$ :

$$Pr \left[ \hat{\beta}_2 - t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_2} < \beta_2 < \hat{\beta}_2 + t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_2} \right] = 1 - \alpha$$

Con lo que podemos escribir el intervalo de confianza del  $(1 - \alpha)$  por ciento para el coeficiente  $\beta_2$  como:

$$IC(\beta_2)_{1-\alpha} = \left( \hat{\beta}_2 \pm 1,993 \hat{\sigma}_{\hat{\beta}_2} \right) = (-7,908 \pm 1,993 \times 1,096) = [-10,092, -5,724]$$

Estimamos que una reducción de 1\$ lleva a un incremento en los ingresos por ventas de entre 5.724\$ y 10.092\$.

- Para la variable gasto en publicidad,  $A$ :

$$Pr \left[ \hat{\beta}_3 - t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_3} < \beta_3 < \hat{\beta}_3 + t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_3} \right] = 1 - \alpha$$

Con lo que podemos escribir el intervalo de confianza del  $(1 - \alpha)$  por ciento para el coeficiente  $\beta_3$  como:

$$IC(\beta_3)_{1-\alpha} = \left( \hat{\beta}_3 \pm 1,993 \hat{\sigma}_{\hat{\beta}_3} \right) = (1,8626 \pm 1,993 \times 0,6832) = [0,501, 3,224]$$

Estimamos que un incremento de 1000\$ en el gasto en publicidad lleva a un incremento en los ingresos por ventas de entre 501\$ y 3.224\$.

## 4.2. Contraste de hipótesis sobre los coeficientes de la regresión

Un problema fundamental de la Econometría es aportar un conocimiento descriptivo de una economía real, los economistas desarrollan teorías sobre el comportamiento económico y las evalúan. Los contrastes de hipótesis son los procedimientos que se usan para evaluar estas teorías. Para ello

vamos a utilizar el modelo  $Y = X\beta + u$  donde consideramos que se cumplen las hipótesis básicas y además la perturbación es normal. La normalidad no es necesaria para estimar por MCO ni para determinar las propiedades del estimador pero si lo es para realizar inferencia dado que al ser  $\hat{\beta}_{MCO}$  lineal en  $u$  tendrá su misma distribución y podremos derivar estadísticos de contraste basándonos en ella.

Por ejemplo, dado que

$$u_i \sim N(0, \sigma^2) \longrightarrow \hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk})$$

si conocemos todos los elementos incluido  $\sigma^2$  podríamos contrastar hipótesis de la forma  $H_0 : \beta_k = c$  con el siguiente estadístico:

$$\frac{\hat{\beta}_k - c}{\sigma \sqrt{a_{kk}}} \stackrel{H_0}{\sim} N(0, 1)$$

En general nosotros lo que queremos es contrastar conjuntos lineales de hipótesis. Podemos realizar contrastes sobre los coeficientes individuales y sobre conjuntos de coeficientes, incluso sobre todos los coeficientes a la vez. Los contrastes más importantes en Econometría son los contrastes de significatividad de los regresores individuales y el contraste de significatividad conjunta. En ellos tratamos de analizar si cada uno de los regresores del modelo de forma individual o conjuntamente son útiles para explicar el comportamiento de la variable endógena. Los veremos a continuación junto con otros de interés.

#### 4.2.1. Contraste de restricciones sobre los coeficientes de regresión individuales

En los contrastes sobre los coeficientes individuales se contrasta la hipótesis nula  $H_0 : \beta_k = c$ , donde la constante  $c$  puede tomar diversos valores. Contrastamos una única restricción. La hipótesis alternativa puede ser a una cola por ejemplo  $H_a : \beta_k > 0$  o a dos colas  $H_a : \beta_k \neq c$ . Para realizar el contraste hemos de derivar el estadístico de contraste y su distribución bajo la hipótesis nula, evaluar el estadístico en la muestra y aplicar la regla de decisión<sup>1</sup>. Para contrastar:

$$H_0 : \beta_k = c \quad \text{frente a} \quad H_a : \beta_k \neq c$$

Bajo las hipótesis básicas y normalidad de las perturbaciones la distribución del estimador  $\hat{\beta}_k$  es la siguiente:

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk})$$

Si  $\sigma^2$  es conocida todo es conocido en la distribución de  $\beta_k$  y el estadístico de contraste sería:

$$\frac{\hat{\beta}_k - c}{\sigma \hat{\beta}_k} \stackrel{H_0}{\sim} N(0, 1)$$

En el resto de ejemplos consideramos el caso más habitual  $\sigma^2$  desconocida, para el cual podemos derivar el siguiente estadístico de contraste<sup>2</sup> y distribución asociada cuando  $\sigma^2$  es estimada con el

<sup>1</sup>Ver Tema ?? para recordar el mecanismo de contraste.

<sup>2</sup>Si  $\sigma^2$  es desconocida habría de ser estimada, bajo la normalidad de las perturbaciones

$$u_i \sim N(0, \sigma^2) \longrightarrow \frac{(N - K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$$

estimador insesgado  $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$ :

$$\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} \stackrel{H_0}{\sim} t_{(N-K)}$$

La regla de decisión es rechazar  $H_0$  si  $\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} > t_{(N-K)|\frac{\alpha}{2}}$ . En este caso contrario no se rechaza.

Si la alternativa es a una cola, por ejemplo:

$$H_0 : \beta_k = c \quad \text{frente a} \quad H_a : \beta_k > c$$

La regla de decisión es rechazar  $H_0$  si  $\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} > t_{(N-K)|\alpha}$ .

### Contraste de significatividad individual

Cuando  $c = 0$  al contraste se le denomina de significatividad individual. En este caso:

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

Podemos derivar el siguiente estadístico de contraste y distribución:

$$\frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \stackrel{H_0}{\sim} t_{(N-K)}$$

Si el estadístico calculado para la muestra es mayor que el estadístico en tablas,  $\frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} > t_{(N-K)|\frac{\alpha}{2}}$  para un  $\alpha$  dado, se rechaza la hipótesis nula. En este caso  $\beta_k \neq 0$  y la variable explicativa asociada  $X_k$  es significativa para explicar el comportamiento de la variable endógena. Por tanto este contraste sirve para decidir si la variable  $X_k$  debe mantenerse en el modelo. Si el estadístico calculado para la muestra es menor que el estadístico en tablas,  $\frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} < t_{(N-K)|\frac{\alpha}{2}}$  para un  $\alpha$  dado, no se rechaza la hipótesis nula. En este caso  $\beta_k = 0$  y la variable explicativa asociada  $X_k$  no es significativa para explicar el comportamiento de la variable endógena.

**Utilización del intervalo de confianza para hacer contraste de hipótesis** En secciones anteriores hablamos de la estimación por intervalo y se mencionó que también podíamos realizar inferencia utilizando intervalos de confianza. Pues bien si recordamos el intervalo de confianza asociado a  $\beta_k$ :

y derivar el correspondiente estadístico de contraste, que sería:

$$\frac{\frac{\hat{\beta}_k - c}{\sigma\sqrt{a_{kk}}}}{\sqrt{\frac{\sum \hat{u}_i^2 / \sigma^2}{N-K}}} \stackrel{H_0}{\sim} t_{(N-K)} \quad \text{si simplificamos} \quad \frac{\hat{\beta}_k - c}{\hat{\sigma}\sqrt{a_{kk}}} \stackrel{H_0}{\sim} t_{(N-K)}$$

$$Pr \left[ \hat{\beta}_k - t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} < \beta_k < \hat{\beta}_k + t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} \right] = 1 - \alpha$$

$$IC(\beta_k)_{1-\alpha} : \left( \hat{\beta}_k \pm t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{\hat{\beta}_k} \right)$$

y la regla de decisión es que si la constante  $c$  (en este caso  $c = 0$ ) pertenece al intervalo, no rechazamos  $H_0$  con un nivel de significación  $\alpha$  y si no pertenece al intervalo, rechazamos  $H_0$  con un nivel de significación  $\alpha$ . Claramente se obtienen exactamente los mismos resultados utilizando los estadísticos de contraste individuales que utilizando los intervalos de confianza.

#### 4.2.2. Contraste de restricciones sobre los coeficientes de regresión

En ocasiones interesa averiguar cuál es el efecto de la combinación de varias variables, por ejemplo nos interesará saber si la combinación de todas las variables es un útil predictor de la variable dependiente.

##### Contraste de significatividad conjunto

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$$

$$H_a : \text{alguna igualdad no se da}$$

En este caso podemos derivar el siguiente estadístico de contraste y distribución asociada:

$$\frac{R^2/K - 1}{1 - R^2/N - K} \stackrel{H_0}{\sim} \mathcal{F}_{(K-1, N-K)}$$

Si  $\frac{R^2/K - 1}{1 - R^2/N - K} > \mathcal{F}_{(q, N-K)|\alpha}$  el estadístico calculado para la muestra es mayor que el estadístico en tablas, para un  $\alpha$  dado, se rechaza la hipótesis nula y se concluye que las variables son conjuntamente significativas para explicar el comportamiento de la variable endógena.

#### Ejemplo 4.2

Vamos a mostrar un ejemplo sobre los contrastes de significatividad individual y conjunto con los resultados de la estimación del modelo (2.4). La ecuación de regresión muestral era:

$$\begin{array}{r} \hat{S}_i \\ (\hat{\sigma}_{\hat{\beta}_k}) \end{array} = \begin{array}{r} 118,914 \\ (6,35164) \end{array} - \begin{array}{r} 7,90785 \\ (1,09599) \end{array} P_i + \begin{array}{r} 1,86258 \\ (0,683195) \end{array} A_i \quad i = 1, \dots, 75$$

$$N = 75 \quad R^2 = 0,448258 \quad \bar{R}^2 = 0,432932$$

Como puede apreciarse en la ecuación anterior, se indica que bajo cada coeficiente estimado aparece su correspondiente desviación típica estimada<sup>3</sup>.

<sup>3</sup>Una alternativa a presentar las desviaciones típicas estimadas de los coeficientes es presentar el valor muestral del estadístico de significatividad individual para el coeficiente de regresión correspondiente o los valores  $p$ .

Contrastes de significatividad individual, contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_k = 0 \\ H_a : \beta_k \neq 0 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \stackrel{H_0}{\sim} t_{(75-3)}$$

- Para la variable  $P$  obtenemos:

$$\left| \frac{-7,90785}{1,09599} \right| = |-7,215| > 1,99346 = t_{(72)|0,025}$$

luego rechazamos  $H_0$  para  $\alpha = 5\%$  y la variable  $P$  es significativa.

- Para la variable  $A$  obtenemos:

$$\frac{1,86258}{0,683195} = 2,726 > 1,99346 = t_{(72)|0,025}$$

luego rechazamos  $H_0$  para  $\alpha = 5\%$  y la variable  $A$  es significativa.

En el contraste de significatividad conjunta, contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_2 = \beta_3 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \end{array} \right\} \text{ con } \frac{R^2/K - 1}{1 - R^2/N - K} \stackrel{H_0}{\sim} \mathcal{F}_{(K-1, N-K)}$$

Evaluated el estadístico en la muestra obtenemos:

$$\frac{0,448258/2}{(1 - 0,448258)/72} = 29,24786 > 3,12391 = \mathcal{F}_{(2,72)|0,05}$$

rechazamos  $H_0$  para  $\alpha = 5\%$ . Concluimos que las variables exógenas  $P$  y  $A$  son conjuntamente significativas.

Además hay otras hipótesis de interés:

- ¿Es la demanda inelástica o elástica con respecto al precio? En este caso queremos saber si:
  - $\beta_2 \geq 0$ , una reducción en el precio conlleva un decrecimiento en los ingresos por ventas, la demanda es inelástica con respecto al precio.
  - $\beta_2 < 0$ , una reducción en el precio conlleva un crecimiento en los ingresos por ventas, la demanda es elástica con respecto al precio.

En general estaremos dispuestos a aceptar que la demanda es elástica cuando existe una fuerte evidencia en los datos para soportar esta hipótesis. Luego lo mejor es que contratemos como hipótesis nula que la demanda es inelástica:

$$\begin{array}{l} H_0 : \beta_2 \geq 0, \text{ la demanda es inelástica} \\ H_a : \beta_2 < 0, \text{ la demanda es elástica} \end{array}$$

En la práctica contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 < 0 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \stackrel{H_0}{\sim} t_{(75-3)}$$

Si rechazamos  $H_0$  para  $\beta_2 = 0$ , además lo rechazamos para  $\beta_2 > 0$ , por lo que asumimos que  $\beta_2 = 0$  es cierto. El estadístico evaluado en la muestra es<sup>4</sup>:

$$\frac{-7,908}{1,096} = -7,215 < -1,666 = t_{(72)}|_{0,05}$$

luego rechazamos  $H_0$  para  $\alpha = 5\%$  y concluimos que la demanda es elástica,  $\beta_2 < 0$ . La evidencia muestral soporta que una reducción en el precio conllevará un incremento en los ingresos por ventas.

- ¿Es efectiva la política de gasto en publicidad? Una hipótesis de interés es si un incremento en el gasto en publicidad conllevará un incremento en los ingresos por ventas que cubra el incremento en el gasto en publicidad:

$$H_0 : \beta_3 \leq 1 \quad H_1 : \beta_3 > 1$$

Luego contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_3 = 1 \\ H_a : \beta_3 > 1 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_3 - 1}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t_{(75-3)}$$

El estadístico evaluado en la muestra es:

$$\frac{1,8626 - 1}{0,6832} = 1,263 < 1,666 = t_{(72)}|_{0,05}$$

luego no rechazamos  $H_0$  para  $\alpha = 5\%$  y  $\beta_3 = 1$ . En nuestra muestra no hay suficiente evidencia para concluir que la publicidad será efectiva.

### Ejemplo 4.3

Vamos a mostrar un ejemplo sobre los contrastes de significatividad individual y conjunto con los resultados de la estimación del modelo (3.5). Primero vamos a escribir los resultados de la estimación de la forma habitual en que se muestran en la literatura:

$$\widehat{PRICE} = 129,062 + 0,154800 SQFT - 21,5875 BEDRMS - 12,1928 BATHS$$

$$\begin{array}{ccccccc} (\hat{\sigma}_{\hat{\beta}_k}) & (88,30) & (0,03) & (27,02) & & (43,25) & \\ N = 14 & R^2 = 0,8359 & \bar{R}^2 = 0,7868 & & & & \end{array}$$

Contrastes de significatividad individual, contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_k = 0 \\ H_a : \beta_k \neq 0 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(14-4)}$$

---

<sup>4</sup>También podemos tomar el estadístico en valor absoluto  $\left| \frac{-7,908}{1,096} \right| = |-7,215| = 7,215 > 1,666 = t_{(72)}|_{0,05}$  luego rechazamos  $H_0$  para  $\alpha = 5\%$ .

- Para la variable  $SQFT$  obtenemos:

$$\frac{0,1548}{0,0319} = 4,8465 > 2,22814 = t_{(10)|0,025}$$

luego rechazamos  $H_0$  para  $\alpha = 5\%$  y la variable  $SQFT$  es significativa.

- Para la variable  $BEDRMS$  obtenemos:

$$\left| \frac{-21,587}{27,0293} \right| = |-0,7987| < 2,22814 = t_{(10)|0,025}$$

luego no rechazamos  $H_0$  para  $\alpha = 5\%$  y la variable  $BEDRMS$  no es significativa.

- Para la variable  $BATHS$  obtenemos:

$$\left| \frac{-12,192}{43,25} \right| = |-0,2819| < 2,22814 = t_{(10)|0,025}$$

luego no rechazamos  $H_0$  para  $\alpha = 5\%$  y la variable  $BATHS$  no es significativa.

En el contraste de significatividad conjunta, contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \text{alguna igualdad no se da} \end{array} \right\} \text{ con } \frac{R^2/K - 1}{1 - R^2/N - K} \stackrel{H_0}{\sim} \mathcal{F}_{(K-1, N-K)}$$

Evaluado el estadístico en la muestra obtenemos:

$$\frac{0,8359/3}{(1 - 0,8359)/10} = 16,989 > 3,70826 = \mathcal{F}_{(3,10)|0,05}$$

rechazamos  $H_0$  para  $\alpha = 5\%$ . Concluimos que las variables exógenas  $SQFT$ ,  $BEDRMS$  y  $BATHS$  son conjuntamente significativas.

#### Ejemplo 4.4

Utilizamos la función de salarios especificada para el año 2002 que se propuso en el Ejemplo 2.7:

$$W_i = \beta_1 + \beta_2 S_{2i} + \beta_3 X_i + u_i \quad i = 1, 2, \dots, N$$

donde  $W_i$  es el salario anual del individuo  $i$ ,  $X_i$  son los años de experiencia del individuo  $i$  y  $S_{2i}$  es una variable ficticia que se define:

$$S_{2i} = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

En este modelo podemos contrastar:

- Si la experiencia es determinante del salario:  $H_0 : \beta_3 = 0$ , si esta hipótesis no se rechaza para un nivel de significatividad dado el salario no depende de los años de experiencia del individuo. Contrastamos:

$$\left. \begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} \stackrel{H_0}{\sim} t_{(N-3)}$$

- Si existe discriminación salarial por sexo:  $H_0 : \beta_2 = 0$ , si esta hipótesis no se rechaza para un nivel de significatividad dado no existe discriminación salarial por sexo. Por ejemplo si la experiencia es cero y  $\beta_2 = 0$ , el salario esperado es  $\beta_1 \forall i$  luego el salario esperado es el mismo para hombres y mujeres.

$$\left. \begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array} \right\} \text{ con el estadístico y distribución } \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \stackrel{H_0}{\sim} t_{(N-3)}$$

### Contraste de combinaciones lineales

Por ejemplo contrastamos la hipótesis:

$$H_0 : \beta_2 + \beta_3 = 1$$

$$H_a : \beta_2 + \beta_3 \neq 1$$

Renombrando  $\hat{w} = \hat{\beta}_2 + \hat{\beta}_3$  y  $c = 1$  se puede expresar la hipótesis nula y alternativa así como el estadístico de contraste y su distribución asociada como:

$$H_0 : w = c \quad H_a : w \neq c$$

$$\frac{\hat{w} - c}{\hat{\sigma}_{\hat{w}}} \stackrel{H_0}{\sim} t_{(N-K)} \quad \text{si } H_0 \text{ es cierta}$$

La distribución del estadístico  $\hat{w} \sim N(\mu_w, \sigma_w^2)$  dado que:

$$\hat{w} = \hat{\beta}_2 + \hat{\beta}_3$$

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2 a_{22})$$

$$\hat{\beta}_3 \sim N(\beta_3, \sigma^2 a_{33})$$

es

$$\mu_w = E(\hat{w}) = E(\hat{\beta}_2 + \hat{\beta}_3) = \beta_2 + \beta_3$$

$$\begin{aligned} \sigma_w^2 &= V(\hat{w}) = E[\hat{w} - E(\hat{w})]^2 = E[(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)]^2 = V(\hat{\beta}_2) + V(\hat{\beta}_3) + 2Cov(\hat{\beta}_2, \hat{\beta}_3) \\ &= \sigma^2(a_{22} + a_{33} + 2a_{23}) \end{aligned}$$

Por tanto

$$\hat{\beta}_2 + \hat{\beta}_3 \sim N(\beta_2 + \beta_3, \sigma^2(a_{22} + a_{33} + 2a_{23}))$$

Luego en términos de los coeficientes estimados originales el estadístico de contraste y distribución es:

$$\frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\sqrt{\hat{V}(\hat{\beta}_2) + \hat{V}(\hat{\beta}_3) + 2\hat{Cov}(\hat{\beta}_2, \hat{\beta}_3)}} \stackrel{H_0}{\sim} t_{(N-K)}$$

o lo que es igual:

$$\frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\hat{\sigma}\sqrt{a_{22} + a_{33} + 2a_{23}}} \stackrel{H_0}{\sim} t_{(N-K)}$$

Con la regla de decisión habitual.

### Ejemplo 4.5

Para contrastar:

$$H_0 : \beta_2 = \beta_3 \quad H_a : \beta_2 \neq \beta_3$$

es equivalente a escribir:

$$H_0 : \beta_2 - \beta_3 = 0 \quad H_a : \beta_2 - \beta_3 \neq 0$$

que podemos contrastar con el estadístico y distribución:

$$\frac{\hat{\beta}_2 - \hat{\beta}_3}{\hat{\sigma}\sqrt{a_{22} + a_{33} - 2a_{23}}} \stackrel{H_0}{\sim} t_{(N-K)}$$

Con la regla de decisión habitual.

### Contraste de un subconjunto de coeficientes

Supongamos el siguiente modelo de regresión:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \dots + \alpha_r Z_{ri} + u_i \quad i = 1, 2, \dots, N$$

y queremos contrastar si el subconjunto de regresores  $Z_{1i}, Z_{2i}, \dots, Z_{ri}$  son conjuntamente significativos para explicar el comportamiento de la variable endógena. La hipótesis de contraste es:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$$

$$H_a : \text{alguna igualdad no se de}$$

El estadístico de contraste y distribución son:

$$\frac{\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / r}{\hat{u}' \hat{u} / (N - K)} \stackrel{H_0}{\sim} \mathcal{F}_{(r, N-K)} \quad (4.1)$$

donde:

- $\hat{u}'_r \hat{u}_r$  es la suma de cuadrados residual del modelo restringido estimado por MCO, siendo el modelo restringido aquel que cumple la hipótesis nula. Luego el modelo restringido es:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_{ri} \quad i = 1, 2, \dots, N$$

- $\hat{u}' \hat{u}$  es la suma de cuadrados residual del modelo no restringido o lo que es igual el modelo de interés estimado por MCO:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \dots + \alpha_r Z_{ri} + u_i \quad i = 1, 2, \dots, N$$

- $r$  es el número de restricciones que se contrastan, en este caso el número de coeficientes  $\alpha_r$ .

La regla de decisión es la habitual, se rechaza la hipótesis nula si:

$$\frac{\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / r}{\hat{u}' \hat{u} / (N - K)} > \mathcal{F}_{(r, N-K) | \alpha}$$

en cuyo caso las variables exógenas  $Z_{ri}$  contribuyen a explicar el comportamiento de la variable endógena, en cuyo caso debemos especificar el modelo no restringido. Si  $\frac{\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / r}{\hat{u}' \hat{u} / (N - K)} < \mathcal{F}_{(r, N-K) | \alpha}$  no rechazamos  $H_0$  en cuyo caso las variables  $Z_{ri}$  no contribuyen a explicar a la variable endógena y debemos especificar el modelo restringido.

### 4.2.3. Contrastes basados en sumas de cuadrados de residuos

Al estadístico:

$$\frac{\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / q}{\hat{u}' \hat{u} / (N - K)} \stackrel{H_0}{\sim} \mathcal{F}_{(q, N-K)}$$

generalmente se le conoce con el nombre de **estadístico de diferencias en las sumas residuales de cuadrados**. Puede ser utilizado para contrastar hipótesis lineales con solo especificar correctamente los modelos restringido y no restringido,  $q$  es el número de restricciones que se contrastan.

Para su aplicación sólo es necesario obtener la SCR del modelo restringido y no restringido. Vamos a estudiarlo en detalle en el ejemplo siguiente.

**Estimación del modelo restringido y cálculo de la  $SCR_r$**  El modelo restringido (MR) es aquel que cumple la hipótesis nula mientras que el modelo no restringido (MNR) es el modelo de interés. Por ejemplo sea el MRLG,

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

donde queremos contrastar la hipótesis nula  $H_0 : \beta_2 + \beta_3 = 1$  sustituyendo la restricción en el modelo encontramos el modelo restringido:

$$\text{MR: } Y_i = \beta_1 + \beta_2 X_{2i} + (1 - \beta_2) X_{3i} + u_{ri}$$

$$\underbrace{Y_i - X_{3i}}_{=Y_i^*} = \beta_1^r + \beta_2^r \underbrace{(X_{2i} - X_{3i})}_{=X_i^*} + u_{ri}$$

$$Y_i^* = \beta_1^r + \beta_2^r X_i^* + u_{ri}$$

La aplicación de MCO en el modelo resultante son los llamados estimadores de Mínimos Cuadrados Restringidos, MCR. Los demás  $\hat{\beta}^r$  se obtienen con las restricciones. En el ejemplo en el modelo restringido se calculan  $\hat{\beta}_1^r$  y  $\hat{\beta}_2^r$  y finalmente se calcula  $\hat{\beta}_3^r = 1 - \hat{\beta}_2^r$ .

En este modelo restringido estimado por MCO se calcula la  $SCR = \hat{u}'_r \hat{u}_r$ . Si escribimos el MR en términos matriciales

$$Y^* = X^* \beta^r + u_r$$

entonces

$$\hat{u}'_r \hat{u}_r = Y^{*'} Y^* - \hat{\beta}^{r'} X^{*'} Y^*$$

donde  $Y^*$  y  $X^*$  son las variables que quedan en el modelo restringido y

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_1^r \\ \hat{\beta}_2^r \end{bmatrix} &= \begin{bmatrix} N & \sum X_i^* \\ \sum X_i^* & \sum X_i^{*2} \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i^* \\ \sum Y_i^* X_i^* \end{bmatrix} \\ &= \begin{bmatrix} N & \sum (X_{2i} - X_{3i}) \\ \sum (X_{2i} - X_{3i}) & \sum (X_{2i} - X_{3i})^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum (Y_i - X_{3i}) \\ \sum (Y_i - X_{3i})(X_{2i} - X_{3i}) \end{bmatrix} \end{aligned}$$

### Ejemplo 4.6

Contraste de significatividad conjunta:  $H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$ . Para esta hipótesis el modelo restringido es

$$Y_i = \beta_1 + u_i$$

si estimamos el MR por MCO obtenemos:

$$\begin{aligned} \text{Min}_{\hat{\beta}_1} \sum \hat{u}_i^2 &= \text{Min}_{\hat{\beta}_1} \sum (Y_i - \hat{\beta}_1)^2 \\ \frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} &= -2 \sum (Y_i - \hat{\beta}_1) = 0 \longrightarrow \hat{\beta}_1^r = \bar{Y} \end{aligned}$$

de donde

$$\begin{aligned} \hat{u}'_r \hat{u}_r &= \sum (Y_i - \hat{Y}_i)^2 = \\ &= \sum (Y_i - \hat{\beta}_1^r)^2 = \sum (Y_i - \bar{Y})^2 = SCT \end{aligned}$$

Así

$$\frac{\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / q}{\hat{u}' \hat{u} / (N - K)} = \frac{(SCT - SCR) / q}{SCR / N - K}$$

dividiendo el numerador y el denominador de entre  $SCT$  obtenemos.

$$F = \frac{(\hat{u}'_r \hat{u}_r - \hat{u}' \hat{u} / q)}{\hat{u}' \hat{u} / N - K} = \frac{R^2 / K - 1}{(1 - R^2) / N - K} \stackrel{H_0}{\sim} \mathcal{F}_{(K-1, N-K)}$$

estadístico que coincide con el obtenido para el contraste de significatividad conjunta.

### Ejercicio 4.1

Sean las variables  $X_t$ , coste medio de las reclamaciones por siniestros realizados en  $t$  y  $Y_t$  los beneficios de la empresa en  $t$ , ambos en miles de millones. Disponemos de la siguiente muestra:

X	2	3	1	5	9
Y	4	7	3	9	17

Para el modelo:  $Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1, \dots, 5 \quad u_t \sim N(0, \sigma^2)$

1. Interpreta los coeficientes  $\beta_1$  y  $\beta_2$ .
2. Escribe la matriz de regresores y el vector de valores de la variable endógena.
3. Calcula la recta de regresión muestral MCO.
4. Interpreta las estimaciones de los parámetros.
5. Calcular los valores de  $\hat{Y}_t$  y  $\hat{u}_t$ .
6. Estima la varianza de la perturbación.
7. Calcula el coeficiente de determinación,  $R^2$  e interpreta su significado.
8. Calcula el coeficiente de correlación simple, ¿qué relación hay entre  $r_{XY}$  y  $R^2$  en este modelo?
9. Contrasta la significatividad de la variable coste medio de las reclamaciones.
10. Considérense los siguientes dos estimadores de la pendiente de la regresión:

$$\hat{\beta}_{2,MCO} \quad \tilde{\beta}_2 = \frac{Y_5 - Y_3}{X_5 - X_3}$$

- a) Calcular los dos estimadores.
- b) Razonar cuál de los dos estimadores es “mejor”.

### Ejercicio 4.2

Supongamos que la variable  $Y$  se relaciona con las variables  $X_1, X_2, X_3$  mediante la ecuación:

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{4.2}$$

Donde  $u_t \sim iid(0, \sigma^2 I_T)$ , los siguientes datos vienen dados por la tabla:

$X_1$	1	1	1	1	1	1	1	1	1
$X_2$	1	0	-1	1	0	-1	1	0	-1
$X_3$	1	1	1	0	0	0	-1	-1	-1
Y	3	6	10	5	10	12	5	10	10

1. Interpreta los parámetros del modelo.

2. Escribe la matriz de regresores.
3. Estima la función de regresión muestral.
4. Estima la varianza de las perturbaciones.
5. Estima la matriz de varianzas y covarianzas de los parámetros estimados.
6. Calcula e interpreta el  $R^2$ .
7. ¿Se cumple en este modelo la igualdad  $r_{XY}^2 = R^2$ ?
8. Calcular el intervalo de confianza para el parámetro  $\beta_2$  y  $\alpha = 5\%$ .
9. Contrasta la significatividad individual y conjunta de los regresores.

### Ejercicio 4.3

Se ha estimado el siguiente modelo de regresión simple, utilizando el criterio de MCO:

$$\hat{V}_t = 1767,61 + 7,48W_t \quad R^2 = 0,80 \quad (4.3)$$

Los estadísticos t para el contraste de que el parámetro correspondiente es igual a cero resultan ser iguales a 1,4 para el término constante y 18,5 para la variable  $W$ . Dado que el valor del estadístico para el contraste de significación del término constante nos lleva a no rechazar la hipótesis nula, se formula y estima el siguiente modelo:

$$\hat{V}_t = 11,71W_t \quad R^2 = 0,91 \quad (4.4)$$

Como el mayor coeficiente de determinación corresponde a la segunda regresión, se opta por elegir el modelo (4.4). ¿Te parece adecuada esta decisión?

### 4.3. Consecuencias del incumplimiento de algunos supuestos: colinealidad

A la hora de estimar un modelo económico, los datos disponibles sobre las variables explicativas o regresores pueden presentar un alto grado de correlación, especialmente en un contexto de series temporales y con series macroeconómicas.

Cuando dos o más variables explicativas de un modelo están altamente correlacionadas en la muestra, es muy difícil separar el efecto parcial de cada una de estas variables sobre la variable dependiente. La información muestral que incorpora una de estas variables es casi la misma que el resto de las correlacionadas con ella. En este tema analizaremos las implicaciones que este fenómeno muestral tiene en la estimación por el método de Mínimos Cuadrados Ordinarios.

- El problema de multicolinealidad es un problema relacionado con la matriz de variables exógenas  $X$ .

- Se refiere no tanto a si existe o no relación lineal entre las variables exógenas del modelo de regresión, que existirá, como al grado de correlación lineal entre las variables explicativas del modelo de regresión lineal.
- En todo momento nosotros vamos a suponer que tenemos un modelo correctamente especificado y que al estimarlo detectamos los problemas en la matriz de datos  $X$ . Así, estamos enfocando el problema como un problema muestral.
- Podemos distinguir dos casos:
  - Multicolinealidad exacta: se produce cuando existe una relación lineal exacta.
  - Alta colinealidad: cuando la correlación entre las variables exógenas es muy alta pero no exacta.

### 4.3.1. Multicolinealidad exacta

Para verlo más claramente vamos a seguir un ejemplo. Sea el modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, \dots, N \quad (4.5)$$

y supongamos que  $X_{3i} = 2X_{2i}$ . Las ecuaciones normales que se obtienen del criterio de estimación MCO forman un sistema de tres ecuaciones pero solo dos son linealmente independientes:

$$\begin{aligned} \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} \\ \sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} \\ \sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \end{aligned}$$

ya que si sustituimos en estas ecuaciones la relación lineal exacta  $X_{3i} = 2X_{2i}$  y reorganizamos, obtenemos:

$$\begin{aligned} \sum Y_i &= N\hat{\beta}_1 + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i} \\ \sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i}^2 \\ 2(\sum Y_i X_{2i}) &= 2\left(\hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i}^2\right) \end{aligned}$$

Se puede observar que la tercera ecuación es la misma que la segunda excepto por un factor de escala igual a 2. Por lo tanto, hay tres incógnitas  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  y  $\hat{\beta}_3$  pero solamente dos ecuaciones linealmente independientes. Dado que  $X_{3i}$  y  $X_{2i}$  son combinación lineal exacta  $rg(X) = K - 1 = 3 - 1 = 2$ , luego  $X$  no es de rango completo y no se cumple una de las hipótesis básicas, la hipótesis de No Multicolinealidad. Consecuentemente, no es posible estimar de forma única todos los coeficientes del modelo. Ahora bien, las dos primeras ecuaciones si podemos resolverlas para  $\hat{\beta}_1$  y la combinación lineal  $(\hat{\beta}_2 + 2\hat{\beta}_3)$ .

Esto mismo se puede comprobar substituyendo  $X_{3i} = 2X_{2i}$  en el modelo (4.5).

$$Y_i = \beta_1 + (\beta_2 + 2\beta_3)X_{2i} + u_i \quad i = 1, 2, \dots, N \quad (4.6)$$

donde podemos estimar de forma separada y única el coeficiente  $\beta_1$  y la combinación lineal  $(\hat{\beta}_2 + 2\hat{\beta}_3)$  pero no cada uno de sus parámetros de forma individual. Además **no** importa la solución arbitraria de las ecuaciones normales, esta combinación lineal tiene siempre un único valor y siempre el mismo.

• Consecuencias de la multicolinealidad exacta:

- Los efectos directos de la correlación exacta entre regresores es que el valor del determinante  $|X'X| = 0$ , por tanto no podemos encontrar  $(X'X)^{-1}$  y por tanto, no podemos estimar el modelo por MCO ya que el estimador se define como  $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$ .
- En este caso lo que ocurre es que tenemos combinaciones lineales en las columnas de la matriz  $X$  con lo que  $rg(X) \neq K$  por lo que  $(X'X)$  es una matriz singular.
- Relajamos la hipótesis básica:

$$rg(X) \neq K \quad \text{tal que} \quad rg(X) \neq K \Rightarrow |X'X| = 0 \Rightarrow \nexists (X'X)^{-1}$$

- Cuando la correlación entre regresores es perfecta el problema de multicolinealidad exacta se convierte en un problema de especificación ya que **no podemos estimar todos los parámetros del modelo de forma individual**. Podremos estimar:
  - individualmente: aquellos parámetros cuyas variables exógenas no están afectadas de correlación exacta con otras variables exógenas del modelo y
  - combinaciones lineales de los parámetros cuyas variables exógenas están implicadas en las relaciones lineales exactas.
- Detección: basta con ver que  $|X'X| = 0$ .

### 4.3.2. Alta colinealidad

En este caso el valor del  $|X'X|$  está muy próximo a cero, pero será distinto de cero, por tanto  $\exists (X'X)^{-1}$  y podremos calcular los estimadores MCO. Además estos estimadores serán **lineales, insesgados y de varianza mínima**. Sin embargo la existencia de alta colinealidad entre variables produce efectos importantes que deben ser tenidos en cuenta y que son los siguientes:

- Varianzas y covarianzas cuantitativamente muy grandes:  
Dado que  $(X'X)$  es casi singular, el valor de  $|X'X|$  será muy pequeño, por lo que,  $(X'X)^{-1}$  tendrá elementos muy grandes. Así, encontraremos varianzas y covarianzas muy grandes, pero estos valores serán los más pequeños que podemos encontrar en estas circunstancias. Cualquier otro estimador tendrá varianza mayor y por tanto el estimador MCO seguirá siendo

de varianza mínima. Aunque como consecuencia del tamaño de  $(X'X)^{-1}$ , las estimaciones sean muy imprecisas<sup>5</sup>.

- Como consecuencia de lo anterior, podremos encontrar  $R^2$  grandes, que indican que las variables exógenas conjuntamente explican mucho de la variabilidad de la variable endógena, unidos a variables explicativas que aportan poco a explicar esta variabilidad.
- Pequeños cambios en los datos producen cambios importantes en las estimaciones de los parámetros.

### ¿Cómo podemos analizar si existe un problema de alta colinealidad?

- Una primera aproximación consiste en obtener los coeficientes de correlación muestral simples para cada par de variables explicativas y ver si el grado de correlación entre estas variables es alto.
- El valor del determinante decrece cuando aumenta la colinealidad, tendiendo a cero cuando esta se hace exacta. Este hecho podemos interpretarlo como un aviso pero no tenemos una medida que nos permita afirmar cuando es grave o muy grave.
- Valores altos del  $R^2$  y en  $(X'X)^{-1}$ , especialmente en su diagonal.
- Otra forma de **detectar la multicolinealidad** consiste en realizar la regresión de cada una de las variables explicativas sobre el resto<sup>6</sup> y analizar los coeficientes de determinación de cada regresión. Si alguno o algunos de estos coeficientes de determinación ( $R_j^2$ ) son altos, estaría señalando la posible existencia de un problema de multicolinealidad.
- Belsley, Kuh y Welsch (1980) consideran una serie de indicadores para analizar el grado de multicolinealidad entre los regresores de un modelo, como por ejemplo los llamados **Tolerancia** (TOL) y **Factor de Inflación de la Varianza** (VIF) que se definen:

$$VIF_j = \frac{1}{(1 - R_j^2)} \quad TOL_j = \frac{1}{VIF_j}$$

siendo  $R_j^2$  el coeficiente de determinación de la regresión auxiliar de la variable  $X_j$  sobre el resto de las variables explicativas y  $1 \leq VIF_j \leq \infty$ .

La varianza de cada uno de los coeficientes de la regresión MCO ( $\hat{\beta}_j$ ) de un modelo de regresión lineal general se puede expresar como:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} \frac{1}{(1 - R_j^2)} = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} VIF_j$$

---

<sup>5</sup>Como veremos en la sección de Contraste de hipótesis el mayor tamaño de las varianzas hará que aumente la probabilidad de no rechazar la hipótesis nula de significatividad individual, cuando en realidad la variable sea significativa, sólo que los datos no permiten detectar esta significatividad.

<sup>6</sup>En cada regresión se incluye el término constante como regresor pero no como variable dependiente.

donde  $\beta_j$ , es el coeficiente que acompaña a la variable  $X_j$  y  $R_j^2$  es el coeficiente de determinación de la regresión auxiliar de la variable  $X_j$  en función del resto de las variables explicativas. Como vemos existe una relación inmediata entre el valor  $VIF_j$  y la varianza del coeficiente estimado. Cuanto más se acerque  $R_j^2$  a la unidad, es decir, cuanto mayor sea la colinealidad de la variable  $X_j$  con el resto, mayor es el valor de  $VIF_j$  y mayor es la varianza del coeficiente estimado, porque tal y como hemos dicho, la multicolinealidad “infla” la varianza. Según estos autores, si  $VIF_j > 10$ , entonces concluiremos que la colinealidad de  $X_j$  con las demás variables es alta.

La utilización de los coeficientes  $TOL$  y  $VIF$  para detectar la presencia de la multicolinealidad ha recibido múltiples críticas, porque la conclusión obtenida con estos valores no siempre recoge adecuadamente la información y problema de los datos. Tal y como hemos visto anteriormente, las varianzas de los estimadores dependen del  $VIF_j$ ,  $\sigma^2$  y  $\sum (X_{ji} - \bar{X}_j)^2$ , por lo que un alto  $VIF_j$  no es condición suficiente ni necesaria para que dichas varianzas sean elevadas ya que es posible que  $\sigma^2$  sea pequeño o  $\sum (X_{ji} - \bar{X}_j)^2$  grande y se compensen.

En la literatura se han propuesto muchas soluciones al posible problema de alta colinealidad y ninguna de ellas es totalmente satisfactoria, por ello parece sensato aprender a convivir con el problema y tener cuidado de no omitir aquellas variables que esconden su significatividad bajo un problema de colinealidad y no incurrir así en un problema de mala especificación. Aunque no es fácil, se pueden considerar las siguientes “soluciones” para intentar resolver el problema:

- Si realmente es un problema muestral, una posibilidad es cambiar de muestra porque puede ser que con nuevos datos el problema se resuelva, aunque esto no siempre ocurre. La idea consiste en conseguir datos menos correlacionados que los anteriores, bien cambiando toda la muestra o simplemente incorporando más datos en la muestra inicial. De todas formas, no siempre resulta fácil obtener mejores datos por lo que muy probablemente debamos convivir con el problema teniendo cuidado con la inferencia realizada y las conclusiones de la misma.
- En ocasiones, si se incorpora información a priori sobre los coeficientes del modelo desaparece el problema. Aún así, sería conveniente tener en cuenta dicha información antes de la detección del problema de multicolinealidad y no posteriormente, ya que así estimaremos el modelo más eficientemente.

#### Ejercicio 4.4

Comenta la siguiente afirmación:

“Si al estimar un modelo de regresión lineal la matriz de datos presenta un alto grado de multicolinealidad, la varianza de los estimadores MCO será muy grande y, por tanto, serán ineficientes.”

#### Ejercicio 4.5

Supón que se formula el siguiente modelo:

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 (X_{1t} + X_{2t}) + \beta_4 X_{1t}^2 + u_t \quad u_t \sim N(0, \sigma^2) \quad (4.7)$$

1. Indica qué parámetros son estimables y cuáles no.
2. Supón que conocemos el valor que toma el parámetro  $\beta_3$ , por ejemplo  $\beta_3 = \beta_{30}$ . Indica si esta información adicional varía la conclusión anterior.

#### Ejercicio 4.6

Supón que se formula el siguiente modelo:

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 (7X_{1t}) + \alpha_3 X_{1t}^2 + u_t \quad (4.8)$$

1. Indica qué parámetros son estimables y cuáles no.
2. Explica intuitivamente porqué no podemos estimar todos los coeficientes.

### 4.4. Consecuencias del incumplimiento de algunos supuestos: omisión de variables relevantes e inclusión de variables irrelevantes

Dentro de las hipótesis básicas hemos supuesto que el modelo estaba correctamente especificado, esto en ocasiones no es así bien porque faltan variables (omisión de variables relevantes) o porque hay más de las necesarias (inclusión de variables irrelevantes). Estas situaciones influyen en las propiedades del estimador MCO y es necesario tenerlo en cuenta.

#### 4.4.1. Omisión de variables relevantes

Suponemos que el modelo correctamente especificado es:

$$Y = X\beta + u = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u = X_1\beta_1 + X_2\beta_2 + u \quad (4.9)$$

donde  $X_1$  es una submatriz de orden  $(N \times K_1)$  y  $X_2$  es una submatriz de orden  $(N \times K_2)$  y por tanto  $\beta_1$  es un subvector de orden  $(K_1 \times 1)$  y  $\beta_2$  es un subvector de orden  $(K_2 \times 1)$ . Pero nosotros estimamos el siguiente modelo incorrectamente especificado:

$$Y = X_1\beta_1 + v \quad \text{donde} \quad v = X_2\beta_2 + u \quad (4.10)$$

El modelo (4.10) incurre en un error de especificación ya que se omiten las variables relevantes recogidas en  $X_2$ . Esto es lo mismo que imponer la restricción vectorial  $\beta_2 = 0$  cuando no es cierta.

El estimador MCO de  $\beta_1$  es  $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$ , y  $\hat{v} = Y - X_1\hat{\beta}_1$ . Consecuencias:

- En general los estimadores son sesgados:

$$E(\hat{\beta}_1) = E((X_1'X_1)^{-1}X_1'Y) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

$Sesgo(\hat{\beta}_1) = (X_1'X_1)^{-1}X_1'X_2\beta_2$  y se anulara si  $X_1'X_2 = 0$ , es decir, si las variables omitidas son ortogonales a las no omitidas. Notar que el sesgo se anula también para  $\beta_2 = 0$  pero esta es una solución trivial dado que al ser  $X_2$  regresores relevantes necesariamente  $\beta_2 \neq 0$ .

- Las matriz de varianzas y covarianzas es  $V(\hat{\beta}_1) = \sigma^2(X_1'X_1)^{-1}$
- El estimador de la varianza de la perturbación es sesgado, y lo es **siempre** incluso cuando los regresores son ortogonales:

$$\hat{\sigma}^2 = \frac{\hat{v}'\hat{v}}{N - K_1} \longrightarrow E(\hat{\sigma}^2) = \frac{E(\hat{v}'\hat{v})}{N - K_1} \neq \sigma^2$$

#### 4.4.2. Inclusión de variables irrelevantes

Este caso formalmente es justo el inverso del anterior. El modelo correctamente especificado es:

$$Y = X_1\beta_1 + u \quad u \sim N(0, \sigma^2 I) \quad (4.11)$$

y el modelo estimado es:

$$Y = X_1\beta_1 + X_2\beta_2 + v \quad (4.12)$$

donde aparecen las variables irrelevantes en la matriz  $X_2$  de orden  $(N \times K_2)$  con unos coeficientes,  $\beta_2$ , de orden  $(K_2 \times 1)$ , que son cero, poblacionalmente. Consecuencias:

- Los estimadores de los coeficientes son insesgados. Podemos escribir el modelo correcto como:

$$Y = X_1\beta_1 + X_2 0 + u \quad (4.13)$$

$$\begin{aligned} E \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= E \left( \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix} + \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'u \\ X_2'u \end{bmatrix} \right) = \\ &= \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix} + \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \underbrace{\begin{bmatrix} X_1'E(u) \\ X_2'E(u) \end{bmatrix}}_0 = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix} \end{aligned}$$

ya que  $X$  es fija y  $E(u) = 0$ . Por lo tanto, el estimador de (4.12) sigue siendo **insesgado** aunque se incluyan variables irrelevantes.

- Las matriz de varianzas y covarianzas es  $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- El estimador de la varianza de las perturbaciones del modelo (4.12) es un estimador **insesgado** de  $\sigma^2$

$$\hat{\sigma}^2 = \frac{\hat{v}'\hat{v}}{N - (K_1 + K_2)}$$

### 4.5. Predicción

Aunque pueda considerarse que la obtención de un buen conjunto de estimaciones es el objetivo principal de la Econometría, a menudo también tiene gran importancia el logro de unas predicciones precisas. Supongamos que con  $N$  observaciones se ha estimado el modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i.$$

Dada una nueva observación de las variables explicativas,

$$X'_p = [ 1 \quad X_{2p} \quad \cdots \quad X_{Kp} ] \quad p \notin \{1, 2, \dots, N\}$$

se puede utilizar el modelo estimado por MCO para predecir el valor que tendrá la variable endógena (desconocido en ese momento). Dado el modelo de regresión, la ecuación para  $Y_p$  es:

$$Y_p = \beta_1 + \beta_2 X_{2p} + \dots + \beta_K X_{Kp} + u_p$$

Para abreviar, utilizaremos la expresión vectorial:

$$Y_p = X'_p \beta + u_p$$

Dada la información muestral disponible (no conocemos  $\beta$  ni  $u_p$ ) la **predicción por punto de  $Y_p$**  es:

$$\hat{Y}_p = X'_p \hat{\beta}_{MCO}$$

O lo que es lo mismo:

$$\hat{Y}_p = \hat{\beta}_1 + \hat{\beta}_2 X_{2p} + \dots + \hat{\beta}_K X_{Kp}.$$

Hay cuatro fuentes potenciales de error al realizar una predicción:

1. El error de especificación. El modelo de regresión en que nos basamos puede ser incorrecto: pueden faltar variables explicativas que afectan de manera clave a  $Y$ , puede que la forma funcional propuesta no sea correcta, puede que se no se cumpla alguna hipótesis básica, etc.
2. Error en los valores de  $X_p$ . La predicción se hace para unos valores dados de  $X_p$ , pero estos pueden ser desconocidos en el momento en que se hace la predicción.
3. El error muestral. No hay más remedio que usar  $\hat{\beta}$  en vez de los valores verdaderos  $\beta$  para hacer la predicción.
4. El error aleatorio.  $Y_p$  dependerá de  $u_p$ , la perturbación aleatoria (desconocida) correspondiente a esa observación. Cuanto más diferente sea de cero, mayor será este error.

Dadas todas estas fuentes de incertidumbre a la hora de predecir  $Y$ , es muy recomendable que la predicción puntual de  $Y$  se acompañe con una medida de lo precisa que esperamos que sea esa predicción. En esto consiste la predicción por intervalo.

#### • Predicción por intervalo del valor de la variable endógena

Es muy difícil que el valor predicho para  $Y_p$ ,  $\hat{Y}_p$  coincida con el valor real. Si la predicción por punto se hace para el mes siguiente, o para el año siguiente, llegará un momento en que conoceremos el error cometido. Este error se denomina **error de predicción** y es igual a

$$e_p = Y_p - \hat{Y}_p$$

En el momento en que hacemos la predicción, tenemos cierta información sobre  $e_p$ , ya que es una variable aleatoria con una distribución conocida. En concreto,

$$e_p \sim N(0, \sigma^2(1 + X_p' (X'X)^{-1} X_p))$$

DEMOSTRACIÓN:

$$\begin{aligned} e_p &= Y_p - \hat{Y}_p = X_p' \beta + u_p - X_p' \hat{\beta} = \\ &= u_p - X_p' (\hat{\beta} - \beta) \end{aligned} \quad (4.14)$$

Buscamos su distribución. Si  $u_p$  es normal el estimador MCO dado que es lineal en la perturbación también lo será y por tanto el error de predicción también lo es. En cuanto a su media y varianza:

$$E(e_p) = E[u_p - X_p' (\hat{\beta} - \beta)] = 0 - X_p' (\beta - \beta) = 0$$

$$\begin{aligned} V(e_p) &= E[e_p - E(e_p)] [e_p - E(e_p)]' = \\ &= E(e_p e_p') = \\ &= E\left[\left(u_p - X_p' (\hat{\beta} - \beta)\right) \left(u_p - X_p' (\hat{\beta} - \beta)\right)'\right] = \\ &= E[u_p u_p'] + E\left[X_p' (\hat{\beta} - \beta) (\hat{\beta} - \beta)' X_p\right] - 2X_p' E\left[(\hat{\beta} - \beta) u_p'\right] = \\ &= E(u_p^2) + X_p' E\left[(\hat{\beta} - \beta) (\hat{\beta} - \beta)'\right] X_p - 2X_p' E\left[(X'X)^{-1} X' u u_p\right] = \\ &= \sigma^2 + \sigma^2 X_p' (X'X)^{-1} X_p - 0 = \\ &= \sigma^2 \left(1 + X_p' (X'X)^{-1} X_p\right) \end{aligned}$$

Por tanto:

$$e_p \sim N(0, \sigma^2 \left(1 + X_p' (X'X)^{-1} X_p\right))$$

Tipificando el error de predicción queda:

$$\frac{e_p - 0}{\sigma \sqrt{1 + X_p' (X'X)^{-1} X_p}} \sim N(0, 1)$$

El problema es que  $\sigma^2$  es desconocida. Utilizando que  $e_p$  y  $\hat{\sigma}^2$  obtenemos

$$\frac{e_p}{\hat{\sigma} \sqrt{1 + X_p' (X'X)^{-1} X_p}} \sim t_{(N-K)}$$

De hecho el denominador final es  $\hat{\sigma}_{e_p}$  (la desviación estimada del error de predicción). Tras sustituir  $e_p = Y_p - \hat{Y}_p$ , se puede utilizar dicha distribución para obtener el siguiente intervalo de predicción para la variable endógena:

$$Pr \left[ -t_{\frac{\alpha}{2}(N-K)} \leq \frac{Y_p - \hat{Y}_p}{\hat{\sigma}_{e_p}} \leq t_{\frac{\alpha}{2}(N-K)} \right] = 1 - \alpha$$

$$Pr \left[ \hat{Y}_p - t_{\frac{\alpha}{2}(N-K)} \cdot \hat{\sigma}_{e_p} \leq Y_p \leq \hat{Y}_p + t_{\frac{\alpha}{2}(N-K)} \cdot \hat{\sigma}_{e_p} \right] = 1 - \alpha$$

$$IC_{1-\alpha}(Y_p) = \left( \hat{Y}_p - t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{e_p}, \hat{Y}_p + t_{\frac{\alpha}{2}(N-K)} \hat{\sigma}_{e_p} \right)$$

### Ejercicio 4.7

Una gestoría comercializa 100 pólizas de seguro. Para tener información acerca del tipo de póliza más solicitada utiliza el siguiente modelo:

$$N_i = \beta_1 H_i + \beta_2 V_i + \beta_3 M_i + \beta_4 A_i + u_i \quad i = 1, \dots, 100$$

donde

- $N_i$  es el número de pólizas suscritas del producto  $i$ -ésimo.
- $H_i = \begin{cases} 1 & \text{si la póliza pertenece a seguros del hogar} \\ 0 & \text{en caso contrario} \end{cases}$
- $V_i = \begin{cases} 1 & \text{si la póliza pertenece al ramo de seguros de vida} \\ 0 & \text{en caso contrario} \end{cases}$
- $M_i = \begin{cases} 1 & \text{si la póliza pertenece al ramo de seguros médicos} \\ 0 & \text{en caso contrario} \end{cases}$
- $A_i = \begin{cases} 1 & \text{si la póliza pertenece a seguros automovilísticos} \\ 0 & \text{en caso contrario} \end{cases}$

Las ventas para el año 1994 se resumen en:

$$\sum_{i \in \text{Hogar}} N_i = 380 \quad \sum_{i \in \text{Vida}} N_i = 150 \quad \sum_{i \in \text{Médicos}} N_i = 70 \quad \sum_{i \in \text{Automóviles}} N_i = 500$$

$$\sum_{i=1}^{100} N_i^2 = 20000 \quad \sum_{i=1}^{100} H_i = \sum_{i=1}^{100} V_i = \sum_{i=1}^{100} M_i = \sum_{i=1}^{100} A_i = 25$$

1. Interpreta los parámetros del modelo propuesto.
2. Estima los parámetros por MCO.
3. Contrasta si hay diferencias significativas en el número de suscripciones según el ramo de seguros al que pertenecen.
4. Halla un intervalo de confianza del 95 % para el número de suscripciones de pólizas médicas.

**Ejercicio 4.8**

Supón que se quieren explicar los rendimientos de los índices bursátiles,  $R_t$ , de varios países a través de su localización geográfica y de los tipos de interés reales,  $I_t$ , considerando para estos últimos que es suficiente con especificar si los tipos son positivos o negativos. Los datos aparecen en la Tabla 4.5.

1. Especifica un modelo para el comportamiento de  $R_t$  con la información que se da.
2. Supón ahora que sólo se considera importante en cuanto a la localización si el país está o no en Europa (i.e. Europa v.s. Resto del Mundo) y reespecifica el modelo.
3. Estima el modelo que has propuesto en el apartado anterior.
4. Supón que, para Portugal, el tipo de interés es positivo y el rendimiento de su índice bursátil del 35%. ¿Consideras que el mercado bursátil en Portugal tiene el mismo comportamiento que en los otros países europeos?

	<i>Rendimiento</i> $R_t$	<i>t/i</i> $I_t$
<b>Europa</b>		
Alemania	10 %	positivo
Francia	20 %	positivo
Reino Unido	15 %	positivo
Italia	40 %	positivo
<b>América</b>		
EEUU	20 %	negativo
Canadá	22 %	negativo
<b>Sudeste Asiático</b>		
Japón	-5 %	positivo
Hong Kong	30 %	positivo
Singapur	35 %	negativo
Taiwán	22 %	negativo

Tabla 4.1: Observaciones sobre rendimiento y t/i por país

**Ejercicio 4.9**

La función de producción Cobb-Douglas tiene la forma  $Y = AK^{\beta_1}L^{\beta_2}$ . Dada una muestra de tamaño  $N$ ,

1. ¿Es el modelo lineal? ¿Qué método se podría usar para estimar los parámetros?
2. Interpreta los coeficientes del modelo.
3. ¿Cómo contrastarías si la empresa tiene rendimientos constantes a escala?
4. Si  $H_0$  es cierta, ¿qué método mejora la estimación de los parámetros? ¿por qué?

### Ejercicio 4.10

Un economista estima el siguiente modelo:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad t = 1, \dots, 36 \quad u_t \sim N(0, \sigma^2) \quad (4.15)$$

Obteniendo los siguientes resultados:

$$\hat{Y}_t = 0,98 + 1,33 X_{2t} + 0,78 X_{3t} \quad R^2 = 0,989$$

(estadístico t)                      (0,27)                      (0,98)

1. ¿Te parece que el modelo puede tener algún problema? ¿Cuál? ¿En qué basas tus respuestas?
2. ¿Cómo cambiarían tus conclusiones si en la estimación anterior el coeficiente de determinación hubiera sido  $R^2 = 0,12$ ?

### Ejercicio 4.11

Supongamos que queremos analizar los salarios de una famosa empresa auditora y para ello hemos tomado una muestra de 500 empleados, de los cuales 300 son licenciados en Ciencias Actuariales. Se postula el siguiente modelo:

$$W_i = \alpha + \beta D_i + u_i \quad i = 1, 2, \dots, 500$$

donde  $W_i$  es el salario del empleado  $i$ -ésimo y  $D_i$  es una variable ficticia que toma el valor 1 si el empleado es licenciado y 0 en el caso contrario.

Si el salario medio **muestral** de los licenciados es de 250.000 pts. al mes, y el salario medio **muestral** de los que no son licenciados es 150.000 pts.:

1. Calcula  $\hat{\alpha}_{MCO}$  y  $\hat{\beta}_{MCO}$ .
2. ¿Cómo contrastarías si existe diferencia salarial entre los licenciados y los que no lo son?
3. Si además en esta muestra dispones de información sobre el sexo del empleado  $i$ -ésimo ¿cómo contrastarías la hipótesis de que no existe en esta empresa discriminación salarial por razones de sexo?

### Ejercicio 4.12

En una muestra de 8 observaciones se quiere estudiar la dependencia del precio de una prima de seguro obligatorio de automóvil respecto de la renta del propietario y su país de matriculación, distinguiendo entre vehículos matriculados en Europa o el resto del mundo. Las observaciones se muestran en la Tabla 4.2.

1. Estima el modelo que exprese dicha dependencia. Interpreta los coeficientes del modelo.

<i>Europa</i>		<i>Resto del Mundo</i>	
<i>Prima pagada</i>	<i>Renta</i>	<i>Prima pagada</i>	<i>Renta</i>
150	8.000	130	5.000
180	11.000	100	3.000
120	9.000	110	6.000
160	12.000	140	10.000

Tabla 4.2: Observaciones muestrales de la prima pagada y renta

2. Contrasta las siguientes hipótesis:

- La renta no es una variable significativa.
- El país de matriculación no es una variable significativa.

### Ejercicio 4.13

Un investigador quiere analizar la inversión ( $I_i$ ) realizada en planes de pensiones en la Comunidad Autónoma Vasca en función del salario percibido ( $W_i$ ) y el sector, privado o público, en el que se trabaja. Con una muestra de 500 individuos, de los cuales la mitad trabaja en el sector público, se ha estimado por MCO el siguiente modelo:

$$\hat{I}_i = 2,7 + 0,31PU_i + 0,47W_i \quad R^2 = 0,71 \quad (4.16)$$

(t - est)                      (0, 22)                      (3, 7)

donde  $PU_i$  toma valor 1 si el individuo  $i$ -ésimo trabaja en el sector público y cero en caso contrario.

- Deriva** los estimadores utilizados en la estimación del modelo (4.16).
- Interpreta los coeficientes del modelo.
- Contrasta la significatividad individual de las variables explicativas.
- Dados los resultados de los contrastes, ¿qué propiedades tienen los estimadores de los coeficientes del modelo (4.16)? ¿por qué?

Más tarde, el investigador sospecha que la variable sexo puede afectar al incremento de la inversión media ante aumentos unitarios en el salario. Con este objetivo estima los siguientes modelos:

$$\hat{I}_i = 3,87 + 2,28PU_i + 0,63W_i + 0,21W_iS_i \quad R^2 = 0,82 \quad (4.17)$$

(t - est)                      (13, 8)                      (7, 26)                      (17, 4)

$$\hat{I}_i = 4,7 + 0,32W_i + 0,15W_iS_i \quad R^2 = 0,75 \quad (4.18)$$

(t - est)                      (5, 82)                      (1, 87)

donde  $S_i$  toma valor 1 si el individuo  $i$ -ésimo es hombre y cero en caso contrario.

5. Dadas las regresiones (4.16), (4.17) y (4.18), selecciona la especificación más adecuada para analizar la inversión en los planes de pensiones. Razona detalladamente tu proceso de selección.
6. ¿Qué sucedería en la estimación de los modelos (4.17) y (4.18) si la muestra estuviera compuesta sólo por hombres?
7. Suponiendo que eliges el modelo (4.18) ¿qué propiedades tienen los estimadores?

#### Ejercicio 4.14

Para analizar la relación entre los gastos de consumo,  $C$ , el ingreso disponible,  $Y$ , y el sexo del cabeza de familia se han estimado, con  $T=12$ , cuatro modelos cuyos resultados han sido:

$$\begin{array}{ll}
 \text{a)} & \hat{C}_t = 1663,6 + \begin{array}{c} 0,75 \\ (21,12) \end{array} Y_t & R^2 = 0,978 \\
 \text{b)} & \hat{C}_t = 186,12 + \begin{array}{c} 0,82 \\ (16,56) \end{array} Y_t + 832,09 D_t & R^2 = 0,984 \\
 & & \quad \quad \quad (1,82) \\
 \text{c)} & \hat{C}_t = 709,18 + \begin{array}{c} 0,79 \\ (18,11) \end{array} Y_t + \begin{array}{c} 0,05 \\ (1,51) \end{array} Y_t D_t & R^2 = 0,983 \\
 \text{d)} & \hat{C}_t = -184,7 + \begin{array}{c} 0,83 \\ (13,65) \end{array} Y_t + 1757,99 D_t - \begin{array}{c} 0,06 \\ (-0,57) \end{array} Y_t D_t & R^2 = 0,985 \\
 & & \quad \quad \quad (1,03)
 \end{array}$$

donde  $D_t$  es una variable ficticia que toma el valor 1 si el cabeza de familia es mujer y 0 en caso contrario. Entre paréntesis aparecen los estadísticos  $t$  muestrales:

1. Interpreta los coeficientes del modelo d).
2. Contrasta si el término constante es diferente según sea el sexo del cabeza de familia.
3. Contrasta si el coeficiente asociado a  $Y$  es diferente según sea el sexo del cabeza de familia.
4. Contrasta si tanto el coeficiente asociado a  $Y$  como el término constante son diferentes según sea el sexo del cabeza de familia.
5. ¿Cuál es el modelo que mejor se ajusta a los datos?

#### Ejercicio 4.15

Se quiere estudiar cómo influye el nivel de educación en los salarios ( $Y$ ) percibidos por los trabajadores de una empresa. En la empresa hay 100 empleados, de los cuales 6 no tienen estudios, 38 tienen únicamente estudios primarios, 15 tienen estudios universitarios y el resto se incorporó a la empresa tras finalizar sus estudios secundarios. Para introducir el nivel de estudios en el modelo se definen las siguientes variables ficticias:

$$P_i = \begin{cases} 1 & \text{si } i \text{ tiene al menos estudios primarios} \\ 0 & \text{en caso contrario} \end{cases}$$

$$S_i = \begin{cases} 1 & \text{si } i \text{ tiene al menos estudios secundarios} \\ 0 & \text{en caso contrario} \end{cases}$$

$$U_i = \begin{cases} 1 & \text{si } i \text{ tiene al menos estudios universitarios} \\ 0 & \text{en caso contrario} \end{cases}$$

De modo que para un empleado que, por ejemplo, no haya realizado estudios primarios  $P_i = 0, S_i = 0$  y  $U_i = 0$  y para un universitario  $P_i = 1, S_i = 1$  y  $U_i = 1$ .

1. ¿Es estimable el modelo  $Y_i = \beta_1 + \beta_2 P_i + \beta_3 S_i + \beta_4 U_i + V_i$ , con  $V_i \sim NID(0, \sigma^2)$ ? Razona la respuesta. En caso de que no sea estimable, propón uno similar que sí lo sea.
2. Interpreta los parámetros  $\beta_1$  y  $\beta_4$  en el modelo anterior, o en el modelo estimable propuesto.
3. Plantea un contraste para la hipótesis de que tener estudios primarios no afecta al salario que se cobra. Escribe la hipótesis nula, el estadístico de contraste y la regla de decisión.

#### Ejercicio 4.16

Un empleado de BBK Peña opina que los pisos de Avenida Mazarredo alcanzan en las subastas un precio considerablemente superior al de cualquier otra zona de Bilbao. Un colega suyo de Lagunaro no está de acuerdo con esta afirmación, por lo que analizan las ventas de pisos de dicha calle realizadas por sus empresas en los últimos años. El número total de pisos subastados fue de 10, de los que 4 estaban enclavados en la calle Mazarredo. Ambos creen que el tamaño del piso y el precio inicial fijado en la subasta son también variables determinantes del precio finalmente alcanzado en la venta.

1. Especifica un modelo que exprese el precio final de venta de los pisos en función de los factores citados y que permita contrastar la hipótesis del empleado de BBK.
2. Escribe la matriz de regresores del modelo que has propuesto.
3. ¿Cómo llevarías a cabo el contraste? Escribe la hipótesis nula, la hipótesis alternativa, el estadístico de contraste, su distribución y la regla de decisión.

#### Ejercicio 4.17

Se quiere estudiar la relación entre el número de incendios forestales y la temperatura en una comarca catalana. Para ello se utilizan datos diarios de ambas variables para el año 1999 y se estiman los siguientes modelos:

$$\begin{array}{ll}
 (1) & Y_t = 105 + 400D_t + 4,7X_t + 8,3D_tX_t + \hat{u}_{1t} & \hat{\sigma}_1^2 = 14,71 \\
 (2) & Y_t = 205 + 43,5X_t + 9,0D_tX_t + \hat{u}_{2t} & \hat{\sigma}_2^2 = 20,15 \\
 (3) & Y_t = 124 + 557D_t + 7,3X_t + \hat{u}_{3t} & \hat{\sigma}_3^2 = 15,14 \\
 (4) & Y_t = 350 + 7,7X_t + \hat{u}_{4t} & \hat{\sigma}_4^2 = 27,05
 \end{array}$$

siendo  $D_t = \begin{cases} 1 & \text{si } t \in \text{Julio, Agosto o Septiembre} \\ 0 & \text{en caso contrario} \end{cases}$

1. Interpreta los coeficientes del modelo (1).
2. El contraste del modelo (3) contra el (4) nos ha llevado a rechazar el modelo (4). Efectúa los contrastes necesarios para decidir cuál es el modelo al que mejor se ajustan los datos.

### Ejercicio 4.18

Para estudiar las ventas de un determinado modelo de moto se especifica el siguiente modelo de regresión lineal:

$$Y_t = \beta_0 + \beta_1 S_{1t} + \beta_2 S_{2t} + \beta_3 S_{3t} + \beta_4 X_t + u_t \quad (4.19)$$

donde  $S_{it} = \begin{cases} 1 & \text{si } t \in \text{trimestre } i\text{-ésimo} \\ 0 & \text{en caso contrario} \end{cases}$

1. Explica qué tipo de comportamiento refleja esta regresión e interpreta cada uno de los coeficientes del modelo.
2. ¿Existe alguna forma alternativa de especificar el comportamiento reflejado por el modelo anterior? Escribe la ecuación correspondiente e interpreta sus parámetros.
3. Encuentra la relación entre los parámetros de la ecuación (4.19) y los de la ecuación propuesta en el apartado anterior.
4. Los resultados obtenidos de estimar el modelo (4.19) con 16 observaciones trimestrales han sido:

$$\hat{Y}_t = \underset{(3,7)}{0,07} + \underset{(3,37)}{0,43} S_{1t} + \underset{(3,4)}{6,55} S_{2t} - \underset{(-3,37)}{2,83} S_{3t} - \underset{(-0,27)}{0,9} X_t \quad R^2 = 0,68 \quad (4.20)$$

A la vista de los resultados ¿se puede concluir que la variable  $Y_t$  presenta un comportamiento estacional?

### Ejercicio 4.19

El dueño de una cadena de cines en una ciudad costera pretende saber cómo influyen en el número de espectadores de sus salas,  $N$  (*miles*), dos factores: el empleo existente

en la ciudad,  $E$  (*cientos de miles*), y el número de otros espectáculos que se programen,  $O$  (*miles*). Propone el siguiente modelo

$$N_t = \alpha_0 + \alpha_1 E_t + \alpha_2 O_t + u_t \quad u_t \sim NID(0, \sigma^2) \quad (4.21)$$

Dispone de información trimestral relativa a las tres variables que va desde el primer trimestre de 1985 hasta el primer trimestre de 2002, ambos inclusive. El resumen de dicha información aparece a continuación:

$$(X'X)^{-1} = \begin{bmatrix} 1,5334 & -0,2122 & 0,00734 \\ & 0,0320 & -0,00379 \\ & & 0,00329 \end{bmatrix} \quad \begin{array}{l} \sum N_t = 2279,4 \\ \sum N_t E_t = 16838,6 \\ \sum N_t O_t = 14193,8 \\ \sum (N_t - \bar{N})^2 = 2435 \end{array}$$

- Estima el modelo propuesto por el método de mínimos cuadrados ordinarios.
- Calcula una medida de la bondad del ajuste e interprétala.
- Estima la matriz de varianzas y covarianzas del estimador de los coeficientes.
- ¿Son las variables explicativas conjuntamente significativas?
- Contrasta al nivel de significación del 5 % que la programación de otros espectáculos tiene influencia sobre el número de espectadores en las salas de cine.
- Para poder planificar la próxima campaña, el dueño hace el supuesto de que va a tener 34.500 espectadores, dado que el número de empleos será de 754.000 y que están previstos otros 8.800 espectáculos. Compruébalo.

Analizando los resultados con un amigo suyo, observan una pauta de comportamiento diferente en los distintos trimestres. Por ello, piensan que no ha tenido en cuenta que la época del año marca también la asistencia al cine, debido a que se va menos cuando hace mejor tiempo. Nuestro hombre decide incluir este factor.

- Define las variables que creas necesarias para tener en cuenta este factor y propón un modelo que incluya el posible efecto de la época del año.
- Interpreta los coeficientes de tu modelo. ¿Cuántas de las variables que has definido incluyes en el mismo? ¿Por qué?

A los pocos días, su amigo le plantea la siguiente duda: a su ciudad acuden muchos turistas en verano, lo que aumenta el empleo temporal que suele cubrirse con gente joven, en general más aficionada al cine. Si esto es cierto, la composición del empleo en verano puede tener influencia sobre la asistencia al cine.

- ¿Cómo definirías una variable para tener en cuenta esta interrelación? ¿Cómo la incluirías en el modelo?

En línea con lo planteado, el empresario escoge y estima tres modelos diferentes de entre todos los posibles, con los siguientes resultados:

$$\hat{N}_t = \underbrace{23,78}_{(\hat{\sigma}_{\beta_i})} + \underbrace{1,371}_{(0,421)} E_t - \underbrace{0,317}_{(0,135)} O_t + \underbrace{4,571}_{(0,658)} D3t, \quad \sum_{t=1}^{69} \hat{u}_t^2 = 359,89 \quad (4.22)$$

$$\hat{N}_t = \underbrace{25,0}_{(\hat{\sigma}_{\beta_i})} + \underbrace{1,204}_{(0,418)} E_t - \underbrace{0,316}_{(0,134)} O_t + \underbrace{0,624}_{(0,088)} EM3t, \quad \sum_{i=1}^{69} \hat{u}_t^2 = 355,23 \quad (4.23)$$

$$\hat{N}_t = \underbrace{25,43}_{(\hat{\sigma}_{\beta_i})} + \underbrace{1,146}_{(0,483)} E_t - \underbrace{0,316}_{(0,135)} O_t - \underbrace{1,613}_{(6,552)} D3t + \underbrace{0,841}_{(0,887)} EM3t, \quad \sum_{i=1}^{69} \hat{u}_t^2 = 354,90 \quad (4.24)$$

donde

- $D3_t$  es una variable ficticia que toma el valor 1 en el tercer trimestre (verano) y cero en los demás casos.
  - $EM3_t$  es una variable que incorpora el efecto estacional del empleo que se ha discutido en el apartado anterior (y tendrá la interpretación que le hayas atribuido anteriormente).
- j) Dados los resultados anteriores, ¿hay evidencia de que existe efecto estacional? Si el efecto estacional existe ¿qué forma toma?, ¿cuál de los modelos propuestos te parece óptimo para recogerlo?, ¿por qué? Lleva a cabo los contrastes precisos.
- k) A la vista de los resultados obtenidos hasta el momento ¿es correcto utilizar los valores de los coeficientes que has obtenido en (4.21)?, ¿por qué?, ¿qué propiedades tienen?
- l) ¿Cómo explicas los resultados del modelo (4.24) a la vista del (4.22) y el (4.23)? ¿Te parece lógico obtener este resultado?, ¿cuál puede ser su causa?

## 4.6. Contraste de hipótesis y predicción en el MRLG con gretl

### • Estimación por Mínimos Cuadrados Ordinarios, MCO:

Una vez abierto el fichero de datos con el que vamos a trabajar, vamos a

*Modelo* → *Mínimos Cuadrados Ordinarios*

Aparecerá una ventana para especificar la parte sistemática del modelo donde debemos:

Seleccionar la variable dependiente pinchando a la izquierda sobre ella y a continuación pinchar en la derecha → *la flecha azul*

Seleccionar las variables independientes pinchando a la izquierda sobre ella-s y a continuación pinchar en la derecha → *la flecha verde*

Para obtener los resultados de la estimación MCO pinchar en *Aceptar*. No pinchar en la indicación *Desviaciones Típicas Robustas*.

En esta ventana aparecerán los resultados básicos de la estimación del modelo. Los podemos guardar como texto plano de la manera habitual o como icono con *Archivo* → *Guardar como icono*.

Los resultados que *gretl* nos devuelve muestran entre otros estadísticos la estimación de los parámetros de la recta de ajuste, sus desviaciones típicas y estadísticos de significatividad individual.

### Ejemplo 4.7

Vamos a utilizar como ejemplo la estimación realizada en el Ejemplo 2.13 sobre el precio de la vivienda y los contrastes desarrollados en el Ejemplo 3.2:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS + u_i \quad i = 1, \dots, 14$$

Los resultados de la estimación MCO mostrados por *gretl* son los siguientes:

Modelo 1: MCO, usando las observaciones 1–14

Variable dependiente: price

	Coefficiente	Desv. Típica	Estadístico <i>t</i>	Valor p
const	129,062	88,3033	1,4616	0,1746
sqft	0,154800	0,0319404	4,8465	0,0007
bedrms	-21,5875	27,0293	-0,7987	0,4430
baths	-12,1928	43,2500	-0,2819	0,7838
Media de la vble. dep.	317,4929	D.T. de la vble. dep.	88,49816	
Suma de cuad. residuos	16700,07	D.T. de la regresión	40,86572	
$R^2$	0,835976	$R^2$ corregido	0,786769	
$F(3, 10)$	16,98894	Valor p (de $F$ )	0,000299	
Log-verosimilitud	-69,45391	Criterio de Akaike	146,9078	
Criterio de Schwarz	149,4641	Hannan–Quinn	146,6712	

En la columna con encabezamiento *Coefficiente* aparece la estimación del coeficiente que acompaña a la correspondiente variable. A continuación aparece su *Desviación Típica* y

el estadístico  $t$  de significatividad individual para el contraste  $H_0 : \beta_k = 0$  así como su correspondiente *valor p*.

A continuación aparecen estadísticos de interés como pueden ser la media de la variable dependiente,  $R^2$  o  $\bar{R}^2$  entre otros. La fila:  $F(3, 10) = 16,98894$ ; Valor  $p$  (de  $F$ ) = 0,000299 se corresponde con el valor muestral del estadístico  $F$  para el contraste de significatividad conjunto y su correspondiente *valor-p*. A continuación aparecen los estadísticos de Akaike, Schwarz y Hannan-Quinn para la selección de modelos.

En la pestaña *Contrastes* que aparece en la pantalla de resultados de la regresión podemos *Omitir u añadir variables, sumar los coeficientes y contrastar combinaciones lineales o restricciones lineales* además podremos realizar contrastes sobre los residuos, de los cuales nos ocuparemos en el último tema del curso.

### Ejemplo 4.8

Por ejemplo para contrastar:

$$H_0 : \beta_3 = \beta_4 \quad \text{versus} \quad H_a : \beta_3 \neq \beta_4$$

cuyo estadístico de contraste y distribución asociada son:

$$\frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\hat{\sigma}_{\hat{\beta}_3}^2 + \hat{\sigma}_{\hat{\beta}_4}^2 - 2 \times \widehat{Cov}(\hat{\beta}_3, \hat{\beta}_4)}} \sim t_{N-4}$$

en la pestaña *Contrastes* seleccionamos *Restricciones lineales* y escribimos  $b_3-b_4=0$  y *gretl* nos devuelve el siguiente resultado:

Restricción:

$$b[\text{bedrms}] - b[\text{baths}] = 0$$

Estadístico de contraste:  $F(1, 10) = 0,0266334$ , con *valor p* = 0,873614 luego no se rechaza la hipótesis nula para  $\alpha$  %.

Además nos proporciona las estimaciones restringidas:

	Coefficiente	Desv. Típica	Estadístico t	Valor p
const	127,736	83,9482	1,522	0,1563
sqft	0,157407	0,0264067	5,961	9,44e-05 ***
bedrms	-18,5060	18,4649	-1,002	0,3378
baths	-18,5060	18,4649	-1,002	0,3378

Desviación típica de la regresión = 39,0158

El modelo restringido es:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 (BEDRMS_i + BATHS) + u_i \quad i = 1, \dots, 14$$

y su FRM es  $\widehat{PRICE}_i = 127,736 + 0,1574 SQFT_i - 18,5060 (BEDRMS_i + BATHS_i)$

En la pantalla de resultados de la estimación aparecen en la barra de menú otros estadísticos o resultados que pueden ser de interés, por ejemplo:

- Podemos hacer gráficos de interés: En la opción *Gráficos* podemos hacer gráficos que nos ayudan a interpretar los resultados de la estimación, por ejemplo  
*Gráficos* → *Gráfico de la variable estimada y observada*  
*Gráficos* → *Gráfico de residuos* → *contra alguna de las variables explicativas del modelo*
- En la pestaña *Guardar* podemos guardar variables como los residuos, los residuos al cuadrado, la suma de cuadrados residual y el coeficiente de determinación entre otros.
- En la pestaña *Análisis* nos muestra las estimaciones de la variable endógena, los intervalos de confianza de los coeficientes y la matriz de varianzas y covarianzas entre otros resultados. Para ver y guardar los valores de  $\hat{Y}$ ,  $\hat{u}$  y otros resultados de utilidad:
  - Ver los valores: Pinchar en *Análisis* → *Mostrar variable* y seleccionar *observada*, *estimada* o *residuos* según nuestro interés.
  - Guardar los valores: Pinchar en *Guardar* → *seleccionar la variable de interés*.

*Gretl* utiliza por defecto la denominación *yhat*, *uhat* para designar a la variable endógena estimada y a los residuos, respectivamente y en la descripción de la variable indicará por ejemplo para *uhat*: *residuos del modelo 1*, donde el valor 1 indica que corresponde con el primer modelo estimado, esto resulta muy útil pues en general trabajaremos con varios modelos a la vez y hay que distinguir claramente las variables de cada uno.

### Ejemplo 4.9

La matriz de varianzas y covarianzas de los coeficientes estimados es:

Matriz de covarianzas de los coeficientes

const	sqft	bedrms	baths	
7797,5	0,67089	-1677,1	-1209,4	const
	0,0010202	-0,075461	-0,99507	sqft
		730,58	-356,40	bedrms
			1870,6	baths

Los intervalos de confianza de los coeficientes son:

$$t(10, 0,025) = 2,228$$

Variable	Coefficiente	Intervalo de confianza 95 %	
const	129,062	-67,6903	325,814
sqft	0,154800	0,0836321	0,225968
bedrms	-21,5875	-81,8126	38,6376
baths	-12,1928	-108,560	84,1742

*Gretl* permite trabajar tanto con variables ficticias cuantitativas como cualitativas y su tratamiento no difiere, solo debemos de ocuparnos de especificar correctamente el modelo. En el caso de que la variable ficticia no esté construida *gretl* permite hacerlo. En la pantalla inicial en *Añadir* podemos añadir *Variables ficticias periódicas* que se ajustarán lógicamente a la periodicidad muestral del conjunto de datos, *Variables ficticias para las variables discretas seleccionadas* donde por ejemplo si tenemos una variable que toma valores 1, 2 y 3 podremos construir tres variables ficticias tal como

$$D_1 = \begin{cases} 1 & \text{si la variable toma valor 1} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{si la variable toma valor 2} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{si la variable toma valor 3} \\ 0 & \text{en caso contrario} \end{cases}$$

Por supuesto también podremos introducirlas con el editor.

Veamos un ejemplo aplicado. Abrimos el fichero de datos data7-3 de Ramanathan, que contiene datos para 14 viviendas sobre el precio de venta de la vivienda (PRICE), pies cuadrados habitables (SQFT), número de habitaciones (BEDRMS) y número de baños (BATHS), y una variable ficticia que toma el valor 1 si la vivienda tiene piscina y 0 en caso contrario (POOL), una variable ficticia que toma el valor 1 si la vivienda tiene sala de estar y 0 en caso contrario (FAMROOM) y una variable ficticia que toma el valor 1 si la vivienda tiene chimenea y 0 en caso contrario (FIREPL). Seleccionamos las variables PRICE y POOL y observamos los valores de estas dos variables:

Obs	price	pool
1	199,9	1
2	228,0	0
3	235,0	1
4	285,0	0
5	239,0	0
6	293,0	0
7	285,0	0
8	365,0	1
9	295,0	0
10	290,0	0
11	385,0	1
12	505,0	1
13	425,0	0
14	415,0	0

Por ejemplo, la primera vivienda de la muestra tiene un precio de 199.900 dólares y tiene piscina (ya que la variable POOL toma el valor 1), mientras que la segunda no tiene piscina (la variable POOL toma el valor 0) y su precio de venta es de 228.000 dólares, etc.

Con los datos anteriores podemos obtener fácilmente que el precio medio de la vivienda es 317.493 dólares:

Estadísticos principales, usando las observaciones 1 - 14  
para la variable price (14 observaciones válidas)

Media	Mediana	Mínimo	Máximo
317,49	291,50	199,90	505,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
88,498	0,27874	0,65346	-0,52983

Sin embargo, también es posible obtener el precio medio para las viviendas que tienen piscina, por un lado, y para las que no la tienen, por otro. Para ello, en primer, lugar se selecciona el precio para aquellas viviendas con piscina. Seleccionamos la variable PRICE, pinchamos en *Muestra* → *Definir a partir de v. ficticia...*, seleccionamos la variable POOL y aceptamos.

De esta forma hemos seleccionado el precio para aquellas viviendas que tienen piscina<sup>7</sup>. A continuación, se obtienen los estadísticos principales:

Estadísticos principales, usando las observaciones 1 - 5  
para la variable price (5 observaciones válidas)

Media	Mediana	Mínimo	Máximo
337,98	365,00	199,90	505,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
122,99	0,36390	0,15896	-1,2798

Para seleccionar el precio de las viviendas que no tienen piscina, pinchamos en *Muestra* → *Restringir a partir de criterio*, introducimos la condición  $POOL = 0$  y aceptamos. Los estadísticos principales son los siguientes:

Estadísticos principales, usando las observaciones 1 - 9  
para la variable price (9 observaciones válidas)

Media	Mediana	Mínimo	Máximo
306,11	290,00	228,00	425,00
Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
68,959	0,225275	0,87575	-0,52255

Por tanto, el precio medio de las viviendas con piscina es de 337.980 dólares frente a los 306.111 de las viviendas sin piscina. Dado el modelo una vivienda con piscina es en promedio 31.869 dólares más

<sup>7</sup>Para restablecer el tamaño muestral inicial pinchar en *Muestra* → *Recuperar el rango completo*.

cara que la que no tiene piscina. Notar que no se están teniendo en cuenta otros factores que pueden afectar al precio de la vivienda (número de pies cuadrados habitables, número de habitaciones, etc.).

El sencillo análisis anterior podemos realizarlo mediante un análisis de regresión. Podemos especificar un modelo econométrico utilizando la variable ficticia POOL como regresor, estimarlo, hacer inferencia e ir incorporando otras características que pueden afectar a los precios de las viviendas. Para comenzar, consideramos el siguiente modelo:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + u_i \quad i = 1, \dots, 14 \quad (4.25)$$

donde

- $\alpha_1$ : precio medio de una vivienda sin piscina.
- $\alpha_1 + \alpha_2$ : precio medio de una vivienda con piscina.
- $\alpha_2$ : diferencia en el precio medio de una vivienda con piscina con respecto a una que no la tiene.

Los resultados de estimar el modelo por Mínimos Cuadrados Ordinarios utilizando *gretl* obtenemos que las estimaciones de los coeficientes son las siguientes:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14  
Variable dependiente: price

Variable	Coefficiente	Desv. típica	Estadístico <i>t</i>	valor p
const	306,111	30,2077	10,1335	0,0000
pool	31,8689	50,5471	0,6305	0,5402
Media de la var. dependiente			317,493	
D.T. de la variable dependiente			88,4982	
Suma de cuadrados de los residuos			98550,5	
Desviación típica de los residuos ( $\hat{\sigma}$ )			90,6231	
$R^2$			0,0320632	
$\bar{R}^2$ corregido			-0,0485982	
Grados de libertad			12	
Log-verosimilitud			-81,880	
Criterio de información de Akaike			167,760	
Criterio de información Bayesiano de Schwarz			169,038	

$$\widehat{PRICE}_i = 306,111 + 31,869 POOL_i \quad i = 1, \dots, 14$$

(10,13)      (0,63)

Para contrastar en el modelo (4.25) si hay diferencias significativas en el precio medio de la vivienda entre aquéllas que tienen piscina y las que no, la hipótesis de contraste es  $H_0 : \alpha_2 = 0$ . Este

contraste se puede realizar utilizando el estadístico  $t$  habitual cuyo p-valor es 0,5405, por lo que no se rechaza la hipótesis nula para un nivel de significación del 5%, es decir, el precio medio de la vivienda no es significativamente diferente por el hecho de tener piscina. Alternativamente, se puede realizar el contraste utilizando el estadístico  $F$  basado en las sumas de cuadrados de los residuos siendo en este caso el modelo (4.25) el modelo no restringido mientras que el modelo restringido es  $PRICE_i = \alpha_1 + u_i \quad i = 1, \dots, 14$ .

Supongamos que ampliamos el modelo (4.25) incorporando regresores que podrían explicar el precio de la vivienda como: el hecho de que la vivienda tenga sala de estar o no, el hecho que tenga chimenea o no, su superficie, el número de habitaciones y el número de baños. Las dos primeras son variables ficticias que pueden definirse así:

$$FIREPL_i = \begin{cases} 1 & \text{si la vivienda } i\text{-ésima tiene chimenea} \\ 0 & \text{en caso contrario} \end{cases}$$

$$FAMROOM_i = \begin{cases} 1 & \text{si la vivienda } i\text{-ésima tiene sala de estar} \\ 0 & \text{en caso contrario} \end{cases}$$

Mientras que la superficie, el número de baños y el número de habitaciones se definen como en los temas anteriores:

$SQFT_i$  tamaño de la vivienda  $i$ -ésima en pies cuadrados  
 $BEDRMS$  número de habitaciones de la vivienda  $i$ -ésima  
 $BATHS$  número de cuartos de baño de la vivienda  $i$ -ésima

Con todas ellas podemos especificar el siguiente modelo para explicar el precio de la vivienda:

$$PRICE_i = \gamma_1 + \gamma_2 POOL_i + \gamma_3 FAMROOM_i + \gamma_4 FIREPL_i$$

$$+ \beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \dots, 14 \quad (4.26)$$

Donde lo primero a notar es que en el modelo (4.26), afectando a la ordenada, conviven tres conjuntos de variables ficticias con dos categorías cada una, el hecho de tener o no piscina, el hecho de tener o no chimenea y el hecho de tener o no sala de estar, de las cuales sólo se incluye una de cada conjunto y se mantiene el término independiente. Esta forma de definir el modelo es muy cómoda ya que sigue manteniendo los resultados de los modelos con término independiente y permite una fácil interpretación de los coeficientes que acompañan a las variables ficticias. Así,  $\gamma_i \quad i = 2, 3, 4$  recogen el diferencial en el valor esperado de una vivienda por el hecho de poseer la característica correspondiente manteniéndose constante el resto de variables. El resultado de la estimación es:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14  
 Variable dependiente: price

Variable	Coefficiente	Desv. típica	Estadístico $t$	valor p
const	39,0571	89,5397	0,4362	0,6758
pool	53,1958	22,0635	2,4110	0,0467
famroom	-21,344	42,8734	-0,4979	0,6338
firepl	26,1880	53,8454	0,4864	0,6416
sqft	0,146551	0,0301014	4,8686	0,0018
bedrms	-7,0455	28,7363	-0,2452	0,8134
baths	-0,263691	41,4547	-0,0064	0,9951
	Media de la var. dependiente		317,493	
	D.T. de la variable dependiente		88,4982	
	Suma de cuadrados de los residuos		9010,24	
	Desviación típica de los residuos ( $\hat{\sigma}$ )		35,8773	
	$R^2$		0,911504	
	$\bar{R}^2$ corregido		0,835650	
	$F(6, 7)$		12,0166	
	valor p para $F()$		0,00221290	
	Log-verosimilitud		-65,134	
	Criterio de información de Akaike		144,269	
	Criterio de información Bayesiano de Schwarz		148,743	

**La interpretación de los coeficientes estimados es la siguiente:**

- $\hat{\gamma}_1 = 39,057$ : el precio medio estimado de las viviendas sin piscina, baños, habitaciones, sala de estar ni chimenea y con 0 pies cuadrados habitables es de 39.057 dólares.
- $\hat{\gamma}_2 = 53,1958$ : la diferencia estimada en el precio medio de las viviendas con piscina con respecto a las que no la tienen, siendo iguales en el resto de características (pies cuadrados habitables, habitaciones, baños, sala de estar y chimenea) es de 53.196 dólares.
- $\hat{\gamma}_3 = -21,34$ : el precio medio estimado de una vivienda con sala de estar es 21.340 dólares inferior al de una sin sala de estar, siendo idénticas en el resto de características. Esto se debe a que, al mantener constante el número de pies cuadrados de la vivienda y el número de habitaciones y baños, incluir una sala de estar hará que el resto de habitaciones o baños sean de menor tamaño.
- $\hat{\gamma}_4 = 26,188$ : el precio medio estimado de una vivienda con chimenea es 26.188 dólares más caro que el de una sin chimenea, siendo idénticas en el resto de características.
- $\hat{\beta}_1 = 0,147$ : el precio medio estimado de una vivienda se incrementa en 147.000 dólares al aumentar en 1 pie cuadrado habitable su superficie, permaneciendo constantes el número de baños y habitaciones.
- $\hat{\beta}_2 = -7,046$ : el precio medio estimado de una vivienda disminuye en 7.046 dólares al aumentar en 1 el número de habitaciones, permaneciendo constantes el número de baños y los pies cuadrados habitables. Esto se debe a que las habitaciones serán de menor tamaño.

- $\hat{\beta}_3 = -0,264$ : el precio medio estimado de una vivienda disminuye en 264 dólares al aumentar en 1 el número de baños, permaneciendo constantes el número de habitaciones y los pies cuadrados habitables. De nuevo, las habitaciones serán de menor tamaño.

### Contraste de hipótesis

Para contrastar, por ejemplo, que no existen diferencias significativas en el precio medio de la vivienda por el hecho de tener chimenea, se realiza un contraste de significatividad individual de la variable FIREPL. En este caso, observando el p-valor correspondiente, 0,6416, se puede concluir que a un nivel de significación del 5%, no existen diferencias significativas en el precio medio de una vivienda por el hecho de tener chimenea.

Si comparamos los modelos (4.25) y (4.26), ninguna de las variables añadidas en el último es significativa individualmente<sup>8</sup>. Además, el  $\bar{R}^2$  es inferior. El contraste de significatividad conjunta para las variables añadidas se puede realizar con el estadístico  $F$  basado en las sumas de cuadrados residuales de los modelos restringido (modelo (4.25)) y no restringido (modelo (4.26)). En este caso, el resultado es:

Contraste de omisión de variables –

Hipótesis nula: los parámetros son cero para las variables

bedrms

baths

famroom

firepl

Estadístico de contraste:  $F(4, 7) = 0,0864517$

con valor p =  $P(F(4, 7) > 0,0864517) = 0,983881$

por lo que no se rechaza la hipótesis nula de que las variables añadidas al modelo (4.26) son conjuntamente no significativas. Al omitir dichas variables el modelo mejora en cuanto a la significación de sus coeficientes y el  $\bar{R}^2$ . Por tanto, manteniendo las variables POOL y SQFT, la inclusión del resto (FIREPL, FAMROOM, BATHS, BEDRMS) no añade capacidad explicativa al modelo.

#### 4.6.1. Predicción en gretl

Para hacer predicción con *gretl* debemos incorporar los nuevos datos ( $X_p$ ) a la base de datos mediante

*Datos* → *Seleccionar todos*

A continuación, pincharemos la opción

---

<sup>8</sup>Un problema añadido es que tenemos un bajo tamaño muestral,  $T=14$ , y hemos aumentado significativamente el número de parámetros a estimar,  $K=7$ , por lo que tenemos muy pocos grados de libertad.

*Datos → Añadir Observaciones*

indicando el número de observaciones que queremos añadir, en este caso 1. En la fila correspondiente incluimos los valores de las variables explicativas en el periodo de predicción, en este caso la observación  $N + 1$ , incorporando cada observación en la casilla correspondiente. Si no incorporamos el valor para la variable  $Y_i$  que es la que vamos a predecir, gretl nos mostrará un aviso (Atención: había observaciones perdidas). Podemos simplemente ignorarlo y darle a aceptar.

Posteriormente, estimaremos el modelo sin considerar esta nueva observación. Para ello, tenemos que especificar el rango muestral, es decir, en la opción

*Muestra → Establecer rango*

especificaremos del rango de observaciones de la muestra para estimar el modelo, en nuestro caso de la 1 a la  $N$  y elegimos *Aceptar*.

Estimaremos el modelo por MCO y en la ventana de los resultados elegimos

*Análisis → Predicciones*

En la nueva ventana podemos determinar el dominio de predicción, es decir el *Inicio* y *Fin* que en este caso es en ambos la observación número  $N + 1$ , y también cuantas observaciones se quieren representar antes de la predicción.

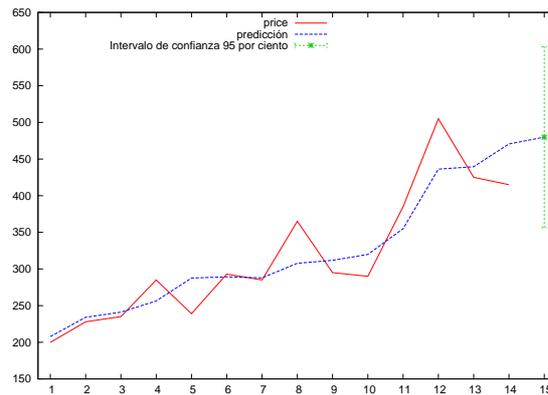
Utilizando los resultados obtenidos en el Ejemplo 5.10 se va a predecir la variable *PRICE*. Los resultados que muestra Gretl son los siguientes:

Para intervalos de confianza 95 %,  $t(10, ,0, 025) = 2, 228$

Observaciones	price	predicción	Desv. Típica	Intervalo de 95 %	
15	500,00	479,91	55,390	356,49	603,32

Estadísticos de evaluación de la predicción

Error medio	20,095
Error cuadrático medio	403,79
Raíz del Error cuadrático medio	20,095
Error absoluto medio	20,095
Porcentaje de error medio	4,0189
Porcentaje de error absoluto medio	4,0189
<i>U</i> de Theil	0



El gráfico que se obtiene junto a los resultados muestra la serie de precios (P) observada en color rojo y estimada con el modelo para las 14 observaciones anteriores a la predicción y la predicción en color azul, junto con su intervalo de confianza en color verde.

La predicción por punto del precio de una vivienda con estas características es de 479,905 miles de euros, mientras que la predicción por intervalo con un nivel de confianza del 95 % es (356,5; 603,3) en miles de euros, por lo que el precio que nos piden, que era de 500 miles de euros por la vivienda, está dentro del intervalo. Este precio para una vivienda de esas características se aceptaría como razonable dado nuestro modelo y la información muestral utilizada para su estimación, con un nivel de confianza del 95 %.

## 4.7. Apéndice

### 4.7.1. Distribuciones que nos interesan

- Distribución de la variable endógena o regresando:

$$Y = X\beta + u$$

$$E(Y) = X\beta$$

$$V(Y) = E[(Y - E(Y))(Y - E(Y))'] = E(uu') = \sigma^2 I_N$$

Dado

$$u \sim N(0, \sigma^2 I_N) \longrightarrow Y \sim N(X\beta, \sigma^2 I_N)$$

- Distribución del vector de estimadores MCO:

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

$$E(\hat{\beta}) = \beta$$

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

Dado

$$\begin{aligned} u \sim N(0, \sigma^2 I_N) &\longrightarrow \hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \\ &\longrightarrow \hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk}) \end{aligned}$$

- Distribución de los residuos:

$$\hat{u} = Mu$$

$$E(\hat{u}) = E(Mu) = ME(u) = 0$$

$$V(\hat{u}) = E[\hat{u}\hat{u}'] = E(Muu'M') = ME(uu')M = \sigma^2 M$$

Dado

$$u \sim N(0, \sigma^2 I_N) \longrightarrow \hat{u} \sim N(0, \sigma^2 M)$$

- Distribución de la SCR:

$$SCR = \hat{u}'\hat{u} = u'Mu$$

Utilizando el resultado 3 del Anexo 4 dado

$$u \sim N(0, \sigma^2 I_N) \longrightarrow \frac{u'Mu}{\sigma^2} \sim \chi_{(N-K)}^2 \longrightarrow \frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi_{(N-K)}^2$$

ya que:  $\hat{u} \sim N(0, \sigma^2 M)$  tenemos

$$\frac{\hat{u}'M^{-1}\hat{u}}{\sigma^2} = \frac{u'MM^{-1}Mu}{\sigma^2} = \frac{u'Mu}{\sigma^2} \sim \chi_{(N-k)}^2$$

## 4.8. Bibliografía del tema

### Referencias bibliográficas básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con *gretl*:

- [1] Ramanathan, R. (2002), *Instructor's Manual to accompany*, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), *Student Solutions Manual*, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

### Referencias Bibliográficas Complementarias:

- [1] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Análisis de regresión con gretl*. Open Course Ware. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Courseisting>).
- [2] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación on-line de la Facultad de C.C. Económicas y Empresariales.
- [3] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones.
- [4] Esteban, MV (2008). *Estadística Actuarial: Regresión Lineal*, Sarriko On Line 3/08. Publicación on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.
- [5] Esteban, M.V. (2007). *Colección de ejercicios y exámenes*. Material docente. Servicio de Publicaciones.
- [6] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). *Ejercicios de Econometría*. Editorial McGraw-Hill.
- [7] Greene, W. (1998), *Análisis Econométrico*, Ed. Prentice Hall, 3ª edición.
- [8] Ramanathan, R. (2002), *Introductory Econometrics with applications*, Ed. South-Western, 5th. edition.
- [9] Verbeek, M. (2004). *A Guide to Modern Econometrics*. Wiley.

## Tema 5

# Modelo de Regresión Lineal General. Diagnóstico y utilización

En este tema vamos a ocuparnos de validar el modelo. Una vez especificado y estimado el modelo de regresión lineal general y realizados los contrastes de interés el modelo puede ser utilizado para la predicción. Esta será más fiable cuanto mejor especificado y estimado esté el modelo. En el Tema 4 nos hemos ocupado de ver las consecuencias de omitir variables relevante e incluir variables irrelevantes y para evitarlo utilizamos los contrastes de significatividad individual y conjunto. En este tema nos ocuparemos de analizar si los coeficientes del modelo son constantes durante todo el periodo muestral.

Por otro lado cuando especificamos las hipótesis básicas de comportamiento, sobre la perturbación supusimos que es homocedástica y no autocorrelada, en este tema estudiaremos como contrastar que efectivamente la perturbación tiene varianza constante y covarianzas cero.

### Competencias a trabajar en estas sesiones:

2. Aplicar la metodología econométrica básica para estimar y validar relaciones económicas en base a la información estadística disponible sobre las variables y utilizando los instrumentos informáticos apropiados.
3. Interpretar razonadamente los resultados obtenidos en la estimación y validación del modelo econométrico con el objetivo de elaborar informes económicos.
4. Presentar de forma clara y concisa, tanto oralmente como por escrito, las conclusiones obtenidas en una aplicación empírica.

### Al final de este tema deberíais ser capaces de:

1. Conocer cuando un modelo es lineal en parámetros y/o en variables.
2. Transformar un modelo no lineal en parámetros en un modelo estimable por MCO.

3. Aplicar el contraste de Ramsey en el contexto adecuado.
4. Explicar que se entiende por un cambio estructural en un modelo econométrico.
5. Contrastar la constancia de los coeficientes de un modelo de regresión lineal general.
6. Explicar que se entiende por un modelo de regresión lineal con heterocedasticidad, cómo modelizar esta característica en el término de perturbación del modelo y sus implicaciones en su matriz de varianzas y covarianzas.
7. Analizar gráficamente la posible existencia de heterocedasticidad y saber contrastarla utilizando el estadístico de White.
8. Describir las propiedades del estimador MCO bajo heterocedasticidad.
9. Explicar que se entiende por un modelo de regresión lineal con autocorrelación, cómo modelizar esta característica en el término de perturbación del modelo y sus implicaciones en su matriz de varianzas y covarianzas.
10. Analizar gráficamente la posible existencia de autocorrelación y saber contrastarla utilizando el estadístico de Durbin Watson.
11. Describir las propiedades del estimador MCO bajo autocorrelación.
12. Utilizar el software gretl para validar el modelo.

### **Bibliografía Recomendada:**

Al final del tema tenéis recogida la bibliografía correspondiente. En particular se os recomienda leer los capítulos correspondientes a la bibliografía básica detallados a continuación:

- Gujarati, D. y Porter, D.C. (2010). Parte I : cap. 6;Parte II: Caps 11 y 12.
- Ramanathan, R. (2002). Caps. 6, 8 y 9.
- Wooldridge, J.M. (2006). Cap. 6.

## 5.1. Forma funcional

La elección de la forma funcional que recoge la relación existente entre la variable dependiente y las variables explicativas es un aspecto de la especificación de un modelo muy importante en el análisis económico. De hecho, la teoría económica no siempre propone relaciones lineales entre variables de interés. Es el caso, por ejemplo, de la función de consumo de un bien que aumenta con la renta pero no de forma indefinida ni a ritmo constante sino, en general, a una tasa decreciente, o de las funciones de costes marginales que suelen tener forma de  $U$ , veáse la Figura 5.1.

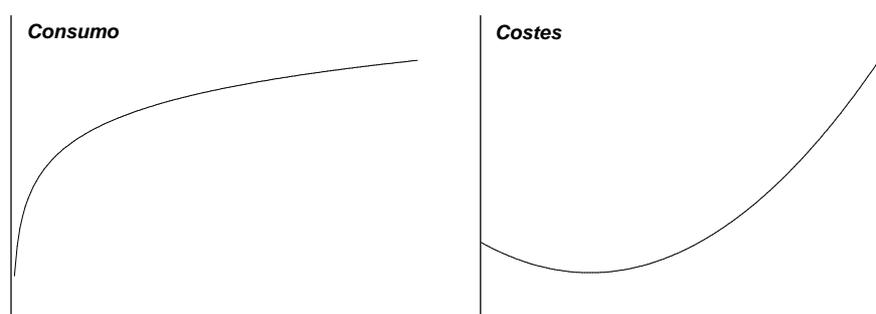


Figura 5.1: Relaciones económicas no lineales

En el Tema 3 se hizo énfasis en el hecho de que el supuesto de *linealidad* del modelo de regresión no implica una relación lineal entre las variables sino un modelo en el que los parámetros entran de forma lineal. Por “lineal en los parámetros” se entiende que los parámetros no se multiplican entre sí, no están elevados a potencias, etc. Sin embargo, como se vió en dicho tema las variables, tanto regresando como regresores, sí se pueden transformar para obtener al final un modelo de regresión lineal que satisfaga los supuestos clásicos. Este hecho hace que el modelo de regresión lineal sea bastante flexible y se pueda utilizar para modelar relaciones entre variables económicas no lineales. Así, tanto la función de consumo como la función de costes marginales de la Figura 5.1 se pueden modelizar utilizando formas funcionales sencillas no lineales en las variables. En el caso de la función de consumo, el supuesto de rendimientos decrecientes se puede representar mediante modelos logarítmicos o semilogarítmicos del tipo:

$$\ln C = \alpha + \beta \ln R + u \quad (5.1)$$

$$C = \alpha + \beta \ln R + u \quad (5.2)$$

y las funciones de costes totales se pueden representar mediante funciones polinómicas:

$$CM = \beta_1 + \beta_2 Q + \beta_3 Q^2 + u \quad (5.3)$$

Los modelos (5.1), (5.2) y (5.3) cumplen el supuesto de linealidad porque son lineales en los parámetros y se pueden analizar dentro del marco del MRLG. Ahora bien, como no son modelos lineales en las variables, el efecto marginal del regresor sobre la variable dependiente no va a ser constante. Por

ejemplo, en el modelo (5.3), el efecto marginal de un incremento unitario de la producción sobre los costes marginales viene dado por:

$$\frac{\partial E(CT)}{\partial Q} = \beta_2 + 2\beta_3 Q$$

Este resultado implica que la pendiente de la función de costes marginales no es constante sino que es una función lineal de  $Q$  que involucra a los parámetros  $\beta_2$  y  $\beta_3$ .

Otra forma de modelar relaciones no lineales entre las variables explicativas y el regresando es incluir términos de interacción, es decir, el producto de varios regresores del modelo. Consideremos, por ejemplo, el siguiente modelo:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 \times X_3) + u$$

Este modelo es lineal en los parámetros, por lo que cumple el supuesto de linealidad. El efecto marginal de  $X_2$  sobre  $Y$  es:

$$\frac{\partial E(Y)}{\partial X_2} = \beta_2 + \beta_4 X_3$$

de forma que el incremento esperado en  $Y$  ante un incremento unitario en  $X_2$  no es constante sino que depende del valor de  $X_3$ .

Los modelos que no cumplen el supuesto de *linealidad* se pueden clasificar en dos grupos. En el primer grupo se encuentran los modelos que no son lineales en los parámetros pero que se pueden linealizar mediante alguna transformación. En este grupo entra por ejemplo la función de producción Cobb-Douglas que no es lineal ni en las variables ni en los parámetros, pero tomando logaritmos se obtiene una función que no es lineal en las variables pero sí es lineal en los parámetros. El segundo grupo lo forman los modelos que no son lineales en los parámetros y que no se pueden linealizar mediante ninguna transformación, por ejemplo,

$$Y = \beta_1 + X_1^{\beta_2\beta_3} + X_2^{\beta_2} + u$$

Este tipo de modelos se estima por mínimos cuadrados no lineales.

### Contraste de Ramsey

El contraste de Ramsey (1969) está diseñado para detectar si la forma funcional está correctamente especificada o no. Una forma sencilla de detectar que la forma funcional de este modelo está mal especificada y la relación no es lineal entre las variables, sería incluir términos no lineales en el modelo anterior y contrastar su significatividad. Ramsey propone añadir términos no lineales del tipo  $\hat{Y}_t^2$ ,  $\hat{Y}_t^3$ , ..., donde  $\hat{Y}_t$  son los valores ajustados por la recta de regresión muestral MCO. De forma que los términos  $\hat{Y}_t^2$ ,  $\hat{Y}_t^3$ , ..., son funciones no lineales de todos los regresores. La lógica del contraste se basa en que, si el modelo no está bien especificado, la introducción de estos términos no lineales mejoraría el ajuste del modelo.

Los pasos para implementar el contraste de Ramsey son los siguientes. En primer lugar, se estima el modelo que incluye tantos términos no lineales como se desee, habitualmente, dos:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \gamma_1 \hat{Y}_t^2 + \gamma_2 \hat{Y}_t^3 + u_t \quad (5.4)$$

La hipótesis nula de correcta especificación del modelo se expresaría en función de los parámetros del modelo (5.4) como

$$H_0 : \gamma_1 = \gamma_2 = 0$$

$$\gamma_1 \neq 0 \quad \text{y/o} \quad \gamma_2 \neq 0$$

Si no se rechaza la hipótesis nula, significa que no se ha detectado mala especificación del modelo. Si, por el contrario, se rechaza, el modelo no es adecuado y se puede mejorar. Ahora bien, una de las limitaciones de este contraste es que no da indicaciones sobre cuál sería la forma funcional apropiada para el modelo. *Gretl* permite obtener el contraste de Ramsey automáticamente.

## 5.2. Sobre constancia de los coeficientes: contraste de cambio estructural

En ocasiones puede ocurrir que la relación entre la variable dependiente y los regresores cambie a lo largo del periodo muestral, es decir, puede que exista un cambio estructural. Por ejemplo, si estamos analizando el consumo de tabaco y durante el período muestral se ha producido una campaña de salud pública informando sobre los peligros que conlleva el consumo de tabaco, podemos pensar que tras dicha campaña el comportamiento de la demanda de tabaco haya cambiado, reduciéndose significativamente. Si esto ocurre no podemos especificar una única función de demanda para todo el período muestral si no que deberíamos especificar dos funciones, una hasta la campaña antitabaco y otra para el período siguiente. Por tanto ante sospechas de que exista un cambio estructural debemos de contrastar la estabilidad de los parámetros de nuestra relación.

El contraste de cambio estructural, llamado habitualmente contraste de Chow puede realizarse de manera sencilla mediante el estadístico de sumas de cuadrados de los residuos sin más que especificar adecuadamente el modelo restringido y el no restringido. También podemos llevarlo a cabo utilizando variables ficticias. Hay cambio estructural cuando el modelo

$$Y_t = \alpha_1 + \beta_1 X_t + u_t \quad t = 1, 2, \dots, t_1 < T \quad (5.5)$$

y el modelo

$$Y_t = \alpha_2 + \beta_2 X_t + u_t \quad t = t_1, t_2, \dots, T \quad (5.6)$$

se verifica sólo desde el momento  $t_1$  hasta  $T$ .

En este caso no podemos escribir una única ecuación del tipo:

$$Y_t = \alpha + \beta X_t + u_t \quad t = 1, 2, \dots, T \quad (5.7)$$

ya que no se verifica durante todo el periodo muestral y nuestro modelo estaría formado por las ecuaciones (5.5) y (5.6). Si existe cambio estructural rechazaríamos  $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$  donde  $q = 2$ . Este contraste podemos llevarlo a cabo utilizando el estadístico  $F$  basado en las sumas de cuadrados de los residuos siendo en este caso el modelo restringido el recogido en la ecuación (5.7) mientras

que el modelo no restringido está constituido por las ecuaciones (5.5) y (5.6). Matricialmente:

$$\text{Modelo no restringido} \Leftrightarrow \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 & 0 & 0 \\ 0 & 0 & i_2 & X_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \Rightarrow \hat{u}'\hat{u} = \hat{u}'_1\hat{u}_1 + \hat{u}'_2\hat{u}_2$$

$$\text{Modelo restringido} \Leftrightarrow \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 \\ i_2 & X_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \Rightarrow \hat{u}'_r\hat{u}_r$$

bajo el supuesto  $u_1 \sim N(0, \sigma_1^2)$ ,  $u_2 \sim N(0, \sigma_2^2)$  el estadístico de contraste y distribución es:

$$\frac{\hat{u}'_r\hat{u}_r - \hat{u}'\hat{u}/2}{\hat{u}'\hat{u}/T_1 + T_2 - 4} \stackrel{H_0}{\sim} F_{(2, T_1+T_2-4)}$$

Notar que para la validez del contraste además de suponer normalidad en las perturbaciones se supone también que  $\sigma_1^2 = \sigma_2^2$ .

### 5.3. Sobre las perturbaciones: contrastes de heterocedasticidad y ausencia de correlación

#### 5.3.1. Contraste de heterocedasticidad

Hasta el momento uno de los supuestos básicos del modelo de regresión lineal es que la varianza de cada término de perturbación  $u_i$  condicionada a los valores de las variables explicativas, es constante e igual a  $\sigma^2$ . Llamábamos a este supuesto *homocedasticidad* y lo denotábamos:  $V(u_i) = \sigma^2$  ó lo que es igual  $E(u_i^2) = \sigma^2 \quad \forall i$ . La varianza  $\sigma^2$  es una medida de dispersión de  $u_i$  alrededor de su media,  $E(u_i) = 0$ , o equivalentemente, una medida de dispersión de la variable dependiente  $Y_i$  alrededor de su media  $\beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ . Así, homocedasticidad significa que la dispersión es la misma a través de todas las observaciones.

Supongamos que disponemos de observaciones sobre consumo y renta para un conjunto de familias, en un año determinado. Las familias con rentas bajas no tienen mucha flexibilidad en sus gastos, en general el grueso de la misma se gastará en cosas básicas, por ello la forma de consumo entre familias de renta baja no variará demasiado. Sin embargo, las familias de rentas altas tienen más posibilidades de consumo, ser grandes consumidores o ahorradores o llevar un gasto equilibrado. En cualquier caso su consumo puede ser muy distinto entre sí por lo que pueden tener una gran dispersión alrededor de su consumo medio mientras que las familias con rentas bajas no. En esta situación suponer que existe homocedasticidad no es sensato, deberíamos suponer que existe heterocedasticidad.

Llamamos heterocedasticidad al caso en que la varianza del término de error varía a través del tiempo si miramos a series temporales,  $V(u_t) = \sigma_t^2$ , o cambia de una observación a otra si miramos datos de sección cruzada, (familias, países, etc.),  $Var(u_i) = \sigma_i^2$ . Seguimos suponiendo que no existe autocorrelación entre perturbaciones por lo que sólo consideramos la existencia de heterocedasticidad. La existencia de heterocedasticidad es más habitual en datos de sección cruzada.

En la Figura 5.2 se puede apreciar la diferencia entre el comportamiento de las perturbaciones homocedásticas, a la izquierda y heterocedásticas, a la derecha. En la figura de la izquierda se puede observar que la varianza condicional de  $Y_i$  a las  $X_i$  permanece igual sin importar los valores que tome la variable  $X$ . Recordar que la varianza condicional de  $Y_i$  es la misma que la de  $u_i$ , por tanto, en el gráfico estamos observando cómo la varianza de la perturbación permanece constante independientemente del valor que tome el regresor. En la figura de la derecha se puede observar que la varianza de  $Y_i$  aumenta a medida que  $X_i$  aumenta y por tanto hay heterocedasticidad:

$$E(u_i^2) = \sigma_i^2$$

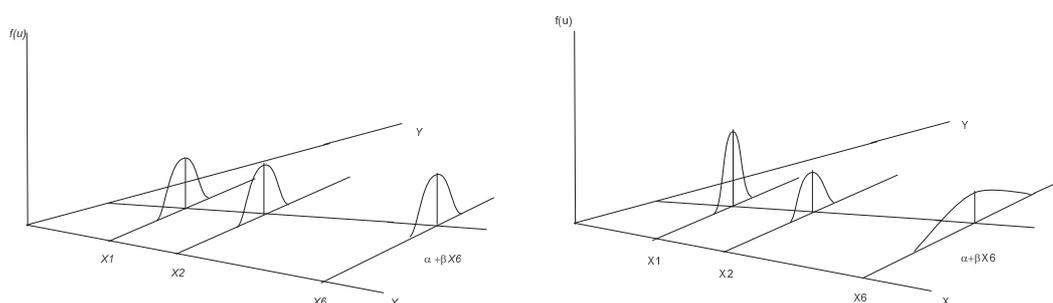


Figura 5.2: Perturbaciones homocedásticas versus heterocedásticas

La existencia de heterocedasticidad puede aparecer en numerosas aplicaciones económicas sin embargo, es más habitual en datos de sección cruzada. A continuación veremos algunas situaciones en las cuales las varianzas de  $u_i$  pueden no ser constantes.

- **En datos de sección cruzada.**

**Ejemplo 5.1** Supongamos que tenemos datos para diferentes comunidades autónomas españolas en el año 2005 sobre gasto sanitario agregado,  $GS$ , renta personal disponible,  $R$ , el porcentaje de población que supera los 65 años,  $SEN$  y población,  $POP$ , con los que estimar el siguiente modelo:

$$GS_i = \beta_1 + \beta_2 R_i + \beta_3 SEN_i + \beta_4 POP_i + u_i \quad i = 1, \dots, N \quad (5.8)$$

Las comunidades con más población y/o mayor porcentaje de población con edad superior a 65 años tendrán mayor gasto sanitario que aquellas con menor población o más joven. En esta situación suponer que la dispersión de los gastos sanitarios es la misma para todas las comunidades con distinto nivel de población y composición de la misma no es realista, y se debería proponer que la varianza de la perturbación sea heterocedástica  $Var(u_i) = \sigma_i^2$ , permitiendo por ejemplo que varíe en función creciente con la población, es decir,  $\sigma_i^2 = \sigma^2 POP_i$ . Incluso podemos pensar que varíe en función creciente con el porcentaje de población mayor de 65 años, en cuyo caso propondríamos  $Var(u_i) = \sigma^2 SEN_i$  o con ambas variables, por lo que la forma funcional pudiera ser  $Var(u_i) = \sigma^2(a POP_i + b SEN_i)$ .

**Ejemplo 5.2** Un ejemplo recurrente para mostrar la heterocedasticidad es el estudio de la relación entre consumo y renta. Supongamos que tenemos datos sobre renta,  $R$ , y gasto en consumo,  $C$ , para  $N$  familias, con los que estimar el modelo:

$$C_i = \beta_1 + \beta_2 R_i + u_i \quad i = 1, \dots, N \quad (5.9)$$

Las familias con mayor renta, una vez satisfechas sus necesidades primordiales tienen mayores posibilidades de decidir cuánto ahorrar y cuánto consumir, por lo que es habitual encontrar una mayor variabilidad en el gasto realizado por familias de renta alta que por familias de renta baja. En esta situación suponer que la dispersión de los gastos de consumo es la misma para todas las familias con distinto nivel de renta no es realista y se debería proponer que la varianza de la perturbación sea heterocedástica  $Var(u_i) = \sigma_i^2$ , permitiendo por ejemplo que varíe en función creciente con la renta de las familias, es decir,  $\sigma_i^2 = \sigma^2 R_i$ .

**Ejemplo 5.3** Un fenómeno parecido ocurre con las empresas que deben decidir qué porcentaje de sus beneficios,  $B$ , deben repartir como dividendos,  $D$ . Las empresas con mayores beneficios tienen un margen de decisión muy superior al fijar su política de dividendos. Al estimar el modelo:

$$D_i = \beta_1 + \beta_2 B_i + u_i \quad i = 1, \dots, N \quad (5.10)$$

cabría esperar que la varianza de  $u_i$  dependa del nivel de beneficios de la empresa  $i$ -ésima y podríamos proponer que por ejemplo,  $E(u_i^2) = \sigma_i^2 = \sigma^2 B_i$ .

- La heterocedasticidad también puede aparecer **como consecuencia de la agregación de datos**. En este caso la varianza puede depender del número de observaciones del grupo.

**Ejemplo 5.4** Supongamos un investigador que desea estimar los coeficientes del siguiente modelo:

$$Y_j = \beta_1 + \beta_2 X_j + u_j \quad j = 1, \dots, N \quad (5.11)$$

donde  $u_j \sim N(0, \sigma^2)$ , es decir, la varianza de la perturbación es homocedástica. Supongamos que el número de observaciones  $N$  es tal que aconseja agrupar las observaciones en  $m$ -grupos de  $n_i$  observaciones cada uno. Supongamos que como observación del grupo  $i$ -ésimo se toma la media aritmética dentro del grupo. El modelo a estimar sería:

$$\bar{Y}_i = \beta_1 + \beta_2 \bar{X}_i + \bar{u}_i \quad i = 1, \dots, m \quad (5.12)$$

y la nueva perturbación  $\bar{u}_i$  seguirá teniendo media cero, pero su varianza no será constante ya que dependerá del número de observaciones dentro del grupo,

$$Var(\bar{u}_i) = \frac{\sigma^2}{n_i} \quad i = 1, \dots, m.$$

Si el número de observaciones dentro del grupo es el mismo en todos los grupos la varianza de la perturbación  $\bar{u}_i$  es homocedástica.

- Otro caso sería la **existencia de un cambio estructural en varianza** recogido por una variable ficticia en la varianza de la perturbación.

**Ejemplo 5.5** Supongamos que se desea estudiar la relación entre producción,  $Y$ , y mano de obra,  $X$ , para un conjunto de 20 trabajadores de los cuales 10 son mujeres y el resto hombres. Si suponemos que la variabilidad de la producción es distinta para los hombres que para las mujeres nuestro modelo a estimar sería:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad i = 1, \dots, 20 \quad (5.13)$$

donde  $u_i \sim (0, \alpha_1 + \alpha_2 D_i)$  siendo  $D_i$  una variable ficticia que toma valor la unidad si la observación corresponde a una mujer y cero en el caso contrario. En este caso:

$$\begin{aligned} \text{Var}(u_i) &= \alpha_1 + \alpha_2 && \text{para las observaciones correspondientes a las mujeres} \\ \text{Var}(u_i) &= \alpha_1 && \text{para las observaciones correspondientes a los hombres} \end{aligned}$$

Suponiendo que las primeras diez observaciones corresponden a mujeres, la matriz de varianzas y covarianzas del vector de perturbaciones sería la siguiente:

$$E(uu') = \begin{bmatrix} (\alpha_1 + \alpha_2)I_{10} & 0 \\ 0 & \alpha_1 I_{10} \end{bmatrix}$$

### Consecuencias de ignorar la heterocedasticidad

Vamos a analizar las consecuencias de utilizar el estimador MCO en presencia de heterocedasticidad:

- **En las propiedades del estimador MCO:** El estimador MCO bajo heterocedasticidad sigue siendo una combinación lineal de las perturbaciones. También sigue siendo insesgado ya que la media de la perturbación es cero. Sin embargo, no va a ser de varianza mínima ya que la matriz de varianzas y covarianzas  $\sigma^2(X'X)^{-1}$  obtenida en el Tema 3 es mínima bajo las hipótesis básicas. Ahora, sin embargo, éstas no se cumplen: estamos considerando el supuesto de heterocedasticidad por tanto  $E(u_i^2) \neq \sigma^2$ , el Teorema de Gauss-Markov no se cumple y el estimador no es de varianza mínima. Ahora la matriz de varianzas y covarianzas de los coeficientes obtenida bajo este supuesto no vendrá dada por la expresión  $\sigma^2(X'X)^{-1}$  y por tanto no será mínima. El estimador no es eficiente.
- **En los contrastes de hipótesis:** Una forma sencilla de pensar en las consecuencias sobre los contrastes de hipótesis es pensar que dado que el estimador no es el mejor de los posibles la inferencia realizada con el mismo no será fiable.

Formalmente lo que está ocurriendo es que el estimador de  $\sigma^2$  propuesto en el Tema 3 ahora no es insesgado por lo que los estadísticos de contraste habituales no tendrán las distribuciones  $t$  y  $F$  habituales. Por tanto, los contrastes no son válidos.

La existencia de heterocedasticidad en  $u_i$  tiene consecuencias en los estimadores MCO, en concreto ya no son los estimadores de menor varianza entre los estimadores lineales e insesgados. Existe otro estimador, el estimador de Mínimos Cuadrados Generalizados que es el de menor varianza entre los lineales e insesgados y para el cual la inferencia es válida. Las consecuencias y soluciones del

problema no forman parte del contenido de este curso. Sin embargo, en la siguiente sección vamos a aprender a detectar la existencia de heterocedasticidad con un estadístico de contraste sencillo y que aparece por defecto en los resultados de estimación MCO de *gretl*. En cursos más avanzados aprenderéis a solucionar el problema.

### Detección de la heterocedasticidad

Sabemos que en presencia de heterocedasticidad el estimador MCO es ineficiente, y los contrastes de hipótesis no son válidos por ello es importante detectar la posible existencia de heterocedasticidad. La determinación de la existencia de heterocedasticidad sólo podremos conseguirla aplicando un test de contraste para heterocedasticidad, sin embargo podemos aproximarnos gráficamente al problema realizando un estudio visual de los residuos del modelo. Los residuos MCO son un estimador insesgado de  $u_i$  aún en presencia de heterocedasticidad. Usaremos el residuo al cuadrado como aproximación al comportamiento de la varianza de la perturbación. Para ver si puede existir un problema de heterocedasticidad podemos empezar por dibujar el cuadrado de los residuos MCO contra la variable de la cual sospechamos que depende  $\sigma^2$ , es decir, que sospechamos causa la heterocedasticidad

Nuestro objetivo es claro: **Detectar la existencia de heterocedasticidad en las perturbaciones de un modelo.** La primera aproximación al objetivo es el estudio de los gráficos de residuos y de las variables del modelo.

#### 5.3.2. Detección gráfica.

La aplicación del estimador de MCG y algunos contrastes de heterocedasticidad requieren conocer la forma funcional de la varianza de la perturbación. Si suponemos que la varianza de la perturbación depende de uno o más regresores, u otras variables conocidas, un instrumento adecuado para aproximarnos a la misma sería llevar a cabo un análisis de los residuos MCO donde no hemos tenido en cuenta la existencia de heterocedasticidad. Aunque  $\hat{u}_{MCO,i}$  no es lo mismo que  $u_i$  la detección de patrones sistemáticos en la variabilidad de los residuos MCO nos indicará la posible existencia de heterocedasticidad en las perturbaciones. Además, puede indicarnos una posible forma funcional de la misma.

Consideramos el modelo (5.8) recogido en el Ejemplo 6.1:

$$GS_i = \beta_1 + \beta_2 R_i + \beta_3 SEN_i + \beta_4 POP_i + u_i \quad i = 1, \dots, N$$

donde suponemos  $E(u_i) = 0 \quad \forall i$  y  $E(u_i u_j) = 0 \quad \forall i, j \quad i \neq j$ . Si sospechamos que  $u_i$  es heterocedástica debido a la variable  $POP$ , podemos intentar detectar la existencia de heterocedasticidad en las perturbaciones del modelo ayudándonos del gráfico de los residuos MCO,  $(\hat{u}_{MCO,i})$ , frente a la variable  $POP_i$ .

Si el gráfico es como el recogido en la Figura 5.3 pensaremos que la variabilidad de los residuos  $\hat{u}_{MCO,i}$  se incrementan con  $POP_i$  y que el incremento es directamente proporcional. Así, podríamos

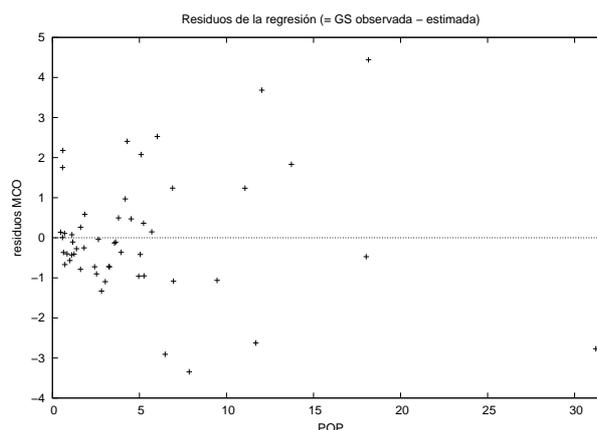


Figura 5.3: Residuos MCO versus  $POP$

proponer, por ejemplo:

$$E(u_i^2) = \sigma^2 POP_i \quad i = 1, 2, \dots, N$$

Si el gráfico de los residuos MCO frente a  $POP$  hubiera sido como el recogido en la Figura 5.4 supondríamos que el aumento en la varianza de  $u_i$  es inversamente proporcional a  $POP_i$  y propondríamos:

$$E(u_i^2) = \sigma^2 POP_i^{-1} \quad i = 1, 2, \dots, N$$

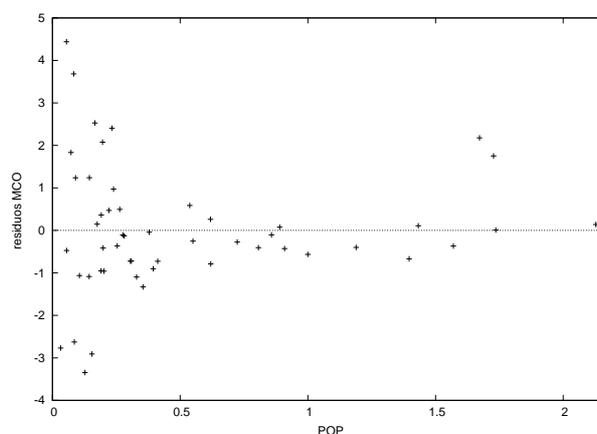


Figura 5.4: Residuos MCO versus  $POP$

También podemos optar por dibujar la serie de los residuos al cuadrado frente a la variable que creemos causa la heterocedasticidad como se muestra en la Figura 5.5. En el gráfico de la izquierda se muestran los pares  $(SEN_i, \hat{u}_{MCO,i}^2)$ , en el gráfico de la derecha se muestran los pares  $(SEN_i, \hat{u}_{MCO,i}^2)$ . Ambos gráficos muestran la misma información, muestran que la variabilidad de los residuos se incrementa con  $SEN$  y podríamos proponer, por ejemplo  $Var(u_i) = E(u_i^2) = \sigma^2 SEN_i$ .

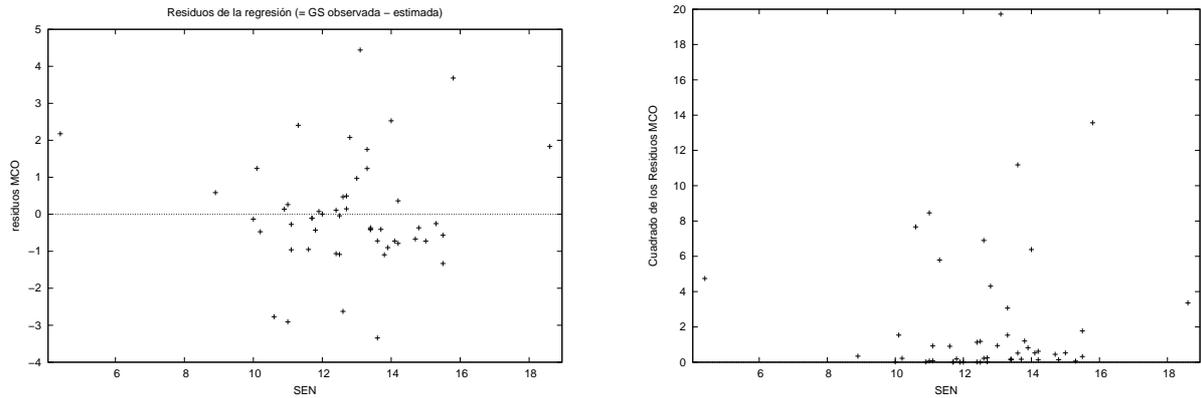


Figura 5.5: Residuos MCO y sus cuadrados versus SEN

En general a priori no se conocerá cuál de las variables exógenas genera la heterocedasticidad por lo que resulta aconsejable estudiar los gráficos de los residuos de MCO, contraponiéndolos a cada una de las variables exógenas del modelo, como estamos haciendo al estudiar los residuos frente a  $POP_i$  y frente a  $SEN_i$ . Notar que ambas variables parecen afectar a la varianza de la perturbación, por ello estaría justificado proponer  $Var(u_i) = (a POP_i + b SEN_i)$ , donde  $a$  y  $b$  son desconocidos y el factor de escala es la unidad,  $\sigma^2 = 1$ .

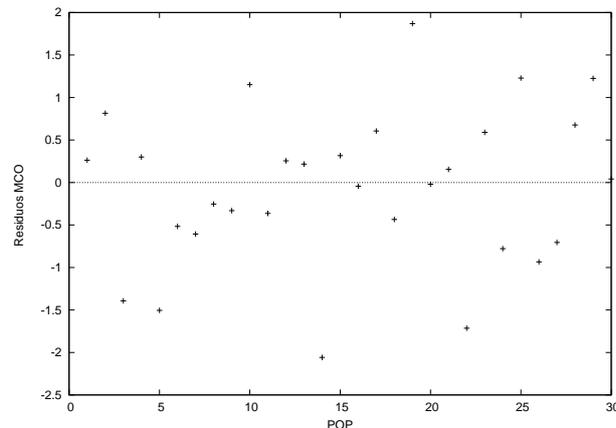


Figura 5.6: Perturbaciones homocedásticas

Si la gráfica entre  $\hat{u}_{MCO,i}$  y  $POP_i$  hubiera resultado como la de la Figura 5.6, concluiríamos que la varianza de la perturbación no depende de  $POP_i$  ya que no se aprecia ningún patrón de comportamiento y parece que hay una distribución aleatoria de los pares  $(POP_i, \hat{u}_i)$ . En esta situación procede analizar los residuos frente al resto de regresores del modelo.

Las formas anteriores no son las únicas. Si recordamos, en el Ejemplo 3.6 se suponía una situación donde hombres y mujeres en una empresa tenían diferente productividad y se suponía que  $Var(u_i) = \alpha_1 + \alpha_2 D_i$  siendo  $D_i$  una variable ficticia que toma valor uno si la observación corresponde a una mujer y cero en caso contrario. En esta situación esperaríamos un gráfico como el recogido en la

Figura 5.7 donde claramente la dispersión de los residuos para las mujeres es mucho mayor que para los hombres.

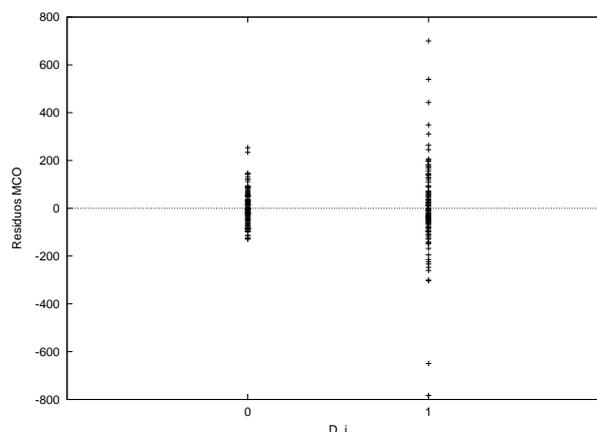


Figura 5.7: Residuos MCO frente a una variable ficticia

Como conclusión diremos que al analizar los gráficos de la relación residuos MCO, o sus cuadrados, con cada uno de los regresores lo que intentaremos detectar visualmente es un crecimiento o decrecimiento en la variabilidad de los residuos con respecto a la variable en cuestión.

Sin embargo el estudio gráfico de los residuos no es determinativo. Para determinar si existe o no heterocedasticidad tendremos que realizar un contraste de existencia de heterocedasticidad con un estadístico adecuado. Estadísticos de contraste de existencia de heterocedasticidad hay muchos y unos se adecúan más a unas situaciones que otros y en general necesitan suponer una forma funcional para  $\sigma_i^2$ . El análisis gráfico no es una pérdida de tiempo ya que la relación entre  $X_{ki}$  y  $\hat{u}_{MCO,i}$  nos indicará una posible forma funcional (de heterocedasticidad) para la varianza de la perturbación y puede indicarnos cuál es el test de contraste más adecuado. En este tema vamos a estudiar un único test de heterocedasticidad que tiene carácter general y no exige supuestos sobre el comportamiento de  $\sigma_i^2$ . Además *gretl* lo proporciona directamente.

### 5.3.3. Contraste de White

El contraste de heterocedasticidad propuesto por White en 1980 es un contraste paramétrico, de carácter general, que no precisa especificar la forma que puede adoptar la heterocedasticidad. En este sentido puede calificarse de robusto. Antes de aplicar el contraste con *gretl* vamos a desarrollar paso a paso el contraste para entender su mecanismo. Para la ilustración vamos a suponer que queremos contrastar la existencia de heterocedasticidad en el modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (5.14)$$

Se procede de la forma siguiente:

1. Estimamos por MCO el modelo original y calculamos los residuos de MCO,  $\hat{u}_{MCO,i}$ .

2. Estimamos la regresión auxiliar: el cuadrado de los residuos mínimo-cuadráticos de la regresión anterior, sobre una constante, los regresores del modelo original, sus cuadrados y productos cruzados de segundo orden, evitando los redundantes:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + \omega_i \quad (5.15)$$

Contrastar la hipótesis nula de homocedasticidad es equivalente a contrastar que todos los coeficientes de esta regresión, exceptuando el término independiente son cero. Es decir:

$$H_0 : \alpha_2 = \alpha_3 = \dots = \alpha_6 = 0$$

3. El estadístico de contraste es  $\lambda = NR^2$  donde  $R^2$  es el coeficiente de determinación de la regresión auxiliar (5.15). Rechazamos  $H_0$  si  $NR^2 > \chi_{(p)}^2$  siendo  $p$  el número de coeficientes en la regresión auxiliar sin incluir el término independiente, en el ejemplo  $p = 5$ .

Observaciones:

1. Este contraste es muy flexible ya que no especifica la forma funcional de heterocedasticidad, pero por otro lado, si se rechaza la hipótesis nula de homocedasticidad no indica cuál puede ser la dirección a seguir.
2. A la hora de incluir los regresores de la regresión auxiliar debemos ser muy cuidadosos para no incurrir en multicolinealidad exacta, por ejemplo en el caso de las variables ficticias con valores 0 y 1, en este caso el cuadrado de la variable coincide con ella misma.
3. También pueden surgir problemas en modelos con un alto número de regresores que puede conllevar que en la regresión auxiliar el número de variables sea tal que no supere al número de observaciones y nos quedemos sin grados de libertad. Si éste es el caso podemos optar por regresar el cuadrado de los residuos MCO sobre  $\hat{Y}_i$  y  $\hat{Y}_i^2$  ya que  $\hat{Y}_i$  es el ajuste de  $Y_i$  usando el estimador MCO con todos los regresores originales.
4. El contraste de White puede recoger otro tipo de problemas de mala especificación de la parte sistemática, omisión de variables relevantes, mala forma funcional etc. Esto es positivo si se identifica cuál es el problema, en caso contrario, la solución que se tome puede estar equivocada. Si la detección de heterocedasticidad se debe a un problema de mala especificación la solución pasa por especificar correctamente el modelo.

### 5.3.4. Contraste de ausencia de correlación temporal

En el modelo de regresión, el término de perturbación engloba aquellos factores que determinando la variable endógena, no están recogidos en la parte sistemática del modelo. Estos factores pueden ser innovaciones, errores de medida en la variable endógena, variables omitidas, etc. Hasta el momento uno de los supuestos básicos del modelo de regresión lineal es que la covarianza entre perturbaciones de distintos periodos es cero. Sin embargo, si estos factores están correlacionados en el tiempo o en el espacio, entonces no se satisface la hipótesis de NO autocorrelación que escribíamos como

$E(u_t u_s) = 0 \quad \forall t, s \quad t \neq s$ . Este fenómeno se conoce con el nombre de **autocorrelación**: *correlación serial*, en el caso de series temporales y *correlación espacial* en el caso de datos de sección cruzada. En general es más habitual en datos de serie temporal.

### Concepto de autocorrelación:

Existe autocorrelación cuando el término de error de un modelo econométrico está correlacionado consigo mismo. Es decir, la covarianza entre las perturbaciones es distinta de cero para diferentes momentos del tiempo (o entre distintos individuos)<sup>1</sup> :

$$E(u_t u_s) \neq 0 \quad \forall t, s \quad t \neq s$$

Esta dinámica, aunque no sea relevante en media, refleja un patrón sistemático que tenemos que considerar a la hora de estimar el modelo.

La existencia de autocorrelación supone el incumplimiento de una de las hipótesis básicas sobre la perturbación de forma similar a la existencia de heterocedasticidad. Esta afecta a las varianzas mientras que la autocorrelación afecta a las covarianzas. En cualquier caso, las consecuencias sobre el estimador MCO son las mismas: el estimador no es de varianza mínima, aunque sigue siendo lineal e insesgado. Los contrastes de hipótesis no son válidos por las mismas razones que en el supuesto de heterocedasticidad. Si sospechamos que en un modelo la perturbación está autocorrelada, primero deberíamos cerciorarnos realizando un contraste y en el caso de que ésta exista es importante estimar el modelo bajo estos nuevos supuestos con un estimador alternativo a MCO que sea de varianza mínima y válido para hacer inferencia. Este estimador queda fuera del contenido de este tema, sin embargo, aprenderemos a detectar el problema mediante un contraste y estudiaremos un proceso sencillo para recoger el comportamiento de la perturbación bajo autocorrelación: el proceso autorregresivo de primer orden.

### Causas de autocorrelación

Como decíamos al iniciar este capítulo el término de perturbación de un modelo engloba aquellos factores que, determinando la variable endógena, no están recogidos en la parte sistemática del modelo. Factores como variables omitidas, mala especificación de la forma funcional o errores de medida, entre otros, son causa de autocorrelación. Repasaremos algunos de ellos:

- **Shocks aleatorios prolongados**

Sea el modelo:

$$R_t = \beta_1 + \beta_2 RM_t + u_t \quad t = 1, 2, \dots, T$$

donde  $R_t$  es la rentabilidad de un activo en el periodo  $t$  y  $RM_t$  es la rentabilidad del mercado en dicho periodo  $t$ . Si en un momento dado se produce una caída del mercado, la rentabilidad

---

<sup>1</sup>No es preciso que  $u_t$  esté correlacionada consigo misma en cada dos instantes distintos del tiempo, sino que basta que la correlación se extienda a algunos periodos.

del activo se verá afectada a la baja y como consecuencia la rentabilidad obtenida será menor que la esperada. Este efecto se prolongará en el tiempo hasta que poco a poco los inversores recuperen la confianza y el mercado vuelva a estabilizarse. El shock se recogerá en el término de perturbación. Si por ejemplo, la caída se produce en  $(t-1)$ , lo que estamos diciendo es que la perturbación en  $t$  dependerá de lo ocurrido en  $(t-1)$  vía  $u_{t-1}$ .

- **Existencia de ciclos y tendencias**

Si estamos analizando un modelo econométrico cuya variable endógena presenta ciclos y/o tendencias que no se explican a través de las variables exógenas, la perturbación recoge dichas estructuras, presentando un comportamiento de autocorrelación. En este caso, los residuos presentan agrupaciones de desviaciones por encima del promedio (en la parte alta del ciclo) y agrupaciones de desviaciones por debajo del promedio (parte baja del ciclo).

- **Relaciones no lineales**

Supongamos que la verdadera relación entre los tipos de interés,  $r_t$ , y el stock de Deuda Pública,  $D_t$ , es cuadrática:

$$r_t = \beta_1 + \beta_2 D_t + \beta_3 D_t^2 + u_t \quad t = 1, 2, \dots, T \quad \beta_2 > 0, \beta_3 < 0$$

Este modelo implica que los tipos de interés aumentan al crecer el stock de deuda pública, aunque menos que proporcionalmente, puesto que se tiene:

$$\frac{\partial r_t}{\partial D_t} = \beta_2 + 2\beta_3 D_t < \beta_2$$

tanto menor cuanto mayor es  $D_t$ . Pero sin embargo se especifica y se estima un modelo lineal:

$$r_t = \beta_1 + \beta_2 D_t + u_t \quad t = 1, 2, \dots, T$$

En este caso la curvatura de la parte sistemática pasa a ser recogida por la perturbación. Los residuos presentarán una racha de residuos negativos seguida de otra racha de residuos positivos para seguir con otra negativa.

- **VARIABLES OMITIDAS RELEVANTES CORRELACIONADAS**

Si el modelo realmente se especifica como:

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t & t = 1, 2, \dots, T \\ u_t &= Y_t - \beta_1 - \beta_2 X_{2t} - \beta_3 X_{3t} \\ \hat{u}_t &= Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t} \end{aligned}$$

Pero estimamos:

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_{2t} + v_t & t = 1, 2, \dots, T \\ v_t &= Y_t - \beta_1 - \beta_2 X_{2t} = u_t + \beta_3 X_{3t} \\ \hat{v}_t &= Y_t - \tilde{\beta}_1 - \tilde{\beta}_2 X_{2t} = \hat{u}_t - (\tilde{\beta}_1 - \hat{\beta}_1) - (\tilde{\beta}_2 - \hat{\beta}_2) X_{2t} + \hat{\beta}_3 X_{3t} \end{aligned}$$

En este contexto de omisión recordemos que los estimadores MCO son sesgados en general. En consecuencia, tras un análisis de los residuos  $\hat{v}_t$  tanto gráfico como mediante tests, es muy probable que si la variable omitida está correlacionada o presenta ciclos o tendencias el investigador llegue a la conclusión de que:  $Cov(v_t, v_s) \neq 0$ . De todas formas hay que tener en cuenta que no siempre que se omite una variable relevante se causa autocorrelación.

### Proceso autorregresivo de primer orden, AR(1)

Existen numerosos procesos capaces de reproducir estructuras de autocorrelación en la perturbación, sin embargo el *proceso autorregresivo de primer orden* es el proceso de autocorrelación más sencillo y uno de los que mejor se suele ajustar a datos económicos<sup>2</sup>. Se especifica como:

$$u_t = \rho u_{t-1} + \epsilon_t \quad t = 1, 2, \dots, T$$

de forma que la perturbación en el periodo  $t$  depende de la perturbación del periodo anterior ( $t - 1$ ) más un término aleatorio (o innovación)  $\epsilon_t$  cuyas características son:

$$\begin{aligned} E(\epsilon_t) &= 0 \quad \forall t \\ E(\epsilon_t^2) &= \sigma_\epsilon^2 \quad \forall t \\ E(\epsilon_t \epsilon_s) &= 0 \quad \forall t, s \quad t \neq s \end{aligned} \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2)$$

y que habitualmente se le llama ruido blanco<sup>3</sup>. La especificación completa del MRLG cuando la perturbaciones presentan autocorrelación es:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_K X_{Kt} + u_t \quad t = 1, 2, \dots, T$$

$$u_t = \rho u_{t-1} + \epsilon_t \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2) \quad |\rho| < 1$$

El coeficiente  $\rho$  mide la correlación entre  $u_t$  y  $u_{t-1}$  y debe cumplir que  $|\rho| < 1$  para que el proceso no sea explosivo. Se le denomina coeficiente de autocorrelación de orden uno (o primer orden) ya que uno es el número de periodos entre  $u_t$  y  $u_{t-1}$ :

$$\rho = \frac{Cov(u_t, u_{t-1})}{\sqrt{Var(u_t)}\sqrt{Var(u_{t-1})}} \quad -1 < \rho < 1$$

Si la covarianza es positiva se le denomina correlación positiva y si es negativa, correlación negativa. Dado que  $u_t = Y_t - E(Y_t | \{X_{it}\}_{i=1}^K)$  la perturbación representa la diferencia entre el comportamiento observado y el comportamiento promedio. Dados los posibles valores de  $\rho$  tenemos que:

- i) Si  $\rho > 0$  entonces un valor elevado de  $u_t$  genera un valor de  $Y_t$  por encima del promedio y tendrá mayor probabilidad de ir seguido por un valor elevado de  $u_{t+1}$  y así sucesivamente.
- ii) Si  $\rho < 0$  un valor alto de  $u_t$  irá seguido por un valor bajo de  $u_{t+1}$  y éste por uno alto de  $u_{t+2}$  y así sucesivamente.

La relación entre la perturbación  $u_t$  y la innovación  $\epsilon_t$  se recoge en el diagrama siguiente:

$$\begin{array}{cccc} u_{t-2} & \longrightarrow & u_{t-1} & \longrightarrow & u_t & \longrightarrow & u_{t+1} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \epsilon_{t-2} & & \epsilon_{t-1} & & \epsilon_t & & \epsilon_{t+1} \end{array}$$

<sup>2</sup>El proceso autorregresivo de orden uno no es el único proceso que recoge dinámica en la perturbación. El proceso autorregresivo más general de todos es el proceso autorregresivo de orden  $p$ , AR(p):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

Además existen otros procesos alternativos a los autorregresivos.

<sup>3</sup>Si nos fijamos,  $\epsilon_t$  cumple las hipótesis básicas sobre la perturbación, es por tanto homocedástica y no autocorrelada.

por lo que cada innovación influye sobre la perturbación en el mismo periodo o periodos posteriores, pero nunca sobre los valores anteriores, es decir:  $E(\epsilon_t u_{t-s}) = 0 \quad s > 0$ . Además en el diagrama se puede observar que  $u_t$  no depende directamente de  $u_{t-2}$  pero sí lo hace a través de  $u_{t-1}$ , por tanto  $u_t$  está correlado con todas las perturbaciones pasadas.

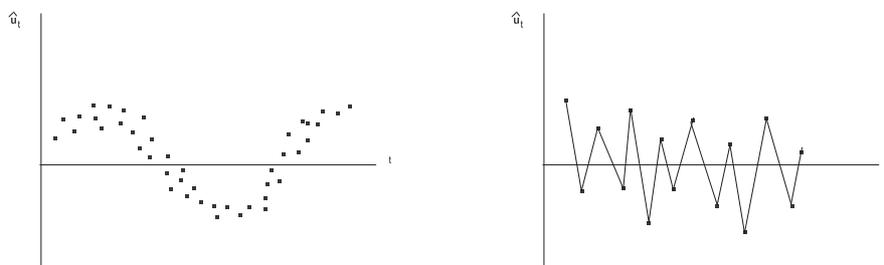


Figura 5.8: Proceso autorregresivo de orden uno

En la izquierda de la Figura 5.8 se observa un proceso autorregresivo de primer orden con parámetro  $\rho$  positivo. En ella podemos observar una racha de residuos positivos seguidos de una racha de residuos negativos y así sucesivamente. En cambio, cuando el parámetro del proceso autorregresivo es negativo, los signos de los residuos se alternan como podemos ver en la figura de la derecha del mismo gráfico.

### Detección de la autocorrelación

En la práctica no se conoce a priori si existe autocorrelación ni cuál es el proceso más adecuado para modelarla. Para determinar su existencia es necesario contrastar dicha hipótesis mediante un estadístico de contraste.

Si embargo, ningún contraste de autocorrelación debe excluir un examen riguroso de los residuos generados en la estimación del modelo. El gráfico de los mismos puede indicarnos la existencia de autocorrelación. Dado que los residuos son una aproximación a la perturbación, la existencia de patrones o comportamientos sistemáticos en los mismos indicaría la posible existencia de autocorrelación en  $u_t$ . Por ejemplo, podemos esperar que el gráfico de la evolución temporal de  $\hat{u}_{t,MCO}$  se comporte de forma similar a lo mostrado por la Figura 5.8. Sin embargo, también podemos dibujar la evolución temporal de  $\hat{u}_{t,MCO}$  contra la de  $\hat{u}_{s,MCO}$  para  $s = t - 1$ . Si encontramos que la mayoría de los puntos en dicho gráfico se hallan en el primer o tercer cuadrante, izquierda de la Figura 5.9, ello es un indicio de autocorrelación positiva. Si se hallan en el segundo y cuarto cuadrante, derecha de la Figura 5.9, indicará autocorrelación negativa.

Tras el análisis gráfico, si sospechamos que existe autocorrelación debemos contrastarla con un estadístico de contraste. Existen varios estadísticos de contraste pero en este tema vamos a estudiar sólo uno, el estadístico de Durbin Watson, específico para contrastar la existencia de un proceso autorregresivo de primer orden.

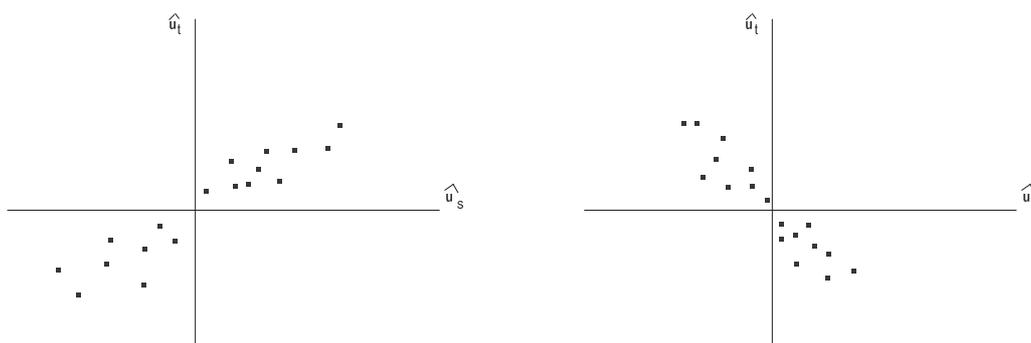


Figura 5.9: Perturbaciones AR(1) positivo versus AR(1) negativo

### Contraste de Durbin Watson

Durbin y Watson propusieron en 1951 un estadístico para contrastar la existencia de un proceso AR(1) en el término de perturbación. La hipótesis nula es la no existencia de autocorrelación:

$$H_0 : \rho = 0$$

frente a la alternativa

$$H_a : \rho \neq 0 \quad \text{en} \quad u_t = \rho u_{t-1} + \epsilon_t \quad \epsilon_t \sim (0, \sigma_\epsilon^2)$$

y se contrasta mediante el estadístico:

$$DW = \frac{\sum_{t=2}^T (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^T \hat{u}_t^2}$$

donde  $\hat{u}_t$  son los residuos mínimo-cuadráticos ordinarios de estimar el modelo original sin tener en cuenta la existencia de autocorrelación en las perturbaciones. *Gretl* proporciona el valor de este estadístico entre los resultados de la estimación MCO. Sin embargo antes de utilizarlo vamos a estudiar su interpretación y manejo sin recurrir a *gretl*.

- Interpretación del estadístico DW:

Si el tamaño muestral es suficientemente grande<sup>4</sup> podemos emplear la relación  $DW \simeq 2(1 - \hat{\rho})$  en base a la cual podemos establecer el siguiente comportamiento en los residuos:

- Si existe autocorrelación positiva de primer orden, valores positivos del término de error  $u_t$  tiendan a ir seguidos de valores positivos y asimismo, valores negativos tiendan a ir seguidos

<sup>4</sup>Teniendo en cuenta las aproximaciones:

$$\sum_{t=2}^T \hat{u}_t^2 \simeq \sum_{t=2}^T \hat{u}_{t-1}^2 \simeq \sum_{t=1}^T \hat{u}_t^2$$

con lo que

$$DW \simeq \frac{2 \sum_{t=2}^T \hat{u}_t^2 - 2 \sum_{t=2}^T \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^T \hat{u}_t^2} \simeq 2 - 2 \frac{\sum_{t=2}^T \hat{u}_t \hat{u}_{t-1}}{\sum_{t=2}^T \hat{u}_{t-1}^2} \simeq 2(1 - \hat{\rho})$$

donde  $\hat{\rho}$  es el estimador de  $\rho$  por MCO en el modelo  $u_t = \rho u_{t-1} + \epsilon_t$ , empleando como aproximación de  $u_t$  el residuo MCO, es decir,  $\hat{u}_t = \rho \hat{u}_{t-1} + \epsilon_t$  y  $\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_{t-1}^2}$ .

de valores negativos. Dado que la aproximación a la perturbación es el residuo, los patrones en la perturbación serán detectados en el residuo. Así, observaremos agrupaciones de residuos positivos seguidas de agrupaciones de residuos negativos. En estas circunstancias, generalmente  $|\hat{u}_t - \hat{u}_{t-1}| < |\hat{u}_t| \Rightarrow (\hat{u}_t - \hat{u}_{t-1})^2 < \hat{u}_t^2$  y el numerador del estadístico “será pequeño” en relación al denominador, con lo que el estadístico “será pequeño”. En consecuencia cuanto más cercano esté el parámetro  $\rho$  a la unidad más próximo a cero estará el DW. En el extremo positivo tenemos que  $\rho \rightarrow 1 \Rightarrow DW \rightarrow 0$ .

- Si existe autocorrelación negativa de primer orden, valores positivos de  $\hat{u}_t$  tienden a ir seguidos de valores negativos, en este caso  $|\hat{u}_t - \hat{u}_{t-1}| > |\hat{u}_t| \Rightarrow (\hat{u}_t - \hat{u}_{t-1})^2 > \hat{u}_t^2$  con lo que el estadístico DW tenderá a tomar valores grandes. En el extremo negativo tenemos que  $\rho \rightarrow -1 \Rightarrow DW \rightarrow 4$ .

A partir de la relación  $DW \simeq 2(1 - \hat{\rho})$  se puede establecer el rango de valores que puede tomar el estadístico DW.

$$\begin{aligned} 0 < \hat{\rho} < 1 & \quad DW \in (0, 2) \\ \hat{\rho} = 0 & \quad DW \simeq 2 \\ -1 < \hat{\rho} < 0 & \quad DW \in (2, 4) \end{aligned}$$

La distribución del estadístico DW bajo  $H_0$  depende de la matriz de regresores  $X$  por lo que los valores críticos del contraste también serán diferentes para cada posible  $X$ . Durbin y Watson tabularon los valores máximo ( $d_U$ ) y mínimo ( $d_L$ ) que puede tomar el estadístico independientemente de cuál sea  $X$ , y tal que  $d_L < DW < d_U$ . La distribución de  $d_L$  y  $d_U$  depende del tamaño de la muestra,  $T$ , y de  $K'$  que denota el número de variables explicativas del modelo exceptuando el término independiente.

• **Contraste de existencia de autocorrelación positiva:**

$$H_0 : \rho = 0$$

$$H_a : \rho > 0 \quad \text{en} \quad u_t = \rho u_{t-1} + \epsilon_t \quad |\rho| < 1 \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2)$$

1. Si  $DW < d_L$  se rechaza la  $H_0$  para un nivel de significatividad  $\alpha$  dado, por tanto existe autocorrelación positiva.
2. Si  $DW > d_U$  no se rechaza la  $H_0$  para un nivel de significatividad  $\alpha$  dado, por tanto no existe autocorrelación positiva.
3. Si  $d_L < DW < d_U$  estamos en una zona de incertidumbre y no podemos concluir si existe o no autocorrelación positiva de primer orden.

• **Contraste de existencia de autocorrelación negativa:**

$$H_0 : \rho = 0$$

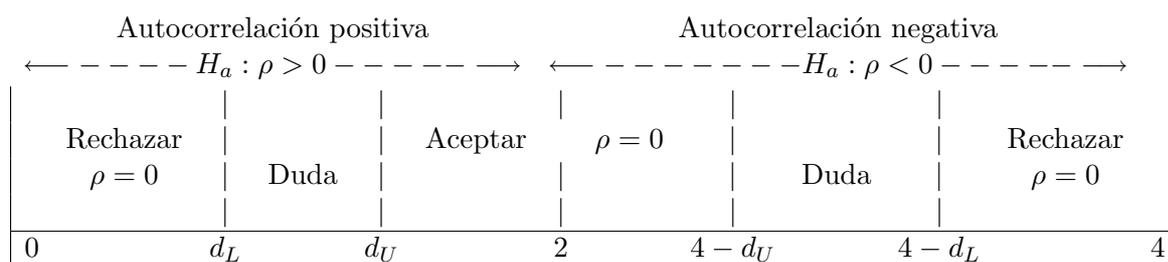
$$H_a : \rho < 0 \quad \text{en} \quad u_t = \rho u_{t-1} + \epsilon_t \quad |\rho| < 1 \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2)$$

1. Si  $DW < 4 - d_U$  no se rechaza la  $H_0$  para un nivel de significatividad  $\alpha$  dado, por tanto no existe autocorrelación negativa.

2. Si  $DW > 4 - d_L$  se rechaza la  $H_0$  para un nivel de significatividad  $\alpha$  dado, por tanto existe autocorrelación negativa.
3. Si  $4 - d_U < DW < 4 - d_L$  estamos en una zona de incertidumbre y como en el caso anterior, no podemos concluir si existe o no autocorrelación negativa de primer orden.

Gráficamente:

$$H_0 : \rho = 0$$



Si el resultado del contraste es que existe autocorrelación y ésta no es debida a una mala especificación del modelo, éste no debe ser estimado por MCO ya que este estimador es ineficiente. Si la autocorrelación es originada por una mala especificación del modelo primero se ha de corregir esta especificación y una vez el modelo esté correctamente especificado analizar las propiedades de la perturbación y actuar en consecuencia.

• Observaciones sobre el contraste de Durbin Watson:

1. El contraste de Durbin Watson también se puede considerar un contraste de mala especificación del modelo. La omisión de variables relevantes correlacionadas, una forma funcional inadecuada, cambios estructurales no incluidos en el modelo, etc., pueden originar un estadístico DW significativo. Esto nos puede llevar a errores si consideramos que hay evidencia de autocorrelación y se modela un proceso AR(1). Por otro lado, si  $u_t$  sigue un proceso distinto de un AR(1), es probable que el estadístico DW lo detecte. Por lo tanto, el estadístico de Durbin Watson es útil porque nos indica la existencia de problemas en el modelo, pero a veces no nos ayuda a establecer cuál es la estructura real. En caso de no rechazar la  $H_0$ , podemos afirmar que no tenemos un AR(1), pero no sabemos si tenemos alguna otra estructura alternativa.
2. Por otro lado el estadístico DW sólo debe aplicarse cuando los regresores son **no estocásticos**, en presencia de regresores aleatorios como la variable endógena retardada no tiene validez.
3. Cuando el estadístico DW cae en zona de duda, y si no podemos llevar a cabo un contraste alternativo, no debemos concluir que no existe autocorrelación. El procedimiento conservador aconseja rechazar la hipótesis nula y estimar por un estimador alternativo a MCO ya que las consecuencias de ignorar su existencia cuando sí la hay son más graves que las correspondientes al caso contrario.

## 5.4. Validación en gretl

### 5.4.1. Contraste de Ramsey con gretl

El contraste de Ramsey es uno de los contrastes que pueden realizarse una vez estimado en *gretl* un modelo. En la pantalla de resultados de la estimación Mínimo Cuadrática Ordinaria pinchar

*Contrastes* → *Contraste RESET de Ramsey*

y elegir las variables que queremos introducir en la especificación para el contraste:

- cuadrados y cubos
- solo cuadrados
- solo cubos
- todas las variantes

Para mostrar cómo contrastar la ausencia de correlación utilizaremos el archivo de datos Ramanathan *data3-3.gdt*. En este archivo de datos se dispone de 34 observaciones para el periodo 1960-1993 (serie temporal por tanto) sobre el número de resultados de patentes en miles, *PATENTES*, y sobre el gasto en *I + D*, en billones de dólares. La relación a estudiar es:

$$PATENTES_t = \alpha + \beta(I + D)_t + u_t \quad (5.16)$$

Los resultados de la estimación MCO son los siguientes:

Modelo 1: MCO, usando las observaciones 1960–1993 ( $T = 34$ )

Variable dependiente: PATENTS

	Coeficiente	Desv. Típica	Estadístico $t$	Valor p
const	34,5711	6,35787	5,4375	0,0000
I+D	0,791935	0,0567036	13,9662	0,0000
Media de la vble. dep.	119,2382	D.T. de la vble. dep.	29,30583	
Suma de cuad. residuos	3994,300	D.T. de la regresión	11,17237	
$R^2$	0,859065	$R^2$ corregido	0,854661	
$F(1, 32)$	195,0551	Valor p (de $F$ )	3,64e-15	
Log-verosimilitud	-129,2704	Criterio de Akaike	262,5408	
Criterio de Schwarz	265,5935	Hannan–Quinn	263,5818	
$\hat{\rho}$	0,945182	Durbin–Watson	0,233951	

Contraste de especificación RESET (cuadrados y cubos) Estadístico de contraste:  $F = 47,161051$ , con valor-p=  $P(F(2, 30) > 47,1611) = 5,48e - 010$ .

Contraste de especificación RESET (cuadrados sólo) Estadístico de contraste:  $F = 14,824629$ , con valor-p=  $P(F(1, 31) > 14,8246) = 0,000553$ .

Contraste de especificación RESET (cubos sólo) Estadístico de contraste:  $F = 18,475400$ , con valor-p=  $P(F(1, 31) > 18,4754) = 0,000158$ . En cualquiera de las posibilidades se rechaza la hipótesis nula luego la especificación lineal no es adecuada, como puede apreciarse en la Figura 5.15 donde se muestra la nube de puntos correspondiente.

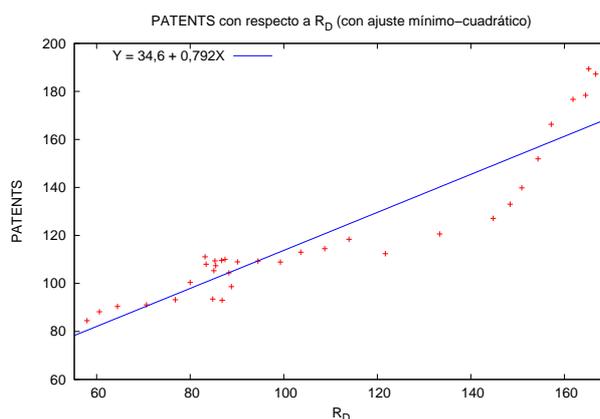


Figura 5.10: Variable endógena versus exógena

#### 5.4.2. Contraste de cambio estructural o Chow con gretl

Utilizando *gretl* una vez abierto el fichero de datos y estimado el modelo correspondiente por MCO, en la ventana de resultados de la estimación haríamos:

*Contrastes* → *Contraste de Chow*

A la pregunta *Observación en la cual dividir la muestra* contestaríamos *fecha correspondiente a T<sub>1</sub>* y automáticamente *gretl* realiza el contraste y nos muestra el resultado.

Por ejemplo el fichero data7-19 del libro de Ramanathan contiene datos para 1960-1988 sobre la demanda de tabaco y sus determinantes en Turquía. Las variables de interés para el ejemplo son las siguientes:

$Q$ : consumo de tabaco por adulto (en kg).

$Y$ : PNB real per cápita en liras turcas de 1968.

$P$ : precio real del kilogramo de tabaco, en liras turcas.

$D82$ : variable ficticia que toma valor 1 a partir de 1982.

A mediados de 1981 el gobierno turco lanza una campaña de salud pública advirtiendo de los peligros de salud que conlleva el consumo de tabaco. Nuestro objetivo es determinar si existen cambios en la demanda de tabaco tras la campaña institucional en cuyo caso la especificación:

$$\ln Q_t = \alpha + \beta \ln Y_t + \gamma \ln P_t + u_t \quad t = 1960, \dots, 1988 \quad (5.17)$$

no es correcta para todo el período muestral y deberíamos especificar dos ecuaciones:

$$\text{Ln}Q_t = \alpha_1 + \beta_1 \text{Ln}Y_t + \gamma_1 \text{Ln}P_t + u_{1t} \quad t = 1960, \dots, 1981 \quad (5.18)$$

$$\text{Ln}Q_t = \alpha_2 + \beta_2 \text{Ln}Y_t + \gamma_2 \text{Ln}P_t + u_{2t} \quad t = 1982, \dots, 1988 \quad (5.19)$$

Si existe cambio estructural rechazaríamos  $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$  y  $\gamma_1 = \gamma_2$

En este caso la contestación a la pregunta *Observación en la cual dividir la muestra* contestaríamos 1982 y el output de *gretl* muestra lo siguiente:

Modelo 1: estimaciones MCO utilizando las 29 observaciones 1960-1988

Variable dependiente: lnQ

Variable	Coefficiente	Desv. típica	Estadístico t	valor p
const	-4,58987	0,724913	-6,332	0,00001***
lnY	0,688498	0,0947276	7,268	0,00001***
lnP	0,485683	0,101394	-4,790	0,00006***

Media de la var. dependiente = 0,784827

Desviación típica de la var. dependiente. = 0,108499

Suma de cuadrados de los residuos = 0,0949108

Desviación típica de los residuos = 0,0604187

R-cuadrado = 0,712058

R-cuadrado corregido = 0,689908

Estadístico F (2, 26) = 32,148 (valor p < 0,00001)

Estadístico de Durbin-Watson = 1,00057

Coef. de autocorr. de primer orden. = 0,489867

Log-verosimilitud = 41,8214

Criterio de información de Akaike (AIC) = -77,6429

Criterio de información Bayesiano de Schwarz (BIC) = -73,541

Criterio de Hannan-Quinn (HQC) = -76,3582

Contraste de Chow de cambio estructural en la observación 1982 -

Hipótesis nula: no hay cambio estructural

Estadístico de contraste: F(3, 23) = 20,1355

con valor p = P(F(3, 23) > 20,1355) = 1,25619e-006

El estadístico calculado es  $F = 20,135 > F_{0,05}(3, 23)$  por lo que rechazamos  $H_0$  para un nivel de significatividad del 5 %, es decir existe cambio estructural, la campaña institucional ha tenido efecto y la demanda de tabaco en Turquía de 1960 a 1988 queda especificada por las ecuaciones (5.18) y (5.19). Los resultados de la estimación mínimo cuadrática de estas ecuaciones son los siguientes:

$$\widehat{\text{Ln}Q}_t = \begin{matrix} -5,024 & + & 0,735 & \text{Ln}Y_t & - & 0,381 & \text{Ln}P_t & & t = 1960, \dots, 1981 & SCR_1 = 0,01654 \\ (-10,614) & & (11,587) & & & (-4,227) & & & & \end{matrix}$$

$$\widehat{\text{Ln}Q}_t = \begin{matrix} 8,837 & + & -0,953 & \text{Ln}Y_t & + & 0,108 & \text{Ln}P_t & & t = 1982, \dots, 1988 & SCR_2 = 0,00965 \\ (2,170) & & (-1,941) & & & (0,654) & & & & \end{matrix}$$

### Cambio estructural utilizando variables ficticias

Alternativamente, el contraste anterior podríamos haberlo realizado mediante la variable ficticia  $D82$  especificando el modelo:

$$\ln Q_t = \beta_1 + \beta_2 \ln Y_t + \beta_3 \ln P_t + \beta_1^* D82_t + \beta_2^* D82_t \cdot \ln Y_t + \beta_3^* D82_t \cdot \ln P_t + u_t \quad t = 60, \dots, 88 \quad (5.20)$$

En el cual, si existe cambio estructural rechazaríamos  $H_0 : \beta_1^* = \beta_2^* = \beta_3^* = 0$ . De nuevo el contraste puede realizarse con el estadístico F habitual de sumas residuales donde el modelo no restringido es el (5.20) y el modelo restringido es

$$\ln Q_t = \beta_1 + \beta_2 \ln Y_t + \beta_3 \ln P_t + u_t \quad (5.21)$$

Utilizando *gretl* el proceso después de abierto el fichero de datos, tomado logaritmos y construido las variables  $D82 \cdot \ln Y$  y  $D82 \cdot \ln P$  sería: estimaríamos el modelo (5.20) por MCO y en la ventana de resultados de la estimación haríamos

*Contrastes* → *Omitir variables*

elegiríamos  $D82$ ,  $D82 \cdot \ln Y$  y  $D82 \cdot \ln P$  y obtendríamos el siguiente resultado:

Modelo 1: estimaciones MCO utilizando las 29 observaciones 1960-1988  
Variable dependiente:  $\ln Q$

Variable	Coefficiente	Desv. típica	Estadístico $t$	valor p
const	-4,58987	0,724913	-6,332	0,00001***
$\ln Y$	0,688498	0,0947276	7,268	0,00001***
$\ln P$	0,485683	0,101394	-4,790	0,00006***

Media de la var. dependiente = 0,784827  
 Desviación típica de la var. dependiente. = 0,108499  
 Suma de cuadrados de los residuos = 0,0949108  
 Desviación típica de los residuos = 0,0604187  
 R-cuadrado = 0,712058  
 R-cuadrado corregido = 0,689908  
 Estadístico F (2, 26) = 32,148 (valor p < 0,00001)  
 Estadístico de Durbin-Watson = 1,00057  
 Coef. de autocorr. de primer orden. = 0,489867  
 Log-verosimilitud = 41,8214  
 Criterio de información de Akaike (AIC) = -77,6429  
 Criterio de información Bayesiano de Schwarz (BIC) = -73,541  
 Criterio de Hannan-Quinn (HQC) = -76,3582

Comparación entre el modelo 10 y el modelo 11:

Hipótesis nula: los parámetros de regresión son cero para las variables

D82

D82Y

D82P

Estadístico de contraste:  $F(3, 23) = 20,1355$ , con valor  $p = 1,25619e-006$

De los 3 estadísticos de selección de modelos, 0 han mejorado.

Dado el p-value rechazamos la hipótesis nula para un nivel de significatividad del 5% y existe cambio estructural. La demanda de tabaco en Turquía de 1960 a 1988 queda especificada por las ecuaciones (5.18) y (5.19).

### 5.4.3. Contraste de heterocedasticidad con gretl

Para ilustrar esta sección y lo que queda del tema vamos a utilizar el conjunto de datos *Data 3.2* del Ramanathan. En este conjunto de datos se dispone de 51 observaciones sobre renta personal, *INCOME*, y gasto sanitarios, *EXPLTH* ambos en billones de dólares para el estado de Washington D.C. en 1993. Se trata por tanto de una muestra de sección cruzada. Queremos analizar la evolución del gasto en función de la renta, así especificamos el modelo:

$$EXPLTH_i = \alpha + \beta INCOME_i + u_i \quad i = 1, \dots, 51 \quad (5.22)$$

suponemos que se cumplen las hipótesis básicas y estimamos el modelo por MCO con los resultados siguientes<sup>5</sup>:

Modelo 2: estimaciones MCO utilizando las 51 observaciones 1-51

Variable dependiente: *explth*

	VARIABLE	COEFICIENTE	DESV.TÍP.	ESTAD.T	2Prob(t >  T )
0)	const	0,325608	0,319742	1,018	0,313515
2)	income	0,142099	0,00196623	72,270	< 0,00001 ***

Media de la var. dependiente = 15,2649

D.T. de la var. dependiente = 17,8877

Suma de cuadrados de los residuos = 148,699

Desviación típica de los residuos = 1,74203

R-cuadrado = 0,990705

R-cuadrado corregido = 0,990516

Grados de libertad = 49

Criterio de información de Akaike (AIC) = 203,307

Criterio de información Bayesiano de Schwarz (BIC) = 207,17

<sup>5</sup>Recordatorio de la secuencia de órdenes para obtener la estimación:

*Archivo* → *Abrir datos* → *Archivo de muestra* → *Data3.2*

*Modelo* → *Mínimos Cuadrados* → *seleccionar la variable endógena y exógenas*

Los resultados se muestran en una ventana llamada *gretl: modelo1*

Para un nivel de significatividad del 5% la variable *INCOME* es significativa, pero el término independiente no es significativamente distinto de cero. El ajuste es muy alto, 99,07%. Gráficamente podemos ver si la forma funcional lineal elegida resulta adecuada para la relación gasto sanitario- renta. Para ello vamos a dibujar la relación gasto sanitario real y ajustado y además el ajuste MCO. Para ello dentro de la venta *gretl*: Modelo1 pinchamos la secuencia:

Gráficos → Gráfico de ajustada-observada → por número de observación

Gráficos → Gráfico de ajustada-observada → contra income

Las figuras obtenidas aparecen a la izquierda y derecha, respectivamente, en la Figura 5.11.

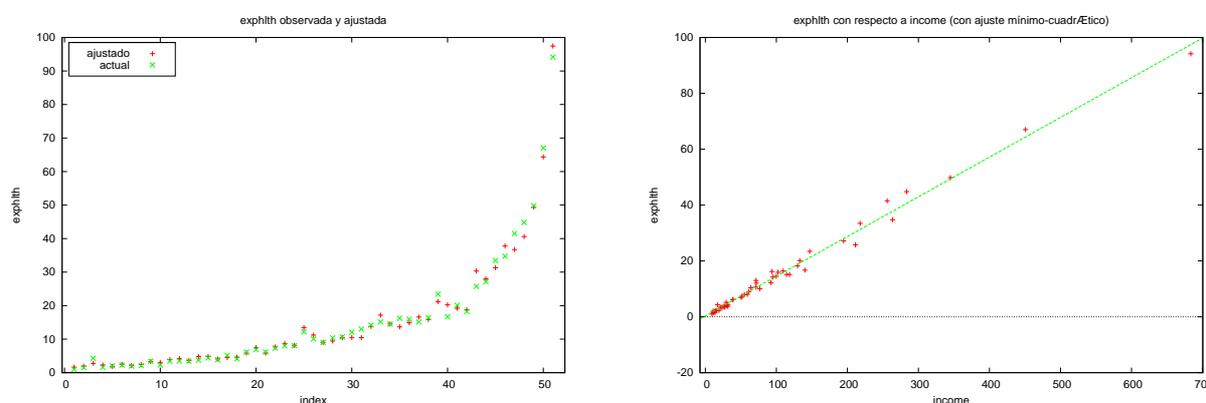


Figura 5.11: Gasto sanitario real y ajustado

Aparentemente el modelo está correctamente especificado y la forma lineal especificada es adecuada. Antes de seguir vamos a guardar los valores ajustados del gasto sanitario los residuos y sus cuadrados añadiéndolos al conjunto de datos para luego poder trabajar con ellos si es necesario. La secuencia de órdenes a realizar en la ventana *gretl*: Modelo1 es:

Datos del modelo → Añadir al conjunto de datos → valores ajustados

Datos del modelo → Añadir al conjunto de datos → residuos

Datos del modelo → Añadir al conjunto de datos → residuos al cuadrado

*gretl* los va a añadir al conjunto de datos con el que trabajamos y los denota respectivamente por *yhat1*, *uhat1* e *usq1* respectivamente. Además añade una leyenda explicativa de la variable. Una vez hecho esto seguimos con el ejercicio.

A pesar del buen ajuste encontrado, no resulta descabellado pensar que la varianza del gasto sanitario, *EXPHLTH*, probablemente dependerá de la renta *INCOME*. Hemos visto que estudiar el gráfico de residuos frente a *INCOME* es un instrumento válido para ver indicios del problema.

Para obtener el gráfico en la ventana *gretl*: Modelo 1 pinchamos:

Gráficos → Gráfico de residuos → contra income

La figura obtenida se recoge en la Figura 5.12:

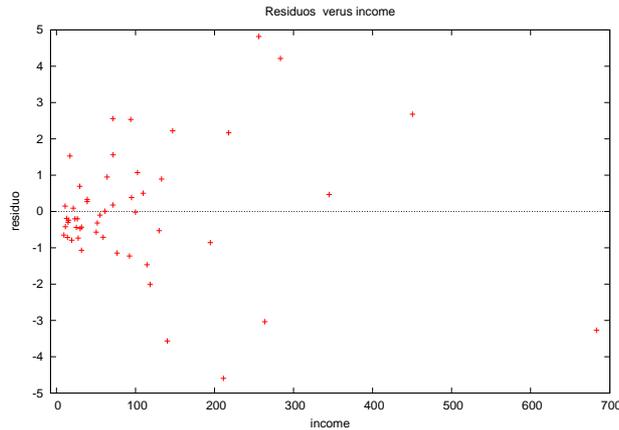


Figura 5.12: Residuos MCO versus RENTA

En el se aprecia que la dispersión de los residuos aumenta a medida que aumenta *INCOME*. Parece que la varianza de *EXPHLTH* aumenta con *INCOME*

Para confirmarlo realizamos el contraste de White y los resultados del mismo son:

Contraste de heterocedasticidad de White estimaciones MCO  
 utilizando las 51 observaciones 1-51 Variable dependiente:  $uhat^2$

VARIABLE	COEFICIENTE	DESV.TÍP.	ESTAD.T	2Prob(t >  T )
0) const	-1,40227	0,986956	-1,421	0,161839
2) income	0,0558410	0,0121349	4,602	0,000031 ***
4) sq_incom	-5,87208E-05	2,10114E-05	-2,795	0,007445 ***

R-cuadrado = 0,421177

Estadístico de contraste:  $TR^2 = 21,480039$ , con valor p =  
 $P(\text{Chi-cuadrado}(2) > 21,480039) = 0,000022$

Los resultados confirman que efectivamente existe heterocedasticidad en la perturbaciones del modelo (5.22). El estimador MCO obtenido no es de varianza mínima y la inferencia realizada de acuerdo a él no es válida.

Esta introducción al problema de heterocedasticidad pretende que hayáis aprendido que nunca se da una especificación por correcta sin un análisis de residuos. Que a pesar de que durante todo el curso hemos trabajado suponiendo que se cumplen unas hipótesis básicas lo habitual es que no sea así y que estas situaciones hay que saber reconocerlas. Ampliaciones sobre este tema podéis encontrar en el Capítulo 8 del Ramanathan del que nosotros solo hemos hecho un esbozo de su introducción.

#### 5.4.4. Contraste de ausencia de correlación con gretl

Para mostrar cómo contrastar la ausencia de correlación utilizaremos el archivo de datos Ramanathan Data3-3. En este archivo de datos se dispone de 34 observaciones para el periodo 1960-1993 (serie temporal por tanto) sobre el número de resultados de patentes en miles, *PATENTES*, y sobre el gasto en *I + D*, en billones de dólares. La relación a estudiar es:

$$PATENTES_t = \alpha + \beta(I + D)_t + u_t \quad (5.23)$$

Los resultados de la estimación MCO son los siguientes:

Modelo 1: MCO, usando las observaciones 1960–1993 ( $T = 34$ )

Variable dependiente: PATENTS

	Coefficiente	Desv. Típica	Estadístico $t$	Valor p
const	34,5711	6,35787	5,4375	0,0000
I+D	0,791935	0,0567036	13,9662	0,0000
Media de la vble. dep.		119,2382	D.T. de la vble. dep.	29,30583
Suma de cuad. residuos		3994,300	D.T. de la regresión	11,17237
$R^2$		0,859065	$R^2$ corregido	0,854661
$F(1, 32)$		195,0551	Valor p (de $F$ )	3,64e-15
Log-verosimilitud		-129,2704	Criterio de Akaike	262,5408
Criterio de Schwarz		265,5935	Hannan–Quinn	263,5818
$\hat{\rho}$		0,945182	Durbin–Watson	0,233951

Los resultados muestran que para un nivel de significatividad del 5% el término independiente es significativamente distinto de cero y el gasto en I+D es una variable significativa para explicar las aplicaciones de las patentes. Además existe un buen ajuste en términos del  $R^2$  (85,9%). Si analizamos los residuos MCO dibujándolos contra el tiempo obtenemos la Figura 5.13. En él podemos ver un primer grupo de residuos positivos que va seguido de un grupo de residuos negativos, otrora positivos y a continuación negativos. Este comportamiento puede indicar la posible existencia de un proceso autorregresivo de primer orden y signo positivo.

También podíamos haber dibujado los pares  $(u_{t-1}, u_t)$ , ver Figura 5.14, en este caso los puntos los

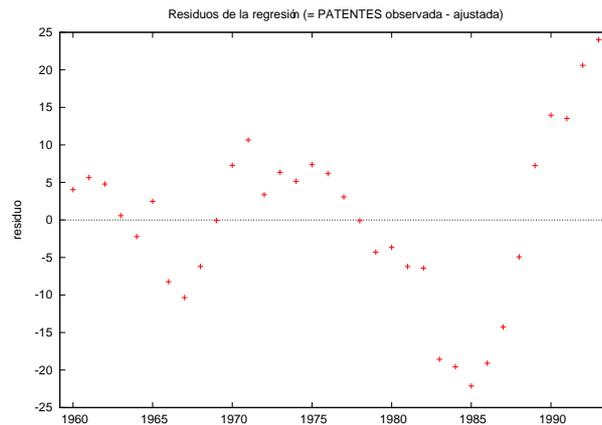


Figura 5.13: Residuos versus tiempo

encontramos en el primer y tercer cuadrante indicando autocorrelación de primer orden de signo positivo<sup>6</sup>.

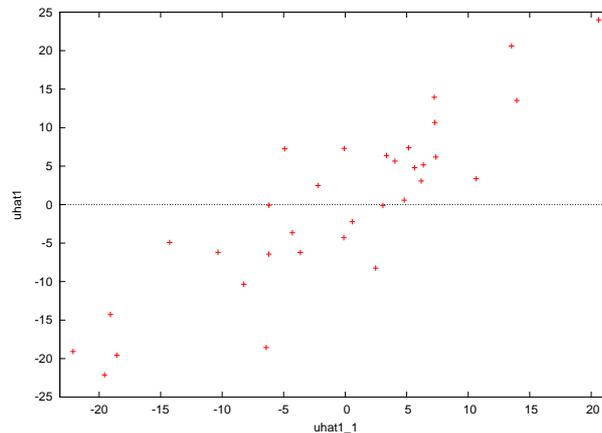


Figura 5.14: Residuos en t versus residuos en t-1

Una vez analizados los gráficos debemos realizar un contraste para cerciorarnos de la existencia del problema. Entre los resultados de la regresión se nos muestra:

Estadístico de Durbin-Watson = 0,233951

que utilizaremos para contrastar la existencia de un proceso autocorregresivo de primer orden y signo positivo en la perturbación, ya que  $DW \in (0, 2)$ .

$$H_0 : \rho = 0$$

---

<sup>6</sup>Para guardar los residuos en la ventana de resultados de la estimación pinchamos *Datos del modelo* → *Añadir al conjunto de datos* → *residuos* y para obtener su retardo  $u_{t-1}$  seleccionamos la variable residuos y pinchamos la secuencia *Datos* → *Añadir al conjunto de datos* → *retardos de las variables seleccionadas*.

$$H_a : \rho > 0 \text{ en } u_t = \rho u_{t-1} + \epsilon_t \quad \epsilon_t \sim (0, \sigma_\epsilon^2)$$

Para ello sólo necesitamos comparar el valor del estadístico con  $d_L$  y  $d_U$  obtenidos en las tablas correspondientes. *Gretl* nos proporciona estas tablas en la opción *Herramientas* que aparece en la primera pantalla una vez se abre el programa. La secuencia a pinchar es:

*Herramientas* → *tablas estadísticas* → *señalar la tabla deseada*

*Gretl* proporciona las tablas estadísticas de la normal, t (t-student), chi-cuadrado, *F* (F-snedercor) y DW (Durbin-Watson). En nuestro caso pinchamos en esta última y se nos despliega una ventana que nos solicita el tamaño de muestra y número de regresores. Se lo damos y pinchamos *Aceptar*. Como resultado *gretl* nos devuelve una ventana con el valor de  $d_L$  y  $d_U$  para el tamaño de muestra dado y diferentes valores de  $K'$ . Para nuestro ejemplo obtenemos:

Valores críticos al 5% del estadístico de Durbin-Watson,  $n = 134$ ,  $k = 1$

$$dL = 1,7028$$

$$dU = 1,7329$$

Por tanto para nuestro ejemplo  $d_L(T = 34, K' = 1) = 1,39$  y  $d_U(T = 34, K' = 1) = 1,51$ . Como  $DW = 0,233 < d_L$  se rechaza la  $H_0$  y por tanto existe autocorrelación positiva de primer orden o bien puede deberse a una mala especificación del modelo. Antes de buscar un estimador alternativo a MCO debemos explorar esta posibilidad e intentar especificar bien el modelo y volver a realizar un estudio de existencia de autocorrelación para el modelo correctamente especificado. Si analizamos la relación entre las variables exógena y endógena, Figura 5.15 vemos que esta no parece ser lineal si no cuadrática al menos en los dos últimos tercios de la muestra, por lo que vamos a proponer la siguiente relación cuadrática:

$$PATENTES_t = \alpha + \beta(I + D)_t + \gamma(I + D)_t^2 + u_t \tag{5.24}$$

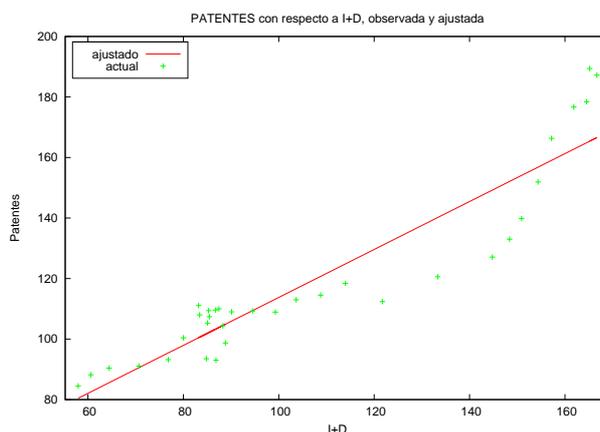


Figura 5.15: Variable endógena versus exógena

Los resultados de su estimación MCO son:

$$\widehat{PATENTES}_t = 121,57 - 0,85 (I + D)_t + 0,007 (I + D)_t^2 \quad R^2 = 0,90 \quad DW = 0,28 \tag{5.25}$$

(t-estad)            (5,23)            (-1,98)            (3,85)

Las variables son significativas para un nivel de significatividad del 5% y el ajuste es bueno, 90%. Sin embargo, para el modelo (5.25) sigue existiendo autocorrelación positiva de primer orden ya que  $DW = 0,28 < d_L(T = 34, K' = 2) = 1,33$ . Si miramos el gráfico de residuos de esta relación, Figura 5.16, encontramos la misma evolución cíclica de grupos de residuos positivos-negativos-positivos. Por tanto una vez especificado correctamente el modelo se sigue manteniendo la autocorrelación en las perturbaciones. El modelo (5.24) debe ser estimado por un estimador alternativo a MCO que sea de varianza mínima y permita realizar inferencia válida.

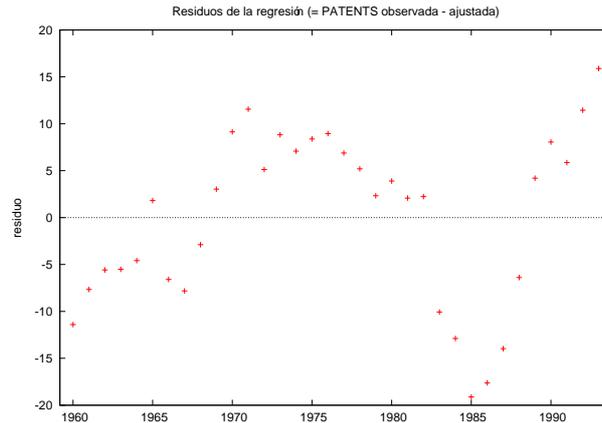


Figura 5.16: Residuos modelo (6.20) versus tiempo

## 5.5. Bibliografía del tema

### Referencias bibliográficas básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con gretl:

- [1] Ramanathan, R. (2002), *Instructor's Manual to accompany*, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), *Student Solutions Manual*, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

### Referencias Bibliográficas Complementarias:

- [1] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Análisis de regresión con gretl*. Open Course Ware. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-gretl/Courseisting>).
- [2] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación on-line de la Facultad de C.C. Económicas y Empresariales.
- [3] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones.
- [4] Esteban, MV (2008). *Estadística Actuarial: Regresión Lineal*, Sarriko On Line 3/08. Publicación on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.
- [5] Esteban, M.V. (2007). *Colección de ejercicios y exámenes*. Material docente. Servicio de Publicaciones.
- [6] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). *Ejercicios de Econometría*. Editorial McGraw-Hill.
- [7] Greene, W. (1998), *Análisis Económico*, Ed. Prentice Hall, 3ª edición.
- [8] Ramanathan, R. (2002), *Introductory Econometrics with applications*, Ed. South-Western, 5th. edition.
- [9] Verbeek, M. (2004). *A Guide to Modern Econometrics*. Wiley.



# Bibliografía

## Referencias Bibliográficas Básicas:

- Teórica:

- [1] Gujarati, D. y Porter, D.C. (2010). *Econometría*. Editorial McGraw-Hill, Madrid. 5ª edición.
- [2] Newbold, P., Carlson, W.L. y Thorne, B. (2008). *Estadística para administración y economía*. Prentice Hall. Madrid.
- [3] Wooldridge, J.M. (2006). *Introducción a la Econometría*. Ed. Thomson Learning, 2ª edición.
- [4] Ruiz Maya, L. y Martín Pliego, F.J. (2005). *Fundamentos de inferencia estadística*, 3ª edición, Editorial AC, Madrid.

- Ejercicios con gretl:

- [1] Ramanathan, R. (2002), Instructor's Manual to accompany, del libro *Introductory Econometrics with applications*, ed. South-Western, 5th edition, Harcourt College Publishers.
- [2] Wooldridge, J. M. (2003), Student Solutions Manual, del libro *Introductory Econometrics: A modern Approach*, ed. South-Western, 2nd edition.

## Referencias Bibliográficas Complementarias:

- [1] Alonso, A., Fernández, J. y Gallastegui, I. (2005), *Econometría*, ed. Pearson: Prentice Hall.
- [2] Dougherty, Ch. (2006), *Introduction to Econometrics*, 3rd. Ed., Oxford University Press.
- [3] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). Análisis de regresión con gretl. Open Course Ware. UPV-EHU. (<http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-gretl/Courseisting>).
- [4] Esteban, M.V.; Moral, M.P.; Orbe, S.; Regúlez, M.; Zarraga, A. y Zubia, M. (2009). *Econometría Básica Aplicada con Gretl*. Sarriko On Line 8/09. <http://www.sarriko-online.com>. Publicación online de la Facultad de C.C. Económicas y Empresariales.
- [5] Esteban, M.V. (2007). *Estadística Actuarial: Regresión*. Material docente. Servicio de Publicaciones. Web de la asignatura.
- [6] Esteban, MV (2008). *Estadística Actuarial: Regresión Lineal*, Sarriko On Line 3/08. Publica-

ción on-line de la Facultad de CC. Económicas y Empresariales, UPV/EHU. <http://www.sarriko-online.com>.

[7] Esteban, M.V. (2007). Colección de ejercicios y exámenes. Material docente. Servicio de Publicaciones. Web de la asignatura.

[8] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[9] Greene, W. (1998), Análisis Econométrico, Ed. Prentice Hall, 3ª edición.

[10] Gujarati, D. (2004), Econometría, ed. McGraw-Hill, 4ª edición.

[11] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

[12] Verbeek, M. (2004). A Guide to Modern Econometrics. Wiley.