Econometría

Autor:
M. Victoria Esteban González

Departamento de Economía Aplicada III. Econometría y Estadística
Facultad de Economía y Empresa
Universidad del País Vasco/Euskal Herriko Unibertsitatea

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 los Modelos de Regresión Lineal Simple y General. El objetivo fundamental del curso es que, al final del mismo, los estudiantes sean capaces de utilizar un 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 seis temas. En el tema introductorio se define la disciplina de la Econometría y se introducen los conceptos básicos relacionados con un modelo econométrico. El segundo tema introduce la especificación del Modelo de Regresión Lineal Simple 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 simple. El estimador de referencia es el estimador de Mínimos Cuadrados Ordinarios. Se estudiaran sus propiedades y cómo compararlo con otros estimadores de interés. El tema cuatro se ocupa del contraste de hipótesis. El tema cinco analiza la especificación, estimación e inferencia en el Modelo de Regresión Lineal General. También se estudian 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 seis 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 *Econometría* del Doble Grado en Administración y Dirección de Empresas y Grado en Derecho así como del Grado en Administración y Dirección de Empresas. Sin embargo, dada su temática básica de estadística y análisis de regresión pueden ser útiles en asignaturas afines de los

Grados en Economía, Marketing, Fiscalidad y Administración Pública. Así mismo sirven de apoyo a estudiantes de master por ejemplo el Master en Ciencias Actuariales y Financieras o el Master Universitario en Banca y Finanzas Cuantitativas.

Las competencias específicas de la asignatura y la evaluación

La asignatura de Econometría 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 las variables.

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[1] :

CT1. Capacidad para emitir juicios razonados apoyándose en los datos obtenidos (M03CM02).

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 (M03CM05).

CT3. Capacidad para la comunicación escrita y oral con fluidez (M03CM09)s.

CT4. Capacidad para el pensamiento analítico y la reflexión crítica(M03CM11).

CT5. Capacidad para comunicarse en una lengua extranjera, preferentemente en inglés, francés o alemán (M03CM13).

Los Resultados de Aprendizaje que se pretende que el alumnado adquiera con los contenidos y metodología de la asignatura son los siguientes:

[1]Los códigos de las competencias transversales se corresponden con las del Módulo Avance en la Administración y Dirección de empresas recogidas en la memoria del grado (www.ehu.eus).

- Comprender la especificación del modelo de regresión lineal y, en particular, el significado y las implicaciones de los supuestos básicos (C1).

- Saber incorporar en el modelo de regresión variables cuantitativas y cualitativas (C1).

- Interpretar los coeficientes del modelo de regresión, incluyendo los de especificaciones no lineales en las variables (C1, C3).

- Organizar y sistematizar información estadística relevante (C3, C4).

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

- Estimar el modelo de regresión por Mínimos Cuadrados Ordinarios (C2).

- Realizar contrastes de hipótesis sobre la relación económica propuesta (C3).

- Predecir valores de interés con un modelo econométrico (C3).

- Comprobar la validez de algunos de los supuestos básicos del modelo de regresión y aprender a modificar el análisis en caso de incumplimiento (C3).

- Seleccionar entre especificaciones alternativas en base a las propiedades de los estimadores (C3).

- Interpretar adecuadamente los resultados obtenidos en la estimación del modelo econométrico (C3, C4).

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[2].
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:

$$http://gretl.sourceforge.net/gretl\_data.html$$

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

$$http://gretl.sourceforge.net/gretl\_espanol.html$$

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.learneconometrics.com/gretl.html

[2]Acrónimo de *Gnu Regression, Econometric and Time Series* (Biblioteca Gnu de Regresión Econometría y Series Temporales)

Contenido

Figuras

Tablas

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 simple**. Este modelo explica el comportamiento de una única variable económica mediante el comportamiento de otra variable. Una vez comprendamos los mecanismos de funcionamiento y relaciones entre las variables de este modelo pasaremos a estudiar un modelo más amplio, el **modelo de regresión lineal general**. A diferencia del Modelo de Regresión Lineal Simple 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.

Objetivo de aprendizaje:
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:

- Stock and Watson, J. M. (2012). Cap.1.
- Wooldridge, J.M. (2006). Cap.1.

1.1. ¿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)*

1.2. 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:

$$V_t = f(p_t, pc_t, c_t) \tag{1.1}$$

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:

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 pc_t + \beta_4 c_t \tag{1.2}$$

Los parámetros o coeficientes de cada variable se representan por β_1, β_2 y β_3. El coeficiente β_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:

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 cp_t + \beta_4 c_t + u_t \tag{1.3}$$

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 β_1, β_2, β_3 y β_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.

1.3. 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 sí...?, en definitiva debe servirnos para dar consejos de política económica.

1.4. Tipología de datos y variables en Econometría

El modelo econométrico genérico completamente especificado tiene la siguiente forma:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_K X_{Kt} + u_t \qquad t = 1, 2, \ldots, T \tag{1.4}$$

Donde Y es la variable a explicar o variable endógena, X_2, X_3, ..., 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 $\beta_k \quad k = 1, 2, \ldots, K$ son los parámetros a estimar, que se suponen constantes. Además es de interés notar que el parámetro β_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 $\beta_k \quad k = 1, 2, \ldots, K$. 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:

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

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

1.4.1. 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

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

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:

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

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

   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 econométrico 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.

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 cp_t + \beta_4 c_t + u_t \tag{1.6}$$

   En el modelo anterior $\beta_1, \beta_2, \beta_3$ y β_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:

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

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:

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

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 basado en datos de corte transversal*: son siempre estáticos.

- **Parámetro, estimador y estimación:**
  En el modelo:
  $$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad t = 1, 2, \ldots, T$$
  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 β_1 y β_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 β_1 y β_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 β_2 sería:

β_2	parámetro poblacional
$\hat{\beta}_2$	estimador
0,5	estimación

  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:
  $$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad t = 1, 2, \ldots, T$$
  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.

1.4.2. 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*.

1.5. 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 postcript (.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 clickear 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 $Z_t = 4 + \epsilon_t \qquad \epsilon_t \sim N(0,1)$. Permite los operadores,

    `+, -, *, /, ^`

    (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:

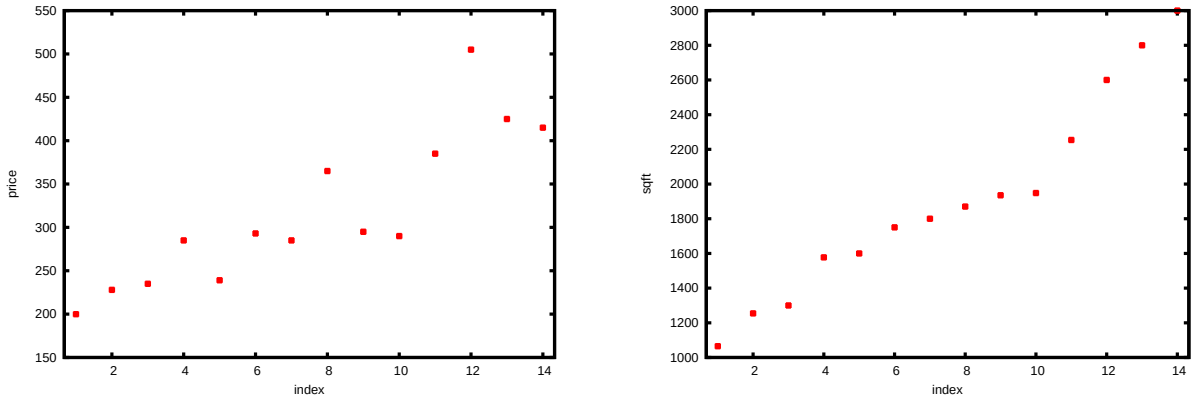


Figura 1.1: Gráficos de las observaciones para las variables *price* y $sqft$

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*.

1.6. Bibliografía del tema

Referencias bibliográficas básicas:

• Teórica:

[1] Stock, James H. y Mark Watson (2012). Introducción a la Econometría. Pearson.

[2] Wooldridge, J.M. (2006). Introducción a la Econometría. Ed. Thomson Learning, 2ª edición.

• 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[4] Gujarati, D. y Porter, D.C. (2010). Econometría. Editorial McGraw-Hill, Madrid. 5ª edición.

[5] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

Tema 2

Modelo de Regresión Lineal Simple. Especificación

En este tema nos ocuparemos de analizar las relaciones entre dos variables y nuestro objetivo fundamental será explicar el comportamiento de una variable, que llamaremos variable a explicar, mediante otra variable económica, que llamaremos explicativa. 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 Simple, 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. Comprender la especificación del modelo de regresión lineal y, en particular, el significado y las implicaciones de los supuestos básicos (C1).
3. Interpretar los coeficientes del modelo de regresión, incluyendo los de especificaciones no lineales en las variables (C1).
4. Saber incorporar en el modelo de regresión variables cuantitativas y cualitativas (C1).

5. Organizar y sistematizar información estadística relevante (C4).
6. 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:

- Stock and Watson, J. M. (2012). Cap. 4.
- Wooldridge, J.M. (2006). Cap. 2

2.1. Especificación del Modelo de Regresión Lineal Simple

Supongamos que nos interesa conocer la relación que hay entre el precio de una vivienda y su superficie. Se trata de cuantificar la influencia que tiene el tamaño de una vivienda en la determinación de su precio de venta mediante un modelo de regresión lineal simple. En este capítulo vamos a especificar, estimar y analizar el ***modelo de regresión lineal simple***. La teoría necesaria para este fin será ilustrada mediante el estudio simultáneo del conjunto de datos *data3-1* disponible en *gretl* dentro del conjunto de datos correspondiente a Ramanathan. Este fichero contiene el precio de venta y la superficie de 14 viviendas vendidas en el área de San Diego. Vamos a comenzar realizando un **análisis gráfico**.

1. Accedemos a este conjunto de datos en *Archivo* → *Abrir datos* →*Archivo de muestra* y en la carpeta de datos de *Ramanathan* seleccionamos *data3-1 House prices and sqft*:

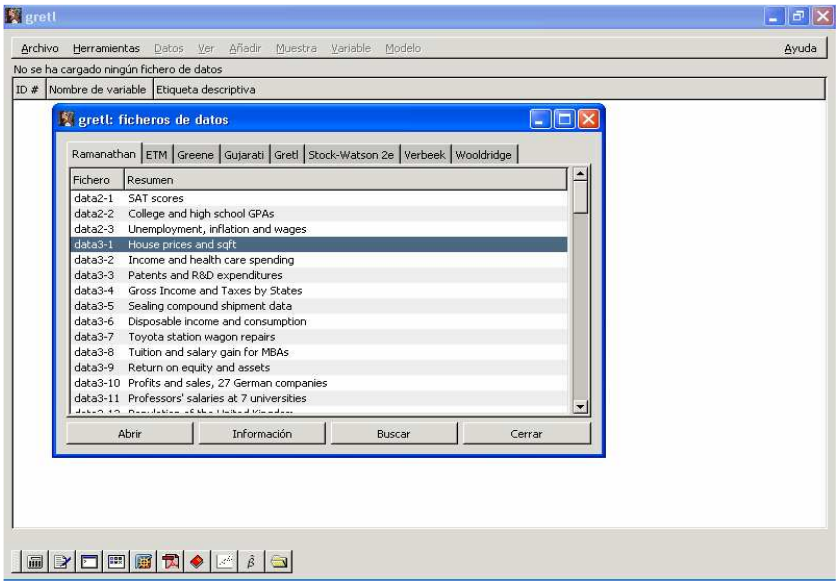


Figura 2.1: Selección de un fichero de muestra

   Se abre un fichero que contiene tres variables, *const*, *price* y *sqft*. La Tabla 2.1 muestra los valores disponibles para cada variable.

2. En *Datos* →*Leer información* aparece la siguiente descripción del conjunto de datos:

```
DATA3-1: Precio de venta y superficie hábil de viviendas
 unifamiliares en la comunidad universitaria de San Diego
 en 1990.
 price = Precio de venta en miles de dólares (Rango 199.9 - 505)
 sqft  = Pies cuadrados de área habitable (Rango 1065 - 3000)
```

3. Seguidamente seleccionamos ambas variables y en *Datos* →*Mostrar valores* vemos los valores muestrales de las variables. Estos valores han sido recogidos en la Tabla 2.1.

i	P_i	SQFT	i	P	SQFT
1	199,9	1065	8	365,0	1870
2	228,0	1254	9	295,0	1935
3	235,0	1300	10	290,0	1948
4	285,0	1577	11	385,0	2254
5	239,0	1600	12	505,0	2600
6	293,0	1750	13	425,0	2800
7	285,0	1800	14	415,0	3000

Tabla 2.1: Conjunto de datos incluidos en *data3.1 House prices and sqft*

4. Abrimos el diagrama de dispersión entre las dos variables (ver la Figura 2.2). En él observamos una relación lineal positiva entre P y $SQFT$.

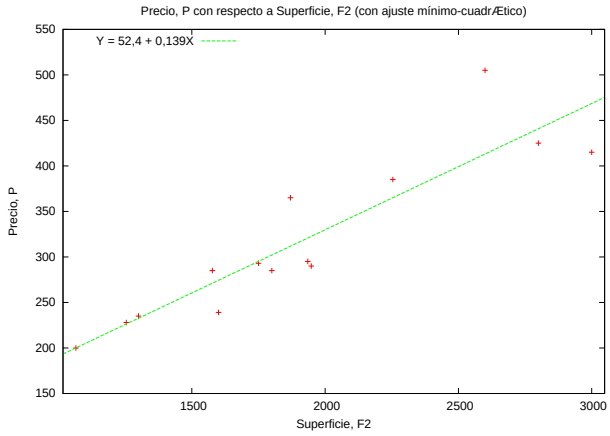


Figura 2.2: Diagrama de dispersión precio-superficie de viviendas

Un modelo sencillo que recoge una relación lineal causa-efecto entre la superficie y el precio de una vivienda es: $P_i = \beta_1 + \beta_2 SQFT_i + u_i$.

Esto quiere decir que el precio de una vivienda depende *únicamente* de su superficie y, por lo tanto, dos viviendas de igual tamaño deben tener *exactamente* el mismo precio. Esta hipótesis es poco realista porque diferencias en otras características, como la orientación de la casa o su estado de conservación, también influyen en su precio. Este modelo que recoge una relación lineal entre únicamente dos variables se denomina modelo de regresión lineal simple.

2.2. Elementos del modelo de regresión simple

El **Modelo de Regresión Lineal Simple (MRLS)** relaciona dos variables de forma lineal,

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad i = 1, \ldots, N \tag{2.1}$$

donde:

- Y es la ***variable a explicar***, ***variable dependiente o endógena***, es decir, la variable que estamos interesados en explicar.

- X es la ***variable explicativa***, ***variable independiente o exógena***.

- La ordenada β_1 y la pendiente β_2 del modelo son los ***coeficientes de la regresión***, son parámetros poblacionales desconocidos. Si definimos K como el *número de coeficientes desconocidos a estimar*, en el modelo de regresión simple tenemos $K = 2$ coeficientes a estimar.

- u es el término de error, variable aleatoria o ***perturbación***.

- El subíndice i denota ***observación***. En general, el subíndice i será empleado cuando la muestra contenga datos de sección cruzada y el subíndice t cuando tengamos observaciones correspondientes a series temporales, aunque esto no es de especial relevancia.

- N es el ***tamaño muestral***, número de observaciones disponibles de las variables de estudio (Y, X). Cuando tratemos con datos temporales T denotará el tamaño muestral.

El error u_i se introduce por varias razones, entre las cuales tenemos:

- Efectos impredecibles, originados por las características de la situación económica o del contexto de análisis, y efectos no cuantificables derivados de las preferencias y los gustos de los individuos o entidades económicas.

- Errores de medida producidos a la hora de obtener datos sobre las variables de interés.

- Errores de especificación ocasionados por la omisión de alguna variable explicativa o bien, por las posibles no linealidades en la relación entre X e Y.

Modelo para la relación precio-tamaño del piso. En este caso planteamos el siguiente modelo de regresión lineal:

$$P_i = \beta_1 + \beta_2\ SQFT_i + u_i \qquad i = 1, \ldots, N \tag{2.2}$$

donde

- P_i es la observación i de la variable dependiente (endógena o a explicar) *precio de venta* de un piso en miles de dólares.

- $SQFT_i$ es la observación i de la variable independiente (exógena o explicativa) *área habitable* del piso en pies cuadrados.

- Los dos coeficientes a estimar son β_1 y β_2, y sospechamos que al menos β_2 tiene valor positivo ya que a mayor superficie habitable de la vivienda su precio lógicamente se esperará sea mayor.

- En este modelo el término de error o perturbación u_i recogería características específicas de los pisos: lugar en el que se sitúa, orientación de la casa, vistas, etc., es decir, características que diferencian el precio de los pisos que tienen la misma superficie habitable.

Un primer objetivo del análisis econométrico es conocer β_1 y β_2, que son los parámetros de la relación entre P y $SQFT$. Del total de viviendas del área objeto de estudio, tenemos una muestra con datos de N= 14 pisos. Por tanto, el objetivo del estudio es *inferir*, a partir de la muestra, la relación precio-tamaño de una vivienda en la población. Para llevar a cabo esta inferencia es necesario determinar la naturaleza aleatoria de las variables que intervienen en el estudio.

Representación del MRLS en forma matricial El modelo

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad i = 1, 2, \ldots, N \tag{2.3}$$

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

$$\left\{\begin{array}{lcll} Y_1 & = & \beta_1 + \beta_2 X_1 + u_1 & i = 1 \\ Y_2 & = & \beta_1 + \beta_2 X_2 + u_2 & i = 2 \\ \vdots & & \vdots & \\ Y_i & = & \beta_1 + \beta_2 X_i + u_i & i = i \\ \vdots & & \vdots & \\ Y_N & = & \beta_1 + \beta_2 X_N + u_N & i = N \end{array}\right.$$

o bien en forma matricial como

$$\underset{(N\times 1)}{Y} = \underset{(N\times K)}{X}\,\underset{(K\times 1)}{\beta} + \underset{(N\times 1)}{u}$$

donde $K = 2$ y

$$\underset{(N\times 1)}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{bmatrix} \quad \underset{(N\times K)}{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_i \\ \vdots & \vdots \\ 1 & X_N \end{bmatrix} \quad \underset{(K\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \underset{(N\times 1)}{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix}$$

Ejemplo 2.1

Siguiendo con el modelo del precio de una vivienda y con los datos recogidos en la Tabla 2.1, tenemos:

$$Y = \begin{bmatrix} 199,9 \\ 228,0 \\ 235,0 \\ 285,0 \\ 239,0 \\ 293,0 \\ 285,0 \\ 365,0 \\ 295,0 \\ 290,0 \\ 385,0 \\ 505,0 \\ 425,0 \\ 415,0 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1065 \\ 1 & 1254 \\ 1 & 1300 \\ 1 & 1577 \\ 1 & 1600 \\ 1 & 1750 \\ 1 & 1800 \\ 1 & 1870 \\ 1 & 1935 \\ 1 & 1948 \\ 1 & 2254 \\ 1 & 2600 \\ 1 & 2800 \\ 1 & 3000 \end{bmatrix}$$

2.2.1. 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 es una variable no observable cuyo valor medio condicionado en X es cero para todo i, $E(u_i|X_i) = 0 \quad \forall i$. La perturbación mide las diferencias con respecto a un promedio, $u_i = Y_i - E(Y_i|X_i)$ 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 $E(u_i|X_i) = 0 \quad \forall i$, cuando miramos al modelo en forma matricial escribimos esta hipótesis como $E(u|X) = \vec{0}$:

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

   - $Var(u_i) = E(u_i^2|X_i) = \sigma_u^2 = \sigma^2 \quad \forall i$ es decir la varianza de la perturbación es desconocida e igual a σ^2 para todas las observaciones. Estamos suponiendo igual dispersión o variabilidad. A esta hipótesis se le conoce con el nombre de *Homocedasticidad*. El caso contrario, cuando la dispersión varía a lo largo de la muestra se denomina *Heterocedasticidad*. La Figura 2.3 ilustra ambas situaciones:

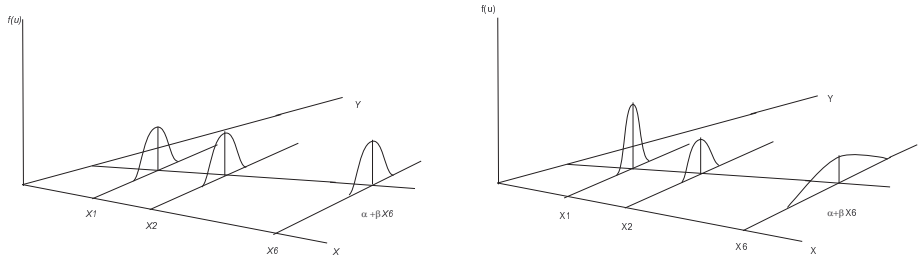


Figura 2.3: Perturbaciones homocedásticas versus heterocedásticas

Hay que notar que generalmente σ^2 será desconocida.

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

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

$$E(uu'|X) = \begin{bmatrix} E(u_1^2|X) & E(u_1u_2'|X) & \dots & E(u_1u_N'|X) \\ E(u_2u_1'|X) & E(u_2^2|X) & \dots & E(u_2u_N'|X) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_Nu_1'|X) & E(u_Nu_2'|X) & \dots & E(u_N^2|X) \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I_N$$

  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 condicionada en X normal.

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

  donde estamos escribiendo la distribución del vector de perturbaciones u y decimos que las perturbaciones siguen una distribución condicionada en X normal, idéntica e independientemente distribuidas, de media cero y varianza constante igual a σ^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 $u_i|X_i \sim N(0, \sigma^2)$.

Estas propiedades pueden también escribirse conjuntamente como

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

2. Hipótesis sobre las variables exógenas X.
   - Condicionamos el análisis a unos valores dados de X. Este proceder es similar a considerar las variables como no aleatorias o regresores fijos.
   - La matriz X es de rango completo e igual a K (en el MRLS $K = 2$) 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.
4. Los coeficientes permanecen constantes a lo largo de toda la muestra.

2.3. Función de Regresión Poblacional. Interpretación de los coeficientes.

Abreviadamente, el modelo con las hipótesis básicas mencionadas se escribe:

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \quad u_i|X \sim NID(0, \sigma^2) \quad \forall i$$

Dado el **supuesto básico** $E(u|X) = 0$:

$$\begin{aligned} E(Y_i|X) &= E(\beta_1 + \beta_2 X_i + u_i|X) \\ &= \beta_1 + \beta_2 X_i + \underbrace{E(u_i|X)}_{=0} = \\ &= \beta_1 + \beta_2 X_i. \end{aligned}$$

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

- $\beta_1 = E(Y_i|X_i = 0)$: valor medio o esperado de la variable endógena cuando el valor que toma la variable exógena es cero.

- $\beta_2 = \dfrac{\Delta E(Y_i)}{\Delta X_i} = \dfrac{\partial E(Y_i)}{\partial X_i}$: Incremento (o decremento) en el valor esperado o valor medio de Y_i cuando la variable explicativa X se incrementa en una unidad. La pendiente mide el efecto de un aumento marginal en la variable explicativa sobre $E(Y_i)$, un aumento unitario en la variable explicativa conlleva un aumento medio de β_2 unidades en la variable endógena.

→ Así, volviendo a nuestro ejemplo tenemos que:

$\beta_1 = E(P_i|SQFT_i = 0)$ es el precio medio de venta en miles de dólares cuando el piso dispone de una superficie de cero pies habitables, que también puede ser considerado como precio mínimo de partida. En este caso, esperaríamos un coeficiente nulo dado que no tiene sentido hablar de un piso sin superficie hábil o bien un precio de partida positivo. No obstante, aunque en este contexto la ordenada no tiene en principio mucho sentido, no debemos de eliminarla a la ligera en aras de obtener resultados fáciles de interpretar.

$\beta_2 = \dfrac{\Delta E(P_i)}{\Delta SQFT_i} = \dfrac{\partial E(P_i)}{\partial SQFT_i}$ indica que, cuando un piso aumenta su superficie hábil en un pie cuadrado, su precio medio aumenta en β_2 miles \$.

Ejemplo 2.2

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:

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

β_2 recoge el incremento en el consumo personal o consumo medio 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. β_1 es el valor esperado o medio del consumo cuando el valor del PIB es cero.

Ejemplo 2.3

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[1]. 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).

[1]Fuente: Statistical Abstract of U.S. (1995), recogida en Ramanthan, Ramu (2002) *Introductory econometrics with applications.* Fichero de datos data8-2.gdt.

Un modelo para analizar si la renta disponible agregada explica el gasto agregado en transporte urbano es el siguiente[2]:

$$EXPTRAV_i = \beta_1 + \beta_2 INCOME_i + u_i \qquad i = 1, \ldots, 51 \tag{2.4}$$

El parámetro β_1 recoge el valor esperado o medio del gasto en transporte cuando la renta es cero, $\beta_1 = E(EXPTRAV_i|INCOME_i = 0)$. La pendiente β_2 recoge el incremento en el valor esperado o valor medio del gasto en transporte cuando la renta se incrementa en una unidad, es este caso cuando se incrementa en un billón de dólares, $\beta_2 = \frac{\partial E(EXPTRAV_i)}{\partial INCOME_i}$. Esperaríamos signo positivo.

Ejemplo 2.4

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

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

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

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

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 o salario medio cuando el individuo es hombre. Esperaríamos signo positivo.
- $E(W_i|S_{2i} = 1) = \beta_1 + \beta_2$ es el salario esperado o salario medio de una mujer. Luego β_2 es el incremento o decremento en el salario medio para un individuo por el hecho de ser mujer. Por tanto β_2 recoge el efecto diferencial en el salario medio 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.

Algunas consideraciones sobre la linealidad en parámetros Cuando decimos que el MRLS es un modelo lineal queremos decir que Y o alguna transformación de Y es lineal en X o en alguna transformación lineal en X. Hay dos tipos de linealidad, linealidad en variables y linealidad en parámetros. Dado que estamos interesados sólo en la linealidad en parámetros también serán considerados lineales los siguientes modelos:

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

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

[2]Son datos de sección cruzada luego utilizamos el subíndice $i = 1, \ldots, N$.

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:
$$Y_i = \beta_1 + \beta_2 LnX_i + u_i \tag{2.5}$$
donde $\beta_2 = (LnB)^{-1}$ y $\beta_1 = (\frac{LnA}{LnB})$ a esta transformación se le llama *semilogarítmica.*

2. Sea el modelo:
$$Y_i = AX_i^B u_i \longrightarrow LnY_i = \beta_1 + \beta_2 LnX_i + u_i \tag{2.6}$$
donde $\beta_1 = LnA$, a esta transformación se le llama *doblemente logarítmica.*

En este modelo en el que todas las variables están medidas en logaritmos, el parámetro de pendiente además de recibir la interpretación habitual pueden interpretarse en términos de elasticidad:
$$\beta_2 = \frac{\partial E(LnY_i)}{\partial LnX_i} = \frac{\partial E(Y_i)}{\partial X_i}\frac{X_i}{Y_i}$$

Es importante notar que para la ecuación (2.5) la interpretación de los parámetros como elasticidades no es posible ya que al no estar la variable Y_i en logaritmos:
$$\beta = \frac{\partial E(Y_i)}{\partial LnX_i} = \frac{\partial E(Y_i)}{\partial X_i}X_i$$

2.4. Utilización de variables explicativas cualitativas

En los ejemplos anteriores 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 una variable cualitativa.

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[3].

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

En este tema ya hemos trabajado con ellas, el Ejemplo 2.4 especificamos la función de salario en función del regresor cualitativo sexo e interpretamos sus parámetros. 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.

En el modelo (2.2) el precio de la vivienda depende exclusivamente de su superficie. Sin embargo hay otras características que pueden influir en el precio como la existencia de piscina, de garaje, el número de habitaciones y/o de baños. Supongamos que tenemos información sobre si la vivenda tiene piscina o no. Podríamos especificar un modelo para el precio de la vivienda suponiendo que este dependa exclusivamente de si la vivienda tiene o no piscina. Esta variable tiene dos categorías o estados de la naturaleza, tener o no piscina, que podemos recoger con las siguientes variables ficticias que dividen la muestra en dos grupos y a las que asignamos un valor arbitrario a cada clase[4]:

$$POOL_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene piscina} \\ 0 & \text{en caso contrario} \end{cases}$$
$$NOPOOL_i = \begin{cases} 1 & \text{si la vivienda i-ésima no tiene piscina} \\ 0 & \text{en caso contrario} \end{cases}$$

Y especificar el modelo:

$$P_i = \beta_1 + \beta_2\ POOL_i + u_i \qquad i = 1, \ldots, N \tag{2.7}$$

Tal que si $E(u_i|X) = 0 \quad \forall i$ la FRP del modelo es $E(P_i|X) = \beta_1 + \beta_2\ POOL_i$
$\longrightarrow$ Si la vivienda no tiene piscina: $E(P_i|POOL_i = 0) = \beta_1$
$\longrightarrow$ Si la vivienda tiene piscina: $E(P_i|POOL_i = 1) = \beta_1 + \beta_2$
Luego β_1 es el precio medio de una vivienda sin piscina, $\beta_1 + \beta_2$ es el precio medio de una vivienda con piscina y β_2 es el diferencial en el precio medio de una vivienda por tener piscina relativamente a no tenerla.

El modelo (2.7) da lugar a dos ecuaciones:

$$P_i = \beta_1 + u_i \quad i = 1, \ldots, N_{NP} \qquad \text{para las viviendas sin piscina}$$
$$P_i = \beta_1 + \beta_2 + u_i \quad i = 1, \ldots, N_N \qquad \text{para las viviendas con piscina}$$

[3]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.

[4]Elegir los valores (0,1) es muy cómodo pero podríamos elegir otros, por ejemplo:
$POOL_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene piscina} \\ 0 & \text{en caso contrario} \end{cases}$ $NOPOOL_i = \begin{cases} 2 & \text{si la vivienda i-ésima no tiene piscina} \\ 0 & \text{en caso contrario} \end{cases}$

En el modelo (2.7) el **grupo de referencia**, el recogido en el término independiente son las viviendas que no tienen piscina. Podríamos haber definido el modelo en base a la variable $NOPOOL$.

• Alternativa de especificación del modelo (2.7):

$$P_i = \alpha_1 NOPOOL_i + \alpha_2 POOL_i + u_i \qquad i = 1, \ldots, N \tag{2.8}$$

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

$\alpha_1 = E(P_i|NOPOOL_i = 1; POOL_i = 0)$ es el precio medio de una vivienda sin piscina
$\alpha_2 = E(P_i|NOPOOL_i = 0; POOL_i = 1)$ es el precio medio de una vivienda con piscina

por tanto estos coeficientes recogen el precio medio de la vivienda dentro del grupo.

En este caso el modelo (2.8) da lugar a dos ecuaciones:

$$\begin{aligned} P_i &= \alpha_1 + u_i \quad i = 1, \ldots, N_P \qquad \text{para las viviendas con piscina} \\ P_i &= \alpha_2 + u_i \quad i = 1, \ldots, N_{NP} \qquad \text{para las viviendas sin piscina} \end{aligned}$$

La relación entre los parámetros del modelo (2.7) y los del modelo (2.8) es la siguiente:

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

¿Cómo sería la matriz X en los modelos anteriores? Supongamos que disponemos de información sobre qué viviendas tiene piscina en la muestra del fichero de datos *data3-1.gdt* y es la siguiente:

i	P_i	$SQFT_i$	Piscina	$POOL_i$	$NOPOOL_i$
1	199,9	1065	si	1	0
2	228,0	1254	no	0	1
3	235,0	1300	si	1	0
4	285,0	1577	no	0	1
5	239,0	1600	no	0	1
6	293,0	1750	no	0	1
7	285,0	1800	no	0	1
8	365,0	1870	si	1	0
9	295,0	1935	no	0	1
10	290,0	1948	no	0	1
11	385,0	2254	si	1	0
12	505,0	2600	si	1	0
13	425,0	2800	no	0	1
14	415,0	3000	no	0	1

Luego para los modelos (2.7) y (2.8) respectivamente tendríamos:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \qquad X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Ejemplo 2.5

Ejemplo para la función de salario. Por ejemplo si queremos estudiar la dependencia del salario (W_i) con respecto al sexo del individuo definiremos dos variables ficticias:

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

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.

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 y podemos especificar el siguiente modelo :

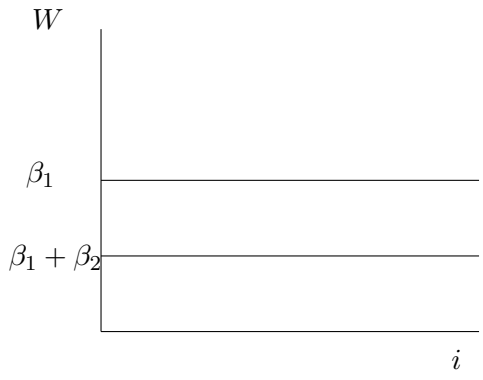
$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \qquad i = 1, \ldots, N_H + N_M \qquad u_i \sim NID(0, \sigma^2) \qquad (2.9)$$

Hay que notar que el modelo (5.9) da lugar a dos ecuaciones:

$$W_i = \beta_1 + u_i \quad i = 1, \ldots, N_H \qquad \text{para los hombres}$$
$$W_i = \beta_1 + \beta_2 + u_i \quad i = 1, \ldots, N_M \qquad \text{para las mujeres}$$

$$\beta_1 = E(W_i | S_{2i} = 0) \text{ es el salario medio de un hombre}$$
$$\beta_1 + \beta_2 = E(W_i | S_{2i} = 1) \text{ es el salario medio de una mujer}$$

β_1 es el salario medio cuando el individuo es hombre, $\beta_1 + \beta_2$ es el salario esperado de una mujer y β_2 recoge el efecto diferencial en el salario medio entre hombres y mujeres. Si no existiera discriminación salarial por sexo, es decir si hombres y mujeres tuvieran el mismo salario medio, su valor sería cero. En el gráfico podemos observar estos efectos donde se supone que β_2 es negativo por razones didácticas.



• Alternativa de especificación del modelo (5.9):

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + u_i \qquad i = 1, \ldots, N_H + N_M \tag{2.10}$$

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

$$\alpha_1 = E(W_i | S_{1i} = 1; S_{2i} = 0) \text{ es el salario medio de un hombre}$$
$$\alpha_2 = E(W_i | S_{1i} = 0; S_{2i} = 1) \text{ es el salario medio de una mujer}$$

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

En este caso el modelo (5.10) da lugar a dos ecuaciones:

$$W_i = \alpha_1 + u_i \quad i = 1, \ldots, N_H \qquad \text{para los hombres}$$
$$W_i = \alpha_2 + u_i \quad i = 1, \ldots, N_M \qquad \text{para las mujeres}$$

La relación entre los parámetros del modelo (5.9) y los del modelo (5.10) 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:

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

donde W_i es el salario del individuo i y

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

¿Qué diferencia hay entre ésta especificación y la especificación del modelo (5.9)?

2.5. Bibliografía del tema

Referencias bibliográficas básicas:

• Teórica:

[1] Stock, James H. y Mark Watson (2012). Introducción a la Econometría. Pearson.

[2] Wooldridge, J.M. (2006). Introducción a la Econometría. Ed. Thomson Learning, 2ª edición.

• 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[4] Gujarati, D. y Porter, D.C. (2010). Econometría. Editorial McGraw-Hill, Madrid. 5ª edición.

[5] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

Tema 3

Modelo de Regresión Lineal Simple. Estimación

En este tema nos ocuparemos de estimar el Modelo de Regresión Lineal Simple. 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.

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. Organizar y sistematizar información estadística relevante (C3).
4. 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:

- Stock and Watson, J. M. (2012). Cap. 4.
- Wooldridge, J.M. (2006). Caps. 2

3.1. Estimación por Mínimos Cuadrados Ordinarios

Una vez descrito el ámbito en el que nos vamos a mover, vamos a obtener un estimador adecuado de los coeficientes del modelo de regresión simple: **el estimador de mínimos cuadrados ordinarios**. En primer lugar, obtendremos el estimador y, a continuación, justificaremos su uso en base a sus propiedades. El modelo simple (2.1) nos indica que cada observación Y_i es una realización de una variable que tiene dos componentes: uno que depende del valor del regresor X_i, cuyo valor observamos, y un componente residual que no observamos. El MRLS desarrolla un sistema de N ecuaciones:

$$\begin{cases} Y_1 = \beta_1 + \beta_2 X_1 + u_1 \\ \vdots \\ Y_i = \beta_1 + \beta_2 X_i + u_i \\ \vdots \\ Y_N = \beta_1 + \beta_2 X_N + u_N \end{cases}$$

La Figura 3.1 representa gráficamente una posible muestra. Los puntos (Y_i, X_i) se sitúan o distribuyen alrededor de la recta $\beta_1 + \beta_2 X_i$. La desviación de cada punto respecto a esta *recta central* viene dada por el valor que tome el término de error no observable u_i. Por ejemplo, en la Figura 3.1, la perturbación es positiva para la primera observación, de modo que Y_1 se encuentra por encima de la recta central. Por otro lado, el punto (Y_2, X_2) se encuentra por debajo de la recta central, es decir, u_2 toma un valor negativo.

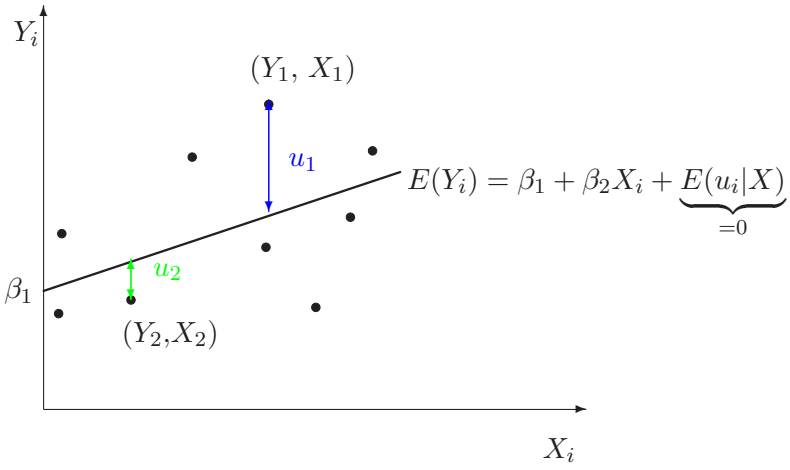


Figura 3.1: Modelo de regresión simple

• Nuestro **objetivo** es estimar los parámetros desconocidos β_1 y β_2 de

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad i = 1, 2, \ldots, N$$

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

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

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad i = 1, 2, \ldots, N$$

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

a la cual denominamos ***Función de Regresión Muestral*** (FRM). La FRM es una estimación de la FRP. Dado que se obtiene para una muestra dada, para cada muestra tendremos una FRM distinta. En la FRM $\hat{\beta}_1$ y $\hat{\beta}_2$ son los estimadores de β_1 y β_2.

- **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. Es una variable aleatoria no observable:

$$u_i = Y_i - E(Y_i|X_i) \qquad i = 1, 2, \ldots, N$$

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

  - 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[1]:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \qquad i = 1, 2, \ldots, N$$

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

    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 β_1 y β_2 introduce un error adicional. Es importante, por tanto, diferenciar y no confundir el residuo con la perturbación.

- **Representación gráfica**:

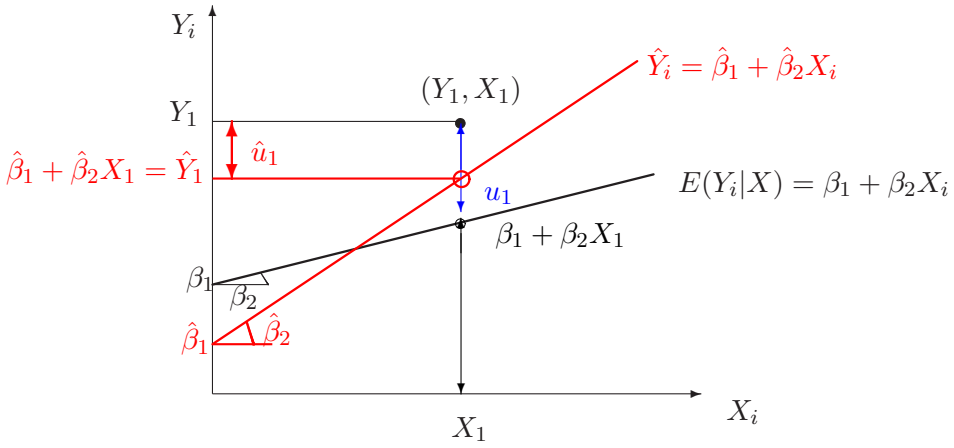


Figura 3.2: Función de regresión poblacional y función de regresión muestral

[1]Los residuos son a la FRM lo que las perturbaciones a la FRP. Sin embargo, no son buenos estimadores de las mismas porque no tienen las mismas propiedades. Tienen media cero pero son heterocedásticos y autocorrelados.

En la Figura 3.2 la función de regresión poblacional está trazada en color negro así como los coeficientes poblacionales, la ordenada (β_1) y la pendiente (β_2). Podemos ver que el valor Y_i se 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. El criterio de estimación mínimo-cuadrático

Dados el modelo y una muestra, debemos decidir cómo obtener la función de regresión muestral, es decir, cómo calcular las estimaciones $\hat{\beta}_1$ y $\hat{\beta}_2$ a partir de los datos. Un método muy utilizado por su sencillez y buenas propiedades es el método de mínimos cuadrados ordinarios. El estimador de ***Mínimos Cuadrados Ordinarios***, o MCO, de los parámetros β_1 y β_2 se obtiene de minimizar la suma de los residuos al cuadrado:

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{N} \hat{u}_i^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \tag{3.1}$$

Las expresiones del estimador de β_1 y β_2 se obtienen de las condiciones de primer orden, para lo cual igualamos las primeras derivadas a cero:

$$\begin{aligned} \frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \\ \frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}_2} &= -2 \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = 0 \end{aligned}$$

Así, obtenemos un sistema de ecuaciones, llamadas **ecuaciones normales** , que vienen dadas por:

$$\sum_{i=1}^{N} (\underbrace{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}_{\hat{u}_i}) = 0 \tag{3.2}$$

$$\sum_{i=1}^{N} (\underbrace{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i}_{\hat{u}_i X_i} = 0 \tag{3.3}$$

Las expresiones de los estimadores MCO para los coeficientes poblacionales β_1 y β_2 se obtienen de resolver las ecuaciones para $\hat{\beta}_1$ y $\hat{\beta}_2$:

$$\hat{\beta}_{2,MCO} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2} = \frac{S_{XY}}{S_X^2} \tag{3.4}$$

$$\hat{\beta}_{1,MCO} = \bar{Y} - \hat{\beta}_2 \bar{X} \tag{3.5}$$

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

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

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:

$$X'Y = X'X\hat{\beta}_{MCO}. \tag{3.6}$$

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

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y \tag{3.7}$$

en el que $X'X$ es una matriz de orden (2×2), $X'Y$ un vector de orden (2×1) y $\hat{\beta}$ un vector de orden (2×1), tales que para el MRLS

$$\underset{(2\times 2)}{X'X} = \begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \qquad \underset{(2\times 1)}{X'Y} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \qquad \underset{(2\times 1)}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}.$$

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.

3.2. La Función de Regresión Muestral. Interpretación de los coeficientes estimados por MCO

En la sección anterior hemos denotado a la Función de Regresión Muestral (FRM) como:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad i = 1, 2, \ldots, N$$
$$\hat{Y} = X\hat{\beta} \qquad \text{en forma matricial,}$$

Los coeficientes estimados tienen la siguiente interpretación:

- $\hat{\beta}_1 = \widehat{E}(Y_i | X_i = 0)$. Valor medio *estimado* de Y_i cuando la variable explicativa es cero.
- $\hat{\beta}_2 = \dfrac{\partial \widehat{E(Y_i)}}{\partial X_i} = \dfrac{\Delta \widehat{E(Y_i)}}{\Delta X_i}$. Incremento medio *estimado* (ó *decremento medio estimado*) en Y_i cuando la variable X se incrementa en una unidad.

Ejemplo 3.1

Siguiendo con el modelo del precio de una vivienda y con los datos recogidos en la Tabla 2.1, tenemos:

i	P_i	$SQFT_i$	$SQFT_i \times P_i$	$SQFT_i^2$	P2
1	199,9	1065	212893,5	1134225	39960,01
2	228,0	1254	285912	1572516	51984
3	235,0	1300	305500	1690000	55225
4	285,0	1577	449445	2486929	81225
5	239,0	1600	382400	2560000	57121
6	293,0	1750	512750	3062500	85849
7	285,0	1800	513000	3240000	81225
8	365,0	1870	682550	3496900	133225
9	295,0	1935	570825	3744225	87025
10	290,0	1948	564920	3794704	84100
11	385,0	2254	867790	5080516	148225
12	505,0	2600	1313000	6760000	255025
13	425,0	2800	1190000	7840000	180625
14	415,0	3000	1245000	9000000	172225
$\sum_{i=1}^{14}$	4444,9	26753	9095985,5	55462515	1513039,01

De donde:

$\bar{P} = \frac{\sum_{i=1}^{14} P_i}{N} = \frac{4444,9}{14} = 317,4928571$
$\overline{SQFT} = \frac{\sum_{i=1}^{14} SQFT_i}{N} = \frac{26753}{14} = 1910,928571$

$$\hat{\beta}_{2,MCO} = \frac{\sum_{i=1}^{N} SQFT_i P_i - N \times \overline{SQFT} \times \bar{P}}{\sum_{i=1}^{N} SQFT_i^2 - N \times \overline{SQFT}^2} = \frac{9095985,5 - 14 \times 317,49 \times 1910,92}{55462515 - 14 \times (1910,92)^2} = 0,1388$$

$$\hat{\beta}_{1,MCO} = \bar{P} - \hat{\beta}_2 \, \overline{SQFT} = 317,49 - 0,1388 \times 1910,92 = 52,3509$$

En forma matricial:

$$\hat{\beta}_{MCO} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum SQFT_i \\ \sum SQFT_i & \sum SQFT_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum P_i \\ \sum SQFT_i P_i \end{bmatrix} =$$

$$= \begin{bmatrix} 14 & 26753 \\ 26753 & 55462515 \end{bmatrix}^{-1} \begin{bmatrix} 4444,9 \\ 9095985,5 \end{bmatrix} =$$

$$= \begin{bmatrix} 0,9129 & -4,4036e-04 \\ -4,4036e-04 & 2,3044e-07 \end{bmatrix} \begin{bmatrix} 4444,9 \\ 9095985,5 \end{bmatrix} = \begin{bmatrix} 52,3509 \\ 0,1388 \end{bmatrix}$$

FRM: $\hat{P}_i = 52,3509 + 0,1388 \, SQFT_i$

$\widehat{\beta_1}$= 52,35 miles de dólares y la estimación de la pendiente es $\widehat{\beta_2} = 0,138750$ miles \$ por pie cuadrado. Es decir, cuando la superficie de la vivienda aumenta en un pie cuadrado, el precio medio de venta **estimado** aumenta en $\widehat{\beta_2} \times 1000 = 138,750$ dólares. La interpretación del término independiente estimado no tiene sentido salvo como precio de partida ya que indica que el precio medio estimado de una vivienda sin superficie es 52.350 dólares.

Algunas equivalencias de notación

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad i = 1, 2, \ldots, N \quad \Leftrightarrow \quad Y = X\beta + u$$

$$E(Y_i) = \beta_1 + \beta_2 X_i \quad i = 1, 2, \ldots, N \quad \Leftrightarrow \quad E(Y) = X\beta$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad i = 1, 2, \ldots, N \quad \Leftrightarrow \quad \hat{Y} = X\hat{\beta}$$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \quad i = 1, 2, \ldots, N \quad \Leftrightarrow \quad Y = X\hat{\beta} + \hat{u}$$

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

Ejercicio 3.1

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

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad t = 1, \ldots, T$$

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.

$$Y_t = \beta X_t + u_t \qquad t = 1, \ldots, T$$

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 $\hat{\beta}$.

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:

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

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 299591,3 billones de dólares.

Ejemplo 3.3

A continuación vamos a estimar el modelo donde suponemos que el precio de venta de una vivienda depende exclusivamente de si tiene piscina o no. En el tema anterior mostramos cómo especificar dicho modelo, recogido en la ecuación (2.7):

$$P_i = \beta_1 + \beta_2\ POOL_i + u_i \qquad i = 1, \ldots, N$$

$$\hat{\beta}_{MCO} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum POOL_i \\ \sum POOL_i & \sum POOL_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum P_i \\ \sum POOL_i P_i \end{bmatrix} =$$

$$= \begin{bmatrix} 14 & 5 \\ 5 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 4444{,}9 \\ 1689,9 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}1111 & -0{,}1111 \\ -0{,}1111 & 0{,}3111 \end{bmatrix} \begin{bmatrix} 4444{,}9 \\ 1689,9 \end{bmatrix} = \begin{bmatrix} 306{,}11 \\ 31{,}86 \end{bmatrix} = \begin{bmatrix} \bar{P}_{NP} \\ \bar{P}_P - \bar{P}_{NP} \end{bmatrix}$$

$$\text{FRM: } \hat{P}_i = 306{,}11 + 31{,}86 POOL_i$$

• Como alternativa de especificación propusimos la ecuación (2.8)

$$P_i = \alpha_1 NOPOOL_i + \alpha_2 POOL_i + u_i \qquad i = 1, \ldots, N$$

En este caso:

$$\hat{\beta}_{MCO} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \sum NOPOOL_i^2 & \sum NOPOOL_i\, POOL_i \\ \sum NOPOOL_i\, POOL_i & \sum POOL_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum NOPOOL_i P_i \\ \sum POOL_i P_i \end{bmatrix} =$$

$$= \begin{bmatrix} N_{NP} & 0 \\ 0 & N_P \end{bmatrix}^{-1} \begin{bmatrix} \sum NOPOOL_i P_i \\ \sum POOL_i P_i \end{bmatrix} = \begin{bmatrix} \frac{\sum NOPOOL_i P_i}{N_{NP}} \\ \frac{\sum POOL_i P_i}{N_P} \end{bmatrix} = \begin{bmatrix} \bar{P}_{NP} \\ \bar{P}_P \end{bmatrix} =$$

$$= \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 2755 \\ 1689{,}9 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}1111 & 0 \\ 0 & 0{,}2 \end{bmatrix} \begin{bmatrix} 2755 \\ 1689{,}9 \end{bmatrix} = \begin{bmatrix} 306{,}1111 \\ 337{,}9800 \end{bmatrix} = \begin{bmatrix} \bar{P}_{NP} \\ \bar{P}_P \end{bmatrix}$$

$$\text{FRM: } \hat{P}_i = 306{,}1111 NOPOOL_i + 337{,}98 POOL_i$$

En este modelo interpretamos los parámetros de la forma siguiente:

$\alpha_1 = E(P_i | NOPOOL_i = 1; POOL_i = 0)$ es el precio medio de una vivienda sin piscina
$\alpha_2 = E(P_i | NOPOOL_i = 0; POOL_i = 1)$ es el precio medio de una vivienda con piscina

por tanto estos coeficientes recogen el precio medio de la vivienda dentro del grupo.
Y hemos obtenido que:

$\hat{\alpha}_1 = \bar{P}_{NP}$ es el precio medio estimado de una vivienda sin piscina
$\hat{\alpha}_2 = \bar{P}_P$ es el precio medio estimado de una vivienda con piscina

por tanto, estos coeficientes estimados son la media muestral de los precios de las viviendas dentro del grupo.

Al ser la relación entre los parámetros del modelo (2.7) y los del modelo (2.8) la siguiente:

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

Tenemos:
$\hat{\beta}_1 = \hat{\alpha}_1 = \bar{P}_{NP} = 306,11$ y $\hat{\beta}_2 = \hat{\alpha}_2 - \hat{\alpha}_1 = \bar{P}_P - \bar{P}_{NP} = 337{,}98 - 306,11 = 31{,}86$ luego $\hat{\beta}_2$ es la diferencia entre las medias muestrales estimadas.

3.2.1. Propiedades de la Función de Regresión Muestral

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: $\hat{Y}'\hat{u} = 0$ ($\hat{u}'\hat{Y} = 0$).

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

   Por tanto los residuos están incorrelados con la variable explicativa y con la variable dependiente estimada.

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: $\sum_{i=1}^{N} \hat{u}_i = 0$. Por tanto la media muestral de los residuos es cero, $\bar{\hat{u}} = 0$

   $$X'\hat{u} = 0 \Leftrightarrow \begin{bmatrix} \sum_1^N \hat{u}_i \\ \sum_1^N X_i\hat{u}_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \sum_{i=1}^{N} \hat{u}_i = 0$$

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

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

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

   $$\sum_{i=1}^{N} \hat{u}_i = 0 \Leftrightarrow \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

   $$\sum Y_i - N\hat{\beta}_1 - \hat{\beta}_2 \sum X_i = 0$$

$$\begin{aligned}\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \\ \frac{1}{N}\sum Y_i &= \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{N}\sum X_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X}\end{aligned}$$

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.

3.3. Bondad del ajuste. Coeficiente de determinación.

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 o residual.

$$SCT = SCE + SCR$$

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,

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

$$SCT = SCE + SCR$$

siendo:

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.

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

Nuestro objetivo es evaluar como se ajusta el modelo estimado a los datos, esto es, cómo explican las variables explicativas del modelo en su conjunto conjunto, la variabilidad de la variable dependiente. Para ello debemos utilizar un estadístico que recoja en un único valor el ajuste del modelo de regresión lineal a los datos una vez que ha sido estimado por MCO. Este estadístico es el Coeficiente de determinación, y mide la variabilidad observada de la variable dependiente que explica el modelo en función de las variables explicativas.

Coeficiente 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. Por tanto el R^2 mide la proporción de la variabilidad observada de la variable dependiente Y que se ha podido explicar por incluir de forma lineal en el modelo la variable explicativa X. Normalmente se interpreta en porcentajes, por ejemplo, se dice que la regresión explica el $100 \times R^2$ por ciento de la variación observada en Y.
- A mayor R^2 mejor ajuste.
- Es fácil comprobar que:
  - El criterio mínimo-cuadrático equivale a maximizar R^2.
  - $R^2 = r^2_{Y\hat{Y}}$, mide la correlación entre el valor observado y el valor predicho o ajustado con la regresión. Como $0 \leq r^2_{Y\hat{Y}} \leq 1$, si $R^2 \simeq 0$ diremos que el ajuste es pobre y, por el contrario, será un buen ajuste cuando este estadístico esté próximo a la unidad.
    Esta propiedad no se cumple en modelos sin término independiente.

Coeficiente de correlación El coeficiente de correlación da una medida estandarizada de la relación lineal entre dos variables. Indica el sentido y el grado de la relación. Mide el grado de asociación lineal entre dos variables. El coeficiente de correlación lineal simple muestral para X e Y se define:

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

El coeficiente de correlación 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.

Además en el MRLS se puede de mostrar que $R^2 = r_{XY}^2$.

Ejemplo 3.4

Con los resultados de la regresión del modelo (2.2) y los datos del fichero *data3-1.gdt* calculamos el coeficiente de determinación:

$$\begin{aligned} SCT &= Y'Y - N\bar{Y}^2 = \sum P_i^2 - N\,\bar{P}^2 = 1513039,01 - 14 \times (317,49)^2 = 101814,9997 \\ SCR &= Y'Y - \hat{\beta}X'Y = \sum P_i^2 - \hat{\beta}X'Y = 1513039,01 - [52{,}3509 \quad 0{,}1388] \begin{bmatrix} 4444{,}9 \\ 9095985{,}5 \end{bmatrix} = \\ &= 1513039,01 - 1494765,4422 = 18273{,}5678 \end{aligned}$$

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{18273{,}5678}{101814,9997} = 0,8205$$

Podemos decir que este ajuste es bueno, ya que la variabilidad muestral de la superficie de la vivienda ($SQFT$) ha explicado el 82 % de la variabilidad muestral de los precios de venta de dichas viviendas (P).

3.4. La estimación MCO en Gretl

En esta sección se va a mostrar cómo utilizar *gretl* para estimar por MCO.
→ Como ejemplo, calcularemos las estimaciones MCO del modelo para el precio de la vivienda, $P_i = \beta_1 + \beta_2 SQFT_i + u_i$, con la muestra del fichero *datos3-1.gdt*. Una forma sencilla de obtener la FRM mínimo-cuadrática es realizar el diagrama de dispersión en el cual la recta de regresión aparece en la parte superior izquierda. En el ejemplo que nos ocupa tenemos que $\hat{\beta}_1 = 52,4$ y $\hat{\beta}_2 = 0,139$, como se puede ver en la Figura 2.2.

Cómo podemos obtener una tabla de resultados detallados: Una vez iniciada la sesión de Gretl y abierto el fichero *datos3-1.gdt*, vamos a

Modelo →Mínimos cuadrados ordinarios...

Figura 3.3: Ventana de especificación del modelo lineal

Aparece la ventana donde se especifica la parte sistemática del modelo:

- Escogemos la variable dependiente, el precio de venta: en el cuadro izquierdo pinchamos sobre P y luego *Elegir* $->$.
- Elegimos la variable independiente, el tamaño: en el cuadro izquierdo pinchamos sobre $SQFT$ y luego *Añadir* $->$. La ventana de especificación aparece en la Figura 3.3.

Tras pinchar en *Aceptar* aparece la ventana de resultados del modelo (ver la Figura 3.4). En esta

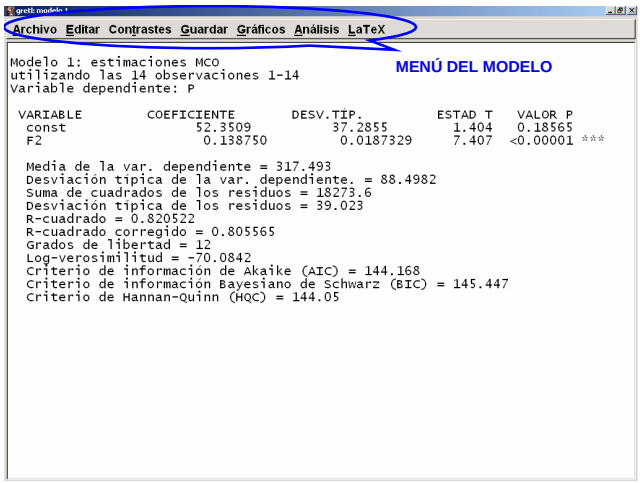


Figura 3.4: Ventana de resultados de estimación MCO

ventana aparecen **todos** los resultados básicos para el análisis del modelo y que se irán explicando a lo largo del curso.

Modelo 1: MCO, usando las observaciones 1–14
Variable dependiente: price

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	52.3509	37.2855	1.4041	0.1857
sqft	0.138750	0.0187329	7.4068	0.0000

Media de la vble. dep.	317.4929	D.T. de la vble. dep.	88.49816
Suma de cuad. residuos	18273.57	D.T. de la regresión	39.02304
R^2	0.820522	R^2 corregido	0.805565
$F(1, 12)$	54.86051	Valor p (de F)	8.20e–06
Log-verosimilitud	−70.08421	Criterio de Akaike	144.1684
Criterio de Schwarz	145.4465	Hannan–Quinn	144.0501

La primera columna muestra las variables explicativas que se han incluido en el modelo, la constante (*const*) y la superficie que posee la vivienda ($SQFT$). En la segunda columna tenemos los coeficientes estimados por MCO correspondientes a cada una de las variables. Como ya vimos, la **estimación** de la ordenada es igual a $\widehat{\beta}_1$= 52,35 miles de dólares y la estimación de la pendiente es $\widehat{\beta}_2 = 0,138750$ miles $ por pie cuadrado. Así la **Función de Regresión Muestral** es:

$$\widehat{P}_i = 52,3509 + 0{,}138750\ SQFT_i \tag{3.8}$$

Es decir, cuando la superficie de la vivienda aumenta en un pie cuadrado, el precio medio de venta **estimado** aumenta en $\widehat{\beta}_2 \times 1000 = 138,750$ dólares. Observar que esta interpretación corresponde a la estimación del coeficiente, no al parámetro poblacional β_2.

La **desviación típica de los residuos** es el error típico $\hat{\sigma}$ y **Suma de cuadrados de los residuos** es $SCR = \sum_i \hat{u}_i^2$.

También encontramos el valor del **coeficiente de determinación**, $R^2 = 0,820522$ Además recordar que en el MRLS $R^2 = r_{XY}^2$ luego $r_{XY} = \sqrt{0,820522} = \pm\, 0,9058$. Si buscamos la matriz de correlación obtenemos: corr(price, sqft) = 0.90582662. Luego ambas variables están correladas, con correlación positiva y elevada. El resto de resultados se irán interpretando según avancemos en la asignatura.

Guardar resultados. Si en el menú de resultados del modelo vamos a *Archivo →Guardar a sesión como icono*, el modelo queda guardado dentro de la carpeta `USER`. Así, podemos recuperarlo siempre que queramos; basta con pinchar sobre el botón *iconos de sesión*, cuarto por la izquierda de la barra de herramientas, y en la ventana que aparece, pinchar dos veces sobre el icono llamado *Modelo 1*. Si posteriormente estimáramos otro modelo y lo guardáramos como icono, Gretl lo denominaría *Modelo 2*.

Algunos gráficos de interés. La opción *Gráficos* de la ventana de resultados del modelo incluye distintas representaciones gráficas tanto de la variable endógena de interés, como de su ajuste y de los errores de su ajuste. Veamos algunos de los más utilizados en regresión con datos de sección cruzada.

- En *Gráficos → Gráfico de variable estimada y observada →contra SQFT* obtenemos el gráfico

de dispersión de las observaciones reales P_i frente a la variable explicativa $SQFT_i$ junto con la función de regresión muestral (3.8). El resultado es la figura izquierda de la Figura 3.5.

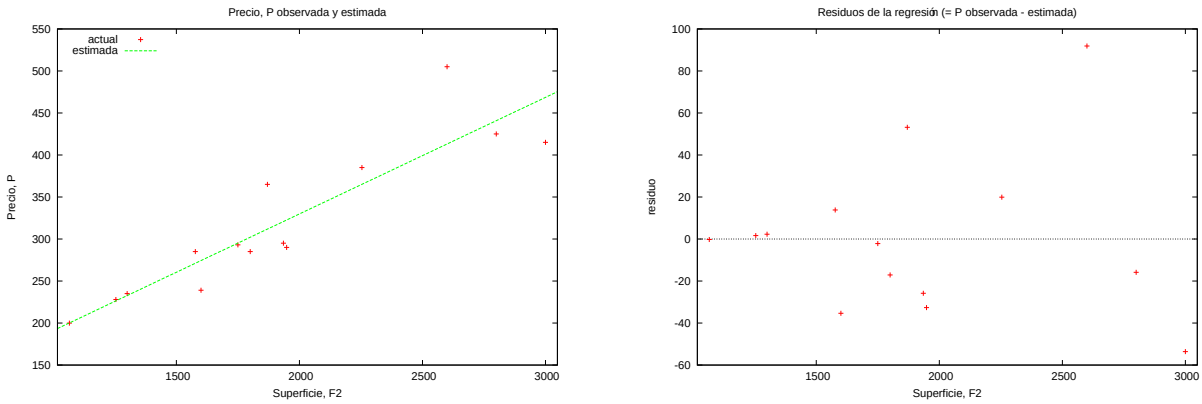


Figura 3.5: Gráficos de resultados de regresión MCO

- Si seleccionamos *Gráficos* →*Gráfico de residuos* →*contra SQFT*, se representan los errores de ajuste $\hat{u}_i$ sobre la variable explicativa $SQFT_i$, es decir, el diagrama de dispersión de los pares de puntos $(SQFT_1, \hat{u}_1), \ldots, (SQFT_{14}, \hat{u}_{14})$, como aparece en la figura derecha de la Figura 3.5. Podemos apreciar que los residuos se distribuyen alrededor del valor cero ($\bar{\hat{u}} = 0$) y que la variación con respecto a esta media crece a medida que aumenta el tamaño de los pisos. Este último resultado podría indicar que la hipótesis básica de varianza constante quizás no sea aceptable.

Variables asociadas a la regresión. Para ver los valores que toman los ajustes $\hat{Y}_i$ y los residuos $\hat{u}_i$, debemos seleccionar *Análisis* →*Mostrar variable observada, estimada, residuos.*
El resultado que obtenemos es la tabla 3.1. Podemos guardar cualquiera de estos valores seleccionando la opción *Guardar* del menú del modelo, tal como muestra la Figura 3.6.

```
Rango de estimación del modelo: 1--14
Desviación típica de los residuos = 39,023
```

Observaciones	P	estimada	residuos	Observaciones	P	estimada	residuos
1	199,9	200,1	−0,2	8	365,0	311,8	53,2
2	228,0	226,3	1,7	9	295,0	320,8	−25,8
3	235,0	232,7	2,3	10	290,0	322,6	−32,6
4	285,0	271,2	13,8	11	385,0	365,1	19,9
5	239,0	274,4	−35,5	12	505,0	413,1	91,9
6	293,0	295,2	−2,2	13	425,0	440,9	−15,9
7	285,0	302,1	−17,1	14	415,0	468,6	−53,6

Tabla 3.1: Residuos de la regresión MCO.

Para almacenar $\hat{P}_i$ hay que elegir *Guardar* →*Valores estimados.* Sale una ventana en la que, por defecto, el valor ajustado o estimado de la variable endógena se llama *yhat1* y en la descripción aparece *valores estimados mediante el modelo 1.* Dado que nuestra variable dependiente es el precio de venta P, cambiamos de nombre a la variable y la renombramos como *phat1.* Si repetimos los

pasos anteriores pero escogemos *Guardar* →*Residuos*, en la ventana correspondiente se nombra a los residuos como *uhat1* y la descripción es *residuos del modelo 1*. Una vez guardadas estas dos series, las encontramos en la ventana principal junto a la variable independiente P y la variable explicativa $SQFT$.

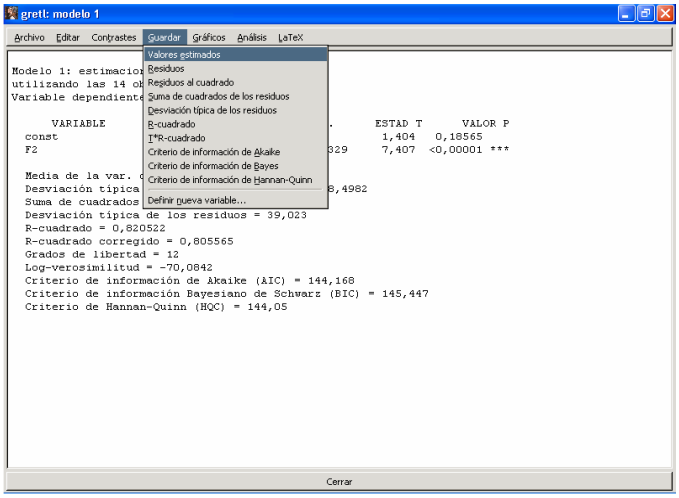


Figura 3.6: Residuos MCO

3.5. Bibliografía del tema

Referencias bibliográficas básicas:

• Teórica:

[1] Stock, James H. y Mark Watson (2012). Introducción a la Econometría. Pearson.

[2] Wooldridge, J.M. (2006). Introducción a la Econometría. Ed. Thomson Learning, 2ª edición.

• 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[4] Gujarati, D. y Porter, D.C. (2010). Econometría. Editorial McGraw-Hill, Madrid. 5ª edición.

[5] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

Tema 4

Modelo de Regresión Lineal Simple. Inferencia

Una vez estimado el Modelo de Regresión Lineal Simple dedicaremos este tema a hacer inferencia sobre el mismo. Aprenderemos a realizar contrates sobre posibles valores de los parámetros poblacionales comenzando con el contraste de significatividad de la variable independiente. Previamente hemos de derivar la distribución del estimador MCO. Para finalizar el tema veremos como realizar inferencia 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. Conocer y saber demostrar las propiedades del estimador de MCO (C2 y C3).
2. Saber derivar la distribución del estimador de MCO (C2).
3. Saber derivar intervalos de confianza y utilizarlos para el contraste de hipótesis (C2 y C3)
4. Saber contrastar la significatividad individual de la variable explicativa (C2 y C3).
5. Utilizar un software econométrico (Gretl) para realizar contraste de hípótesis 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:

- Stock and Watson, J. M. (2012). Cap. 5.
- Wooldridge, J.M. (2006). Caps. 2

4.1. Propiedades del estimador de 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 θ, 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.

4.1.1. Propiedades del estimador de MCO

Sea el modelo de regresión lineal general

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

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.** El estimador MCO, condicionando en X, se puede expresar como una función lineal de Y o de u que serían los elementos aleatorios.

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

- **Insesgadez.** Dado que $E(u|X) = 0$ el estimador MCO es insesgado es decir, su valor esperado es igual al vector de coeficientes del modelo.

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

- **Matriz de varianzas y covarianzas.** Dado que $E(u|X) = 0$ y $E(uu'|X) = \sigma^2 I_N$)

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

Matricialmente para el MRLS:

$$\underset{(2\times 2)}{V(\hat{\beta})} = \begin{bmatrix} V(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & V(\hat{\beta}_2) \end{bmatrix} = \sigma^2 \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \sigma^2(X'X)^{-1}$$

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

La matriz de varianzas y covarianzas $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ es **mínima** y nos lo garantiza el Teorema de Gauss-Markov.

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

Notar que para derivar la matriz de varianzas y covarianzas del estimados MCO hemos utilizado todas las hipótesis básicas sobre la perturbación salvo la hipótesis de normalidad.

4.1.2. 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** de σ^2:

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

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 salvo la hipóteis de normalidad, **un estimador insesgado de la matriz de varianzas y covarianzas**, de $\hat{\beta}_{MCO}$ es

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

Ejemplo 4.1

Con los datos disponibles en el fichero *data3-1.gdt* y los resultados de la estimación del modelo (2.2),

$$P_i = \beta_1 + \beta_2\ SQFT_i + u_i \qquad i = 1, \ldots, N$$

se calcula la siguiente matriz de varianzas y covarianzas estimada:

$$\hat{\sigma}^2 = \frac{SCR}{N-K} = \frac{Y'Y - \hat{\beta}'X'Y}{N-K} = \frac{\sum P_i^2 - \hat{\beta}X'Y}{N-K} = \frac{18273{,}5678}{12} = 1522,79$$

$$\begin{aligned} \widehat{V}(\hat{\beta}_{MCO}) &= 1522,79 \times \begin{bmatrix} 14 & 26753 \\ 26753 & 55462515 \end{bmatrix}^{-1} = \\ &= \begin{bmatrix} 1390{,}21 & -0{,}670583 \\ & 3{,}50920e-04 \end{bmatrix} \end{aligned}$$

4.2. Distribución del estimador de MCO bajo Normalidad

Si $Y = X\beta + u$, donde $u|X \sim N(0, \sigma^2 I_N)$, el estimador MCO, dado que es lineal en las perturbaciones, también seguirá una distribución Normal Multivariante, con vector de medias $E(\hat{\beta}|X) = \beta$ y matriz de varianzas y covarianzas $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$. Es decir,

$$\hat{\beta}_{MCO}|X \sim N(\beta, \sigma^2(X'X)^{-1})$$

Para el k-ésimo coeficiente,

$$\hat{\beta}_k|X \sim N(\beta_k, \sigma^2 a_{kk})$$

donde a_{kk} es el elemento (k,k) de la matriz $(X'X)^{-1}$.

Luego para los coeficientes β_2 y β_1

$$\hat{\beta}_2|X \sim N(\beta_2, \sigma^2 a_{22})$$

donde a_{22} es el elemento $(2,2)$ de la matriz $(X'X)^{-1}$.

$$\hat{\beta}_1|X \sim N(\beta_1, \sigma^2 a_{11})$$

donde a_{11} es el elemento $(1,1)$ de la matriz $(X'X)^{-1}$.

4.3. Estimación por intervalo

Para el k-ésimo coeficiente,

$$\hat{\beta}_k|X \sim N(\beta_k, \sigma^2 a_{kk})$$

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 $\hat{\sigma}\sqrt{a_{kk}} = \hat{\sigma}_{\hat{\beta}_k} = \widehat{desv}(\hat{\beta}_k)$.

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 $\beta_k \quad k = 1, 2$ 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 β_k ó intervalo de confianza $(1-\alpha)$ para β_k. Un intervalo de confianza nos dice que, con probabilidad $(1-\alpha)$ se estima que el parámetro β_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 β_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 β_k. Además, como veremos más adelante, los intervalos sirven para realizar contraste de hipótesis.

4.4. Contraste de hipótesis. Estadístico t

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.

Un contraste de hipótesis tiene tres etapas: formulación de dos hipótesis opuestas; derivación de un estadístico de contraste y su distribución muestral y determinación de un criterio de decisión para elegir una de las dos hipótesis planteadas.

Una **hipótesis** estadística es una afirmación sobre la distribución de una o varias variables aleatorias. En un contraste se trata de decidir cuál, entre dos hipótesis planteadas, es la que mejor se adecúa a los datos[1]. La hipótesis de interés se denomina **hipótesis nula**, H_0, y la supondremos cierta mientras no haya evidencia en contra. La hipótesis frente a la que se contrasta la nula se llama **hipótesis alternativa**, H_1.

Tanto las hipótesis nulas como alternativas pueden ser simples o compuestas. Las hipótesis simples especifican un único valor para el parámetro poblacional y por tanto en ellas la distribución de probabilidad queda perfectamente definida. En general especificaremos hipótesis nulas simples. En la hipótesis compuesta se especifica un rango de valores para el parámetro poblacional. La hipótesis alternativa puede ser a una cola o a dos colas. La hipótesis alternativa a una cola envuelve todos los posibles valores del parámetro poblacional a un lado o a otro del valor especificado en la H_0. La hipótesis alternativa a dos colas envuelve todos los valores posibles del parámetro poblacional excepto el especificado por la H_0.

La elección entre las hipótesis se basa en un **estadístico de contraste**, que es una función de los datos que mide la discrepancia entre estos y H_0. A continuación veremos en detalle el mecanismo de contraste. 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.

[1] El establecimiento de una hipótesis sobre el parámetro desconocido θ divide su espacio paramétrico en dos partes, una integrada por los valores que cumplan la hipótesis, le llamaremos Θ_0 y otra formada por el conjunto de valores que no la cumplen y que llamaremos Θ_1. Θ_0 y Θ_1 son disjuntos por definición, $\Theta_0 \cup \Theta_1 = \Theta$.

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. 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:

$$\hat{\beta}_k | X \sim N(\beta_k, \sigma^2 a_{kk})$$

Si σ^2 es conocida todo es conocido en la distribución de β_k y el estadístico de contraste sería:

$$\frac{\hat{\beta}_k - c}{\sigma_{\hat{\beta}_k}} \overset{H_0}{\sim} N(0,1)$$

En el resto de ejemplos consideramos el caso más habitual σ^2 desconocida, para el cual podemos derivar el siguiente estadístico de contraste[2] y distribución asociada cuando σ^2 es estimada con el estimador insesgado $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$:

$$\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(N-K)}$$

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 $\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} > t_{(N-K)|\,\alpha}$.

4.4.1. Contraste de significatividad individual en el MRLS

Cuando $c = 0$ al contraste se le denomina de significatividad individual. En este caso:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

[2] Si σ^2 es desconocida habría de ser estimada, bajo la normalidad de las perturbaciones

$$u_i | X \sim N(0, \sigma^2) \longrightarrow \frac{(N-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$$

y derivar el correspondiente estadístico de contraste, que sería:

$$\frac{\frac{\hat{\beta}_k - c}{\sigma\sqrt{a_{kk}}}}{\sqrt{\frac{\sum \hat{u}_i^2/\sigma^2}{N-K}}} \overset{H_0}{\sim} t_{(N-K)} \qquad \text{si simplificamos} \qquad \frac{\hat{\beta}_k - c}{\hat{\sigma}\sqrt{a_{kk}}} \overset{H_0}{\sim} t_{(N-K)}$$

Podemos derivar el siguiente estadístico de contraste y distribución:

$$\frac{\hat{\beta}_2}{\widehat{des}(\hat{\beta}_2)} \overset{H_0}{\sim} t_{(N-K)}$$

Si el estadístico calculado para la muestra es mayor que el estadístico en tablas, $\frac{\hat{\beta}_2}{\widehat{des}(\hat{\beta}_2)} > t_{(N-K)|\frac{\alpha}{2}}$ para un α dado, se rechaza la hipótesis nula. En este caso $\beta_2 \neq 0$ y la variable explicativa asociada X es significativa para explicar el comportamiento de la variable endógena. Por tanto este contraste sirve para decidir si la variable X debe mantenerse en el modelo y es en realidad un contraste de especificación. Si el estadístico calculado para la muestra es menor que el estadístico en tablas, $\frac{\hat{\beta}_2}{\widehat{des}(\hat{\beta}_2)} < t_{(N-K)|\frac{\alpha}{2}}$ para un α dado, no se rechaza la hipótesis nula. En este caso $\beta_2 = 0$ y la variable explicativa asociada X no es significativa para explicar el comportamiento de la variable endógena.

→ Continuamos con el ejemplo de la relación entre precio y superficie de vivienda. Veamos si la superficie de la vivienda es un factor relevante para determinar su precio:

$$\begin{cases} H_0 \colon \beta_2 = 0 \\ H_a \colon \beta_2 \neq 0 \end{cases} \qquad t = \frac{\widehat{\beta}_2}{\widehat{des}(\widehat{\beta}_2)} \overset{H_0}{\sim} t_{(14-2)}$$

El valor muestral del estadístico t_c es:

$$t_c = \frac{\hat{\beta}_2}{\widehat{des}(\hat{\beta}_2)} = \frac{0,13875}{0,0187329} = 7,4068$$

El valor crítico del contraste para el nivel de significación del 5 % es $t_{(14-2)0,05/2} = 2,179$. Como resultado tenemos que $7,4068 > 2,179$, por lo que t_c pertenece a la región crítica y, en consecuencia, rechazamos H_0 a un nivel de significación del 5 %. Podemos concluir que la variable $SQFT$ es significativa o relevante para determinar el precio medio de la vivienda.

4.4.2. Otros contrastes sobre β_2.

Como hay evidencia estadística de que β_2 es distinto de cero y, por lo tanto, la variable explicativa X es significativa, nos puede interesar saber qué valor puede tomar. Vamos a generalizar el procedimiento de contraste anterior. Veamos dos ejemplos.

→ **Ejemplo 1**. Ante un aumento de la superficie de la vivienda de un pie cuadrado, ¿podría el precio medio de venta de la vivienda aumentar en 100 dólares? Planteamos el contraste:

$$\begin{cases} H_0 \colon \beta_2 = 0,1 \\ H_a \colon \beta_2 \neq 0,1 \end{cases}$$

El estadístico de contraste y distribución asociada es:

$$t = \frac{\widehat{\beta}_2 - 0,1}{\widehat{des}(\widehat{\beta}_2)} \overset{H_0}{\sim} t_{(N-K)}$$

El valor muestral del estadístico calculado es:

$$t_c = \frac{0,138750 - 0,1}{0,0187329} = 2,068$$

El valor crítico es $t_{(14-2)0,05/2} = 2,179$. Como el valor calculado cae fuera de la región crítica, $2,068 < 2,179$, no rechazamos la H_0 a un nivel de significación del 5 %. Por tanto, es posible un incremento de 100 dólares en el precio medio de la vivienda ante un aumento unitario en la superficie.

→ **Ejemplo 2**. Ante el mismo aumento unitario en la superficie, ¿podría el precio medio de venta de la vivienda aumentar en 150 dólares? Planteamos el contraste y, al igual que en el caso anterior, llegamos al estadístico de contraste:

$$\begin{cases} H_0 : \beta_2 = 0,15 \\ H_a : \beta_2 \neq 0,15 \end{cases} \qquad t = \frac{\widehat{\beta_2} - 0,15}{\widehat{des}(\widehat{\beta_2})} \overset{H_0}{\sim} t_{(N-K)}$$

El estadístico de contraste en este caso toma el valor

$$t_c = \frac{0,138750 - 0,15}{0,0187329} = -0,6005 \Rightarrow |-0,6005| < 2,179$$

con $2,179 = t_{(12)0,025}$. Así, no rechazamos H_0 a un nivel de significación del 5 % y también es posible que si $\Delta SQFT = 1$, entonces el precio medio de la vivienda aumente en 150$.
Notar que en este caso el valor muestral del estadístico es negativo por lo que se toma en valor absoluto para seguir utilizando la cola derecha de la distribución t-student al tomar la regla de decisión.

4.4.3. 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, el intervalo de confianza asociado a β_2:

$$Pr\left[\hat{\beta}_2 k - t_{\frac{\alpha}{2}(N-K)}\hat{\sigma}_{\hat{\beta}_2} < \beta_2 < \hat{\beta}_2 + t_{\frac{\alpha}{2}(N-K)}\hat{\sigma}_{\hat{\beta}_2}\right] = 1 - \alpha$$

$$IC(\beta_2)_{1-\alpha} : \left(\hat{\beta}_2 \pm t_{\frac{\alpha}{2}(N-K)}\,\hat{\sigma}_{\hat{\beta}_2}\right)$$

y la regla de decisión es que si la constante c pertenece al intervalo, no rechazamos H_0 con un nivel de significación α y si no pertenece al intervalo, rechazamos H_0 con un nivel de significación α. Claramente se obtienen exactamente los mismos resultados utilizando los estadísticos de contraste individuales que utilizando los intervalos de confianza.

4.5. Inferencia en gretl

Para mostrar cómo hacer inferencia en gretl seguimos utilizando el ejemplo: $P_i = \beta_1 + \beta_2 SQFT_i + u_i$, con la muestra del fichero *datos3-1.gdt*. Los resultados de la estimación que muestra *gretl* son:

Modelo 1: MCO, usando las observaciones 1–14
Variable dependiente: price

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	52.3509	37.2855	1.4041	0.1857
sqft	0.138750	0.0187329	7.4068	0.0000

Media de la vble. dep.	317.4929	D.T. de la vble. dep.	88.49816
Suma de cuad. residuos	18273.57	D.T. de la regresión	39.02304
R^2	0.820522	R^2 corregido	0.805565
$F(1, 12)$	54.86051	Valor p (de F)	8.20e–06
Log-verosimilitud	−70.08421	Criterio de Akaike	144.1684
Criterio de Schwarz	145.4465	Hannan–Quinn	144.0501

→ **Contraste de significatividad individual**:

$$\begin{cases} H_0\colon \beta_2 = 0 \\ H_a\colon \beta_2 \neq 0 \end{cases} \qquad t = \frac{\widehat{\beta_2}}{\widehat{des}(\widehat{\beta_2})} \overset{H_0}{\sim} t_{(14-2)}$$

El valor muestral del estadístico t_c se incluye en los resultados de estimación, es la cuarta columna, encabezada por *Estadístico t*. Es decir,

$$t_c = 7,4068 = \frac{columna\ COEFICIENTE}{columna\ DESV.TIP.} = \frac{\hat{\beta}_2}{\widehat{desv}(\hat{\beta}_2)} = \frac{0,13875}{0,0187329}$$

El valor crítico del contraste para el nivel de significación del 5 % es $t_{(14-2)0,05/2} = 2,179$. Como resultado tenemos que $7,4068 > 2,179$, por lo que t_c pertenece a la región crítica y, en consecuencia, rechazamos H_0 a un nivel de significación del 5 %. Podemos concluir que la variable $SQFT$ es significativa o relevante para determinar el precio medio de la vivienda. En el tema siguiente, veremos cómo la columna *valor p* de la tabla de resultados de Gretl informa sobre la conclusión del contraste.

Hay que tener en cuenta que la columna *Estadístico t* de los resultados de estimación de Gretl, corresponde al valor muestral del estadístico para $H_0\colon \beta_2 = 0$ exclusivamente. Por tanto, para cualquier otra nula tenemos que calcular el valor muestral del estadístico de contraste o utilizar el intervalo de confianza para realizar el contraste.

Utilización del intervalo de confianza para hacer inferencia Vamos a obtener los intervalos de confianza para los dos coeficientes de regresión. Para ello, vamos a *Análisis →Intervalos de confianza para los coeficientes*. El resultado mostrado es:

Variable	Coeficiente	Intervalo de confianza 95 %	
const	52.3509	−28.8872	133.589
sqft	0.138750	0.0979349	0.179566

En esta tabla de resultados, la segunda columna ofrece las estimaciones por punto, esto es, $\widehat{\beta_1} = 52,3509$ y $\widehat{\beta_2} = 0,138750$. La tercera indica los límites de los intervalos a una confianza del 95 %, esto es:

$$IC(\beta_1)_{0,95} = [-28,887\ ;\ 133,587]$$

$$IC(\beta_2)_{0,95} = [0,0979349\ ;\ 0,179566]$$

Por tanto, podemos afirmar con un nivel de confianza del 95 % que, ante un aumento de la superficie de la vivienda de un pie cuadrado, el precio medio de venta de dicha vivienda aumentará entre 97,9349 y 179,566 dólares.

Para hacer inferencia utilizando el intervalo de confianza solo tenemos que ver si el valor del parámetro en la hipótesis nula cae dentro o fuera del intervalo de confianza. Si cae dentro no rechazamos la hipótesis nula y si cae fuera rechazamos, para un nivel de significatividad $\alpha = 5\,\%$. Por ejemplo para el último caso:

$$0,15\ \in\ [0,0979349\ ;\ 0,179566]$$

no rechazamos H_0 a un nivel de significación del 5 %.

En la ventana de resultados de la estimación no aparece la varianza de la perturbación estimada, pero se puede calcular:

- De su relación con la desviación típica de los residuos; $\hat{\sigma}^2 = 39,0230^2 = 1522,8$.
- Dividiendo la SCR entre los grados de libertad $N - 2$.

$$\hat{\sigma}^2 = \frac{SCR}{N-2} = \frac{18273,6}{14-2} = 1522,8$$

También es posible obtener la estimación de la **matriz de varianzas y covarianzas** de los coeficientes de regresión seleccionando en el menú del modelo *Análisis →Matriz de covarianzas de los coeficientes*. El resultado para el conjunto de 14 observaciones es:

```
Matriz de covarianzas de los coeficientes de regresión
        const          sqft
      1390,21     -0,670583     const
                   3,50920e-04  sqft
```

Tabla 4.1: Estimación de varianzas y covarianza de $\hat{\beta}_1$ y $\hat{\beta}_2$.

es decir, $\widehat{var}(\hat{\beta}_1) = 1390,21$, $\widehat{var}(\hat{\beta}_2) = 3,5092 \times 10^{-4}$ y $\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0,670583$.

Los errores típicos de estimación y de la regresión dependen de las unidades de medida, es decir, las podemos reducir o agrandar cuanto queramos con sólo cambiar de escala las variables dependiente e independiente.

4.6. Resumen. Presentación de los resultados

Los resultados de la estimación de un modelo se suelen presentar de forma resumida, incluyendo tanto la recta de regresión como un conjunto de estadísticos útiles para evaluar los resultados. Una forma habitual de presentar la estimación es la siguiente:

$$\underset{(\widehat{des})}{\widehat{P}} = \underset{(37,285)}{52,3509} + \underset{(0,018733)}{0,138750}\, F2$$

$$N = 14 \qquad R^2 = 0,82 \qquad \hat{\sigma} = 39,023$$

Bajo cada coeficiente estimado aparece su error típico de estimación. Otra opción es incluir los estadísticos t_c de significatividad individual o los grados de libertad. Por ejemplo,

$$\underset{(estad.\, t)}{\widehat{P}} = \underset{(1,404)}{52,3509} + \underset{(7,407)}{0,138750}\, F2$$

$$\text{Grados libertad} = 12 \qquad R^2 = 0,82 \qquad \hat{\sigma} = 39,023$$

4.7. Bibliografía del tema

Referencias bibliográficas básicas:

• Teórica:

[1] Stock, James H. y Mark Watson (2012). Introducción a la Econometría. Pearson.

[2] Wooldridge, J.M. (2006). Introducción a la Econometría. Ed. Thomson Learning, 2ª edición.

• 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[4] Gujarati, D. y Porter, D.C. (2010). Econometría. Editorial McGraw-Hill, Madrid. 5ª edición.

[5] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

Tema 5

Modelo de Regresión Lineal General

En este tema nos ocuparemos de generalizar el Modelo de Regresión Lineal Simple para analizar las relaciones entre un conjunto de variables. Nuestro objetivo fundamental será explicar el comportamiento de una variable, que llamamos variable a explicar, mediante un conjunto de variables económicas, que llamamos explicativas. Especificaremos el Modelo de Regresión Lineal General, poniendo especial cuidado en el tratamiento de las variables explicativas cualitativas.

A continuación estimaremos el modelo por Mínimos Cuadrados Ordinarios, MCO, que bajo ciertas hipótesis de comportamiento sobre los distintos elementos del modelo nos proporciona estimadores con buenas propiedades, lineales, insesgados y de mínima varianza. Una vez estimado el modelo veremos como realizar contraste de restricciones lineales que recojan hipótesis relevantes desde el punto de vista económico dentro del Modelo de Regresión Lineal General. Aprenderemos a contrastar no sólo si las variables son relevantes individualmente sino si también lo son conjuntamente para explicar el comportamiento de la variable objetivo y a hacer contraste de combinaciones lineales, entre otros contrastes de interés.

Finalmente 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:

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.

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. Explicar y entender el alcance de las hipótesis básicas sobre el comportamiento del modelo de regresión lineal general (C1).
2. Interpretar los coeficientes del modelo de regresión, incluyendo los de especificaciones no lineales en las variables (C1).
3. Saber especificar correctamente modelos que incluyan variables cualitativas (C1).
4. Aplicar el estimador de Mínimos Cuadrados Ordinarios, MCO (C2).
5. Interpretar los coeficientes estimados del modelo de regresión (C2).
6. Distinguir entre la perturbación y el residuo u error de estimación. Conocer las distribuciones respectivas (C2).
7. Conocer y saber demostrar las propiedades del estimador de MCO. Derivar la distribución del estimador de MCO (C2 y C3).
8. Saber contrastar la significatividad individual de las variables explicativas (C2 y C3).
9. Saber contrastar la significatividad conjunta de las variables explicativas (C2 y C3).
10. Saber contrastar restricciones lineales de parámetros (C2 y C3).
11. Saber contrastar restricciones múltiples (C2 y C3).
12. 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).
13. Organizar y sistematizar información estadística relevante (C4).
14. Utilizar un software econométrico (Gretl) para realizar contraste de hípótesis 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:

- Stock and Watson, J. M. (2012). Cap. 6, 7 y 8.
- Wooldridge, J.M. (2006). Caps. 2, 3, 4, 6 y 7.

5.1. 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 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.
- El salario es una función del nivel de estudios, la experiencia, la edad y el puesto 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, $X_1, X_2, \ldots, X_K$ mediante la especificación de un modelo lineal.

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

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

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

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

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

donde habitualmente $X_{1i} = 1 \; \forall i$, de forma que β_1 es un término independiente y entonces,

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

Elementos del MRLG

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

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + u_i \qquad i = 1, 2, \ldots, N \tag{5.1}$$

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

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

o bien en forma matricial como

$$\underset{(N \times 1)}{Y} = \underset{(N \times K)}{X} \underset{(K \times 1)}{\beta} + \underset{(N \times 1)}{u}$$

donde

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

5.1.1. Hipótesis básicas.

1. Hipótesis sobre la perturbación aleatoria

   - La media de la perturbación condicionada en X es cero, para todo i, $E(u_i|X_i) = 0 \quad \forall i$. Para la perturbación en i lo escribimos como $E(u_i|X_i) = 0 \quad \forall i$, cuando miramos al modelo en forma matricial escribimos esta hipótesis como $E(u|X) = \vec{0}$.
   - $V(u_i) = E(u_i^2|X_i) = \sigma_u^2 = \sigma^2 \quad \forall i$ es decir la varianza de la perturbación condicionada en X es desconocida e igual a σ^2 para todas las observaciones. Estamos suponiendo igual dispersión o variabilidad. A esta hipótesis se le conoce con el nombre de *Homocedasticidad*.

     Hay que notar que generalmente σ^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 σ^2.
   - $Cov(u_i, u_j) = E(u_i u_j|X) = 0 \quad \forall i, j \quad i \neq j$. La covarianza entre perturbaciones de distintas observaciones es cero. 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.

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

$$E(uu'|X) = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I_N$$

     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*.
   - La distribución de las perturbaciones condicionada en X es normal:

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

     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 σ^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 $u_i|X_i \sim N(0, \sigma^2)$.

Estas propiedades pueden también escribirse conjuntamente como

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

ó en forma matricial,

$$\underset{(N \times 1)}{u|X} \sim N\left(\underset{(N \times 1)}{0_N}, \underset{(N \times N)}{\sigma_u^2 I_N}\right)$$

2. Hipótesis sobre las variables exógenas X.

   - Condicionamos el análisis a unos valores dados de X. Este proceder es similar a considerar las variables como no aleatorias o regresores fijos.
   - La matriz X es de rango completo e igual a K con $K < N$, $\quad rg(X) = K$, 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 $X_1, X_2, \ldots, X_K$ 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.

5.2. Función de Regresión Poblacional. Interpretación de los coeficientes.

Dados los supuestos básicos del MRLG,

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

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

- $\beta_1 = E(Y_i|X_{2i} = \ldots = X_{Ki} = 0)$. Valor medio o esperado de Y_i cuando las variables explicativas son todas cero.

- $\beta_k = \dfrac{\partial E(Y_i)}{\partial X_{ki}} = \dfrac{\Delta E(Y_i)}{\Delta X_{ki}} \quad \forall k = 2, \ldots, K$. 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 β_k unidades en la variable endógena, ceteris paribus.

Ejemplo 5.1

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:

$$PRICE_i = \beta_1 + \beta_2\, SQFT_i + \beta_3\, BEDRMS_i + \beta_4\, BATHS_i + u_i \quad i = 1, 2, \ldots, N \quad (5.2)$$

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 β_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 β_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 β_4 miles de dólares.

Ejemplo 5.2

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 caeran (demanda ineslástica en precio) pero si una reducción en el precio conlleva un gran incremento en la cantidad vendida, los ingresos por ventas creceran (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:

$$S = \beta_1 + \beta_2 P + \beta_3 A \tag{5.3}$$

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.

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:

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + u_i \qquad i = 1, \ldots, 75 \tag{5.4}$$

β_1, β_2 y β_3 son los parámetros desconocidos cuyo valor queremos estimar.

- El coeficiente β_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 β_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 β_3 miles de dólares.

Ejemplo 5.3

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

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

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

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

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 β_2 es el incremento o decremento en el salario esperado para un individuo por el hecho de ser mujer. Por tanto β_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 5.4

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

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

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:

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

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 β_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 β_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 5.5

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

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

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 β_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 β_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 segundo trimestre. Luego β_4 es el diferencial entre las ventas esperadas en el cuarto trimestre y el primer trimestre.

Algunas consideraciones sobre la linealidad en parámetros Hay dos tipos de linealidad, linealidad en variables y linealidad en parámetros. Nosotros estamos interesados en la linealidad en parámetros. Existen relaciones que aunque en principio no son lineales pueden transformarse en lineales y por tanto son perfectamente estimables en nuestros términos. Un ejemplo específico de un modelo no lineal linealizable 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 β_1, β_2 y β_3:

$$Q_t = AL_t^{\beta_2} K_t^{\beta_3} e^{u_t} \longrightarrow LnQ_t = \beta_1 + \beta_2 LnL_t + \beta_3 LnK_t + u_t \tag{5.5}$$

siendo $\beta_1 = LnA$. Una ventaja de este tipo de modelos como el recogido en la ecuación (5.5), 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:

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

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

Es decir $\beta_k \quad k = 2, 3$, 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 β_2 y β_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.

5.2.1. 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.

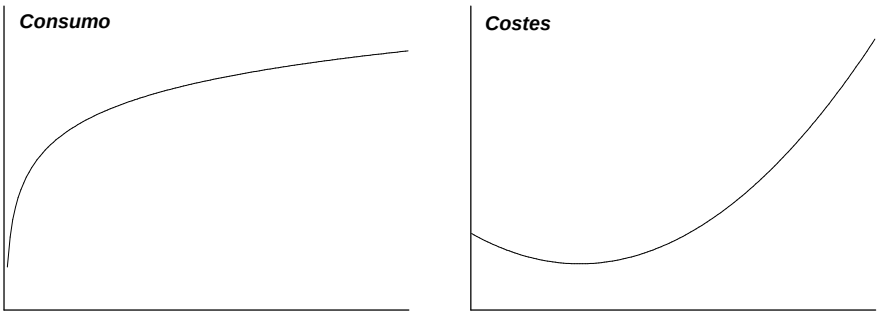


Figura 5.1: Relaciones económicas no lineales

Es necesario hacer enfásis 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:

$$\ln C = \alpha + \beta \ln R + u \tag{5.6}$$

$$C = \alpha + \beta \ln R + u \tag{5.7}$$

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

$$CM = \beta_1 + \beta_2 Q + \beta_3 Q^2 + u \tag{5.8}$$

Los modelos (5.6), (5.7) y (5.8) 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 (5.8), 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 β_2 y β_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.

5.3. 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[1].

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

[1]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.

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

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

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

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.

En este tema ya hemos trabajado con ellas, el Ejemplo 5.3 especificamos la función de salario en función del regresor cualitativo sexo e interpretamos sus parámetros. En el Ejemplo 5.4 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.

5.3.1. 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:

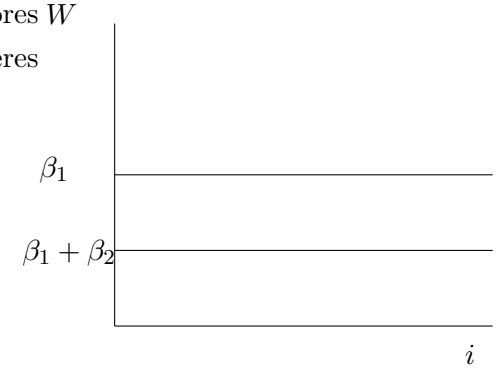
$$S_{1i} = \begin{cases} 1 & \text{si el individuo i es hombre} \\ 0 & \text{en caso contrario} \end{cases} \qquad S_{2i} = \begin{cases} 1 & \text{si el individuo i es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

y podemos especificar el siguiente modelo como ya se hizo en el Ejemplo 2.5:

$$W_i = \beta_1 + \beta_2 S_{2i} + u_i \qquad i = 1, \ldots, N_H + N_M \qquad u_i \sim NID(0, \sigma^2) \tag{5.9}$$

Hay que notar que el modelo (5.9) da lugar a dos ecuaciones:

$$W_i = \beta_1 + u_i \quad i = 1, \ldots, N_H \qquad \text{para los hombres } W$$
$$W_i = \beta_1 + \beta_2 + u_i \quad i = 1, \ldots, N_M \qquad \text{para las mujeres}$$



β_1 es el salario esperado cuando el individuo es hombre, $\beta_1+\beta_2$ es el salario esperado de una mujer y β_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 β_2 es negativo por razones didácticas.

• Alternativa de especificación del modelo (5.9):

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + u_i \qquad i = 1, \ldots, N_H + N_M \tag{5.10}$$

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 (5.10) da lugar a dos ecuaciones:

$$W_i = \alpha_1 + u_i \quad i = 1, \ldots, N_H \qquad \text{para los hombres}$$
$$W_i = \alpha_2 + u_i \quad i = 1, \ldots, N_M \qquad \text{para las mujeres}$$

La relación entre los parámetros del modelo (5.9) y los del modelo (5.10) 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 5.2

Interpreta los coeficientes de la siguiente regresión:

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

donde W_i es el salario del individuo i y

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

¿Qué diferencia hay entre ésta especificación y la especificación del modelo (5.9)?

5.3.2. 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:

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

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

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

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

$$W_i = \mu + \alpha_2 S_{2i} + \beta_2 E_{2i} + \beta_3 E_{3i} + u_i \qquad i = 1, \ldots, N_H + N_M \tag{5.11}$$

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

$E(W_i/S_{2i} = E_{2i} = E_{3i} = 0) = \mu$, salario esperado de un hombre con estudios primarios.
$E(W_i/E_{2i} = 1; S_{2i} = E_{3i} = 0) = \mu + \beta_2$, salario esperado de un hombre con estudios secundarios.
$E(W_i/E_{3i} = 1; S_{2i} = E_{2i} = 0) = \mu + \beta_3$, salario esperado de un hombre con estudios universitarios.
$E(W_i/S_{2i} = 1; E_{2i} = E_{3i} = 0) = \mu + \alpha_2$, salario esperado de una mujer con estudios primarios
$E(W_i/S_{2i} = E_{2i} = 1; E_{3i} = 0) = \mu + \alpha_2 + \beta_2$, salario esperado de una mujer con estudios secundarios.
$E(W_i/S_{2i} = E_{3i} = 1; E_{2i} = 0) = \mu + \alpha_2 + \beta_3$, 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}	μ	$\mu + \beta_2$	$\mu + \beta_3$
S_{2i}	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_3$

y podemos interpretar los parámetros como sigue:

- μ Base de comparación.
- α_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.
- β_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.
- β_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:

$$W_i = \mu^* + \alpha_1^* S_{1i} + \alpha_2^* S_{2i} + \beta_1^* E_{1i} + \beta_2^* E_{2i} + \beta_3^* E_{3i} + u_i \qquad i = 1, \ldots, N_H + N_M \tag{5.12}$$

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 (5.12) es:

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

5.3.3. 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:

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + \beta X_i + u_i \qquad i = 1, \ldots, N_H + N_M \tag{5.13}$$

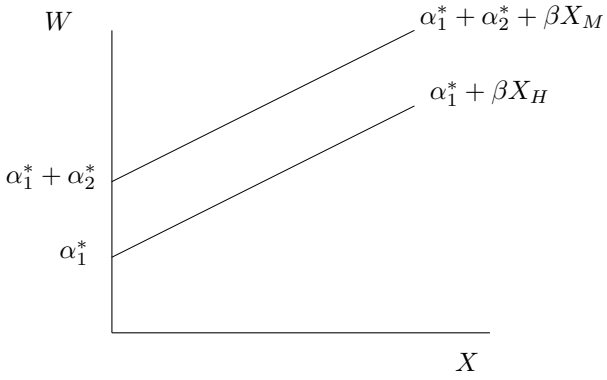
Donde el coeficiente β se interpreta de la forma habitual, $\beta = \frac{\partial E(W_i)}{\partial X_i}$. En forma matricial el modelo sería:

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

La especificación alternativa correspondiente sería:

$$W_i = \alpha_1^* + \alpha_2^* S_{2i} + \beta X_i + u_i \qquad (5.14)$$

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



Donde el coeficiente β se interpreta de la forma habitual. En forma matricial el modelo sería:

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

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

5.3.4. 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:

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

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:

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

5.3.5. Efectos de interacción

Entre factores cualitativos y cuantitativos En las ecuaciones (5.13) y (5.14) 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:

$$W_i = \alpha_1 S_{1i} + \alpha_2 S_{2i} + \beta_1 (S_{1i} \times X_i) + \beta_2 (S_{2i} \times X_i) + u_i \qquad i = 1, \ldots, N_H + N_M \qquad (5.15)$$

$$E(W_i/S_{1i}=1;S_{2i}=0)=\alpha_1+\beta_1 X_i$$
$$E(W_i/S_{1i}=0;S_{2i}=1)=\alpha_2+\beta_2 X_i$$

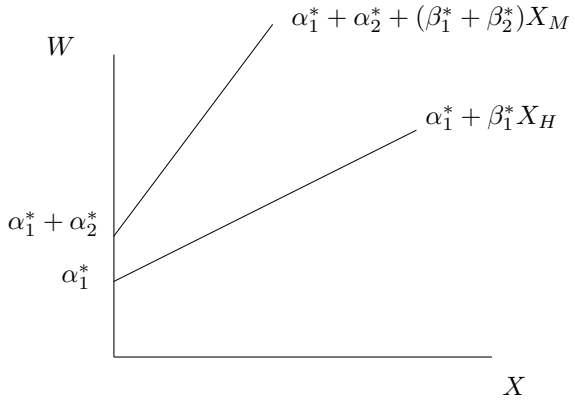
donde β_1 y β_2 recogen el incremento en el salario medio ante un aumento unitario en las horas trabajadas, para los hombres y para las mujeres respectivamente.

Una especificación alternativa sería:

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

$$i=1,\ldots,N_H+N_M \qquad (5.16)$$

siendo α_2^* el incremento salarial en media por el hecho de ser mujer y β_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 (5.11) 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:

$$W_i=\mu+\alpha_2 S_{2i}+\beta_2 E_{2i}+\beta_3 E_{3i}+\gamma_2(S_{2i}\times E_{2i})+\gamma_3(S_{2i}\times E_{3i})+u_i \quad i=1,\ldots,N_H+N_M \,(5.17)$$

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}	μ	$\mu+\beta_2$	$\mu+\beta_3$
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:

μ	base de comparación.
β_2	Efecto diferencial en el salario medio debido a tener un nivel de estudios secundarios, con respecto a tener estudios primarios, para los hombres.
β_3	Efecto diferencial en el salario medio debido a tener un nivel de estudios universitarios, con respecto a tener estudios primarios, para los hombres.
α_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.

5.4. Estimación por Mínimos Cuadrados Ordinarios (MCO)

• Nuestro **objetivo** es estimar los parámetros desconocidos $\beta_k, \quad k = 1, \ldots, K$ de

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

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

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

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

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

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

• Estimador MCO del MRLG

Criterio:

$$\min_{\hat{\beta}_1,\ldots,\hat{\beta}_K} \sum_{i=1}^{N} \hat{u}_i^2 = \min_{\hat{\beta}_1,\ldots,\hat{\beta}_K} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 =$$

$$\min_{\hat{\beta}_1,\ldots,\hat{\beta}_K} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_K X_{Ki})^2 \tag{5.18}$$

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

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

de donde se obtienen las ecuaciones normales:

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

que pueden escribirse como:

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

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

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

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:

$$X'Y = X'X\hat{\beta}_{MCO}. \tag{5.19}$$

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

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y \tag{5.20}$$

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

$$\underset{(K\times K)}{X'X} = \begin{bmatrix}
N & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{Ki} \\
\sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & \cdots & \sum X_{2i}X_{Ki} \\
\sum X_{3i} & \sum X_{3i}X_{2i} & \sum X_{3i}^2 & \cdots & \sum X_{3i}X_{Ki} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum X_{Ki} & \sum X_{Ki}X_{2i} & \sum X_{Ki}X_{3i} & \cdots & \sum X_{Ki}^2
\end{bmatrix}$$

$$\underset{(K\times 1)}{X'Y} = \begin{bmatrix} \sum Y_i \\ \sum X_{2i}Y_i \\ \sum X_{3i}Y_i \\ \vdots \\ \sum X_{Ki}Y_i \end{bmatrix} \qquad \underset{(K\times 1)}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}.$$

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.

Algunas equivalencias de notación

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

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

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

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

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

Interpretación de los coeficientes estimados por MCO

- $\hat{\beta}_1 = \widehat{E}(Y_i | X_{ki} = 0, \forall k = 2, \ldots, K)$. Valor esperado *estimado* de Y_i cuando las variables explicativas son todas cero.
- $\hat{\beta}_k = \dfrac{\partial \widehat{E(Y_i)}}{\partial X_{ki}} = \dfrac{\Delta \widehat{E(Y_i)}}{\Delta X_{ki}} \quad \forall k = 2, \ldots, K$. 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 5.6

Vamos a retomar ahora el Ejemplo 5.1 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[2]:

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:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS + u_i \quad i = 1, \ldots, 14 \quad (5.21)$$

Para estimar el modelo se utilizan las observaciones disponibles en el fichero data4-1.gdt y que son las siguientes[3]:

[2]Fuente: Ramanathan, Ramu (2002) *Introductory econometrics with applications.* Conjunto de datos data4-1.gdt

[3]Puedes acceder a estos datos ejecutando *gretl* → *En Archivo* → *Abrir datos* → *Archivo de muestra* → Elige Ramanathan, el fichero data4-1.gdt.

Obsv.	$PRICE$	$SQFT$	$BEDRMS$	$BATHS$
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 5.1: Datos de características de viviendas. Fichero 4-1.gdt.

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

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

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

• 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 redunda 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:

$$\triangle \widehat{PRICE}_i = 0,1548 \triangle SQFT_i - 21,588 \triangle BEDRMS_i - 12,192 \triangle BATHS_i =$$

$$= (0,1548 \times 500) - 21,588 \times 1 - 12,192 \times 0) = 77,4000 - 21,588 = 55,812$$

Ejemplo 5.7

Vamos a retomar ahora el Ejemplo 5.2 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[4]:

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:

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + u_i \quad i = 1, \ldots, 75 \tag{5.22}$$

Para estimar el modelo se utilizan las observaciones disponibles en el fichero *andy.gdt* y que son las siguientes[5]:
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:

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

• Interpretación de los signos obtenidos:
Para la variable P el signo es negativo lo que indica que la demanda es elástica.

[4]Fuente: Ramanathan, Ramu (2002) *Introductory econometrics with applications.* Carpeta PoE, conjunto de datos andy.gdt

[5]Puedes acceder a estos datos ejecutando *gretl* → *En Archivo* → *Abrir datos* → *Archivo de muestra* → Elige PoE, el fichero *andy.gdt.*

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:

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

5.4.1. 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: $\bar{Y} = \bar{\hat{Y}}$.

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

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

$$\sum_{i=1}^{N} \hat{u}_i = 0 \Leftrightarrow \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_K X_{Ki}) = 0$$

$$\sum Y_i - N\hat{\beta}_1 - \hat{\beta}_2 \sum X_{2i} - \ldots - \hat{\beta}_K \sum X_{Ki} = 0$$

$$\begin{aligned} \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \ldots + \hat{\beta}_K \sum X_{Ki} \\ \frac{1}{N}\sum Y_i &= \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{N}\sum X_{2i} + \ldots + \hat{\beta}_K \frac{1}{N}\sum X_{Ki} \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2\bar{X}_2 + \ldots + \hat{\beta}_K\bar{X}_K \end{aligned}$$

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.

5.4.2. 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.

$$SCT \quad = \quad SCE + SCR$$

siendo:

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.

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

Coeficiente de determinación, R^2

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

- 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, $\bar{R}^2$.

Coeficiente de determinación corregido, $\bar{R}^2$.

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

- 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 $K = 1$, $R^2 = \bar{R}^2$.
- 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.

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

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

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 $X_k \qquad k = 1, 2, \dots K$
- $r_{kk} = 1$, 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.

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Ejemplo 5.8

Con los datos de la Tabla 5.1 y los resultados de la estimación del modelo (5.21) calculamos el coeficiente de determinación y el coeficiente de determinación corregido:

$$\begin{aligned} SCT &= Y'Y - N\bar{Y}^2 = 1512980 - 14 \times 317,493^2 = 101754,7293 \\ SCR &= Y'Y - \hat{\beta}X'Y = 1512980 - 1496279,9 = 16700,1 \end{aligned}$$

$$\begin{aligned} R^2 &= 1 - \frac{SCR}{SCT} = 1 - \frac{16700,1}{101754,7293} = 0,835976 \\ \bar{R}^2 &= 1 - \frac{(N-1)}{(N-K)}(1 - R^2) = 1 - \frac{14-1}{14-4}(1 - 0,835976) = 0,786769 \end{aligned}$$

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 $BATHS$.

Ejemplo 5.9

Con los resultados de la regresión del modelo (5.22) y los datos del fichero *andy.gdt* calculamos el coeficiente de determinación y el coeficiente de determinación corregido:

$$\begin{aligned} SCT &= Y'Y - N\bar{Y}^2 = 452128{,}4100 - 75 \times 77,375^2 = 3111,6131 \\ SCR &= Y'Y - \hat{\beta}X'Y = 452128{,}4100 - 450409{,}4671 = 1718,9429 \end{aligned}$$

$$\begin{aligned} R^2 &= 1 - \frac{SCR}{SCT} = 1 - \frac{1718,943}{3111,6131} = 0,448258 \\ \bar{R}^2 &= 1 - \frac{(N-1)}{(N-K)}(1-R^2) = 1 - \frac{75-1}{75-3}(1-0,448258) = 0,432932 \end{aligned}$$

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

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

5.5. Propiedades de los estimadores MCO

Sea el modelo de regresión lineal general

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

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.

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

- Es insesgado.

$$E(\hat{\beta}|X) \quad = \quad E((\beta + (X'X)^{-1}X'u)|X) = \beta$$

  Donde para demostrarlo hemos utilizado $E(u|X) = 0$.

- Tiene varianza mínima entre todos los estimadores lineales e insesgados
  Dado que $E(u|X) = 0$ y $E(uu'|X) = \sigma^2 I_N$

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

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

$$\underset{(K\times K)}{V(\hat{\beta})} = \begin{bmatrix} V(\hat{\beta}_1) & Cov(\hat{\beta}_1,\hat{\beta}_2) & Cov(\hat{\beta}_1,\hat{\beta}_3) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_K) \\ Cov(\hat{\beta}_2,\hat{\beta}_1) & V(\hat{\beta}_2) & Cov(\hat{\beta}_2,\hat{\beta}_3) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_K) \\ Cov(\hat{\beta}_3,\hat{\beta}_1) & Cov(\hat{\beta}_3,\hat{\beta}_2) & V(\hat{\beta}_3) & \cdots & Cov(\hat{\beta}_3,\hat{\beta}_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_K,\hat{\beta}_1) & Cov(\hat{\beta}_K,\hat{\beta}_2) & Cov(\hat{\beta}_K,\hat{\beta}_3) & \cdots & V(\hat{\beta}_K) \end{bmatrix} =$$

$$= \sigma^2 \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1K} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2K} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{K1} & a_{K2} & a_{K3} & \cdots & a_{KK} \end{bmatrix} = \sigma^2(X'X)^{-1}$$

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".

5.5.1. 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** de σ^2:

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

Por tanto podremos utilizarlo como el estimador apropiado de la varianza de la perturbación. En términos de las variables observables mediante las matrices Y, X, podemos expresarlo:

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

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

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

Ejemplo 5.10

Con los datos de la Tabla 5.1 y los resultados de la estimación del modelo (5.21) se calcula la siguiente matriz de varianzas y covarianzas estimada:

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

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

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

Ejemplo 5.11

Con los datos disponibles en el fichero andy.gdt y los resultados de la estimación del modelo (5.22) se calcula la siguiente matriz de varianzas y covarianzas estimada:

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

$$\widehat{V}(\hat{\beta}_{MCO}) = 23,8742 \times \begin{bmatrix} 75,0000 & 426,5400 & 138,3000 \\ 426,5400 & 2445,7074 & 787,3810 \\ 138,3000 & 787,3810 & 306,2100 \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} 40,34330 & -6,79506 & -0,74842 \\ -6,79506 & 1,20120 & -0,01974 \\ -0,74842 & -0,01974 & 0,46675 \end{bmatrix}$$

Ejemplo 5.12

Vamos a retormar el Ejemplo 2.5 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:

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

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

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

Recordemos que β_1 es el salario esperado cuando el individuo es hombre, $\beta_1 + \beta_2$ es el salario esperado de una mujer y β_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:

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

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

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.

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

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

$$= \begin{bmatrix} N_H + N_M & N_M \\ N_M & N_M \end{bmatrix}^{-1} \begin{bmatrix} \sum W_H + \sum W_M \\ \sum W_M \end{bmatrix} = \begin{bmatrix} \bar{W}_H \\ \bar{W}_M - \bar{W}_H \end{bmatrix}$$

que sería el equivalente a estimar cada ecuación por separado, en las dos ecuaciones a las que da lugar el modelo (5.9):

$$W_i = \beta_1 + u_i \quad i = 1, \ldots, N_H \qquad \text{para los hombres}$$
$$W_i = \beta_1 + \beta_2 + u_i \quad i = 1, \ldots, N_M \qquad \text{para las mujeres}$$

• Alternativa de especificación :

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

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:

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

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

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

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

$$= \begin{bmatrix} N_H & 0 \\ 0 & N_M \end{bmatrix}^{-1} \begin{bmatrix} \sum W_H \\ \sum W_M \end{bmatrix} = \begin{bmatrix} \sum W_H / N_H \\ \sum W_M / N_M \end{bmatrix} = \begin{bmatrix} \bar{W}_H \\ \bar{W}_M \end{bmatrix}$$

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

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

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

5.6. Distribución del estimador MCO. Estimación por intervalo

5.6.1. Distribución del estimador de MCO bajo Normalidad

Si $Y = X\beta + u$, donde $u|X \sim N(0, \sigma^2 I_N)$, el estimador MCO, dado que es lineal en las perturbaciones, también seguirá una distribución Normal Multivariante

$$\hat{\beta}_{MCO}|X \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Para el k-ésimo coeficiente,

$$\hat{\beta}_k|X \sim N(\beta_k, \sigma^2 a_{kk})$$

donde a_{kk} es el elemento (k,k) de la matriz $(X'X)^{-1}$.

5.6.2. Estimación por intervalo

Para el k-ésimo coeficiente,

$$\hat{\beta}_k|X \sim N(\beta_k, \sigma^2 a_{kk})$$

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 β_k como:

$$IC(\beta_k)_{1-\alpha} = \left(\hat{\beta}_k \pm t_{\frac{\alpha}{2}(N-K)}\,\hat{\sigma}_{\hat{\beta}_k}\right)$$

Las estimaciones por intervalo 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 β_k. Además, como veremos más adelante, los intervalos sirven para realizar contraste de hipótesis.

Ejemplo 5.13

Para los valores estimados del modelo (5.22) 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 β_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 β_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$.

5.7. 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

$$u_i|X \sim N(0, \sigma^2) \longrightarrow \hat{\beta}_k \sim N(\beta_k, \sigma^2 a_{kk})$$

si conocemos todos los elementos incluido σ^2 podríamos contrastar hipótesis de la forma $H_0 : \beta_k = c$ con el siguiente estadístico:

$$\frac{\hat{\beta}_k - c}{\sigma\sqrt{a_{kk}}} \overset{H_0}{\sim} N(0,1)$$

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.

5.7.1. Contraste de restricciones sobre los coeficientes de regresión individuales. Estadístico t

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. 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 σ^2 es conocida todo es conocido en la distribución de β_k y el estadístico de contraste sería:

$$\frac{\hat{\beta}_k - c}{\sigma_{\hat{\beta}_k}} \overset{H_0}{\sim} N(0,1)$$

El caso más habitual es que σ^2 sea desconocida, en este caso podemos derivar el siguiente estadístico de contraste y distribución asociada cuando σ^2 es estimada con el estimador insesgado $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$:

$$\frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(N-K)}$$

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:

$$\frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(N-K)}$$

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 α 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 α 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 β_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 pertenece al intervalo, no rechazamos H_0 con un nivel de significación α y si no pertenece al intervalo, rechazamos H_0 con un nivel de significación α. Claramente se obtienen exactamente los mismos resultados utilizando los estadísticos de contraste individuales que utilizando los intervalos de confianza.

5.7.2. Contraste de restricciones sobre los coeficientes de regresión. Estadístico F

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

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

$$H_a : \text{alguna igualdad no se da}$$

En este caso podemos derivar el siguiente estadístico de contraste y distribución asociada:

$$\frac{R^2/K - 1}{1 - R^2/N - K} \overset{H_0}{\sim} \mathcal{F}_{(K-1,N-K)}$$

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 α dado, se rechaza la hipótesis nula y se concluye que las variables son conjuntamente significativas para explicar el comportamiento de la variable endógena.

Ejemplo 5.14

Vamos a mostrar un ejemplo sobre los contrastes de significatividad individual y conjunto con los resultados de la estimación del modelo (5.21). Primero vamos a escribir los resultados de la estimación de la forma habitual en que se muestran en la literatura:

$$\underset{(\hat{\sigma}_{\hat{\beta}_k})}{\widehat{PRICE}} = \underset{(88{,}30)}{129{,}062} + \underset{(0{,}03)}{0{,}154800}\, SQFT - \underset{(27{,}02)}{21{,}5875}\, BEDRMS - \underset{(43{,}25)}{12{,}1928}\, BATHS$$

$$N = 14 \quad R^2 = 0{,}8359 \quad \bar{R}^2 = 0{,}7868$$

Contrastes de significatividad individual, contrastamos:

$$\left.\begin{array}{l} H_0 : \beta_k = 0 \\ H_a : \beta_k \neq 0 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(14-4)}$$

- 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:

$$\left.\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \text{alguna igualdad no se da} \end{array}\right\} \quad \text{con} \quad \frac{R^2/K - 1}{1 - R^2/N - K} \overset{H_0}{\sim} \mathcal{F}_{(K-1,N-K)}$$

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 5.15

Vamos a mostrar un ejemplo sobre los contrastes de significatividad individual y conjunto con los resultados de la estimación del modelo (5.22). La ecuación de regresión muestral era:

$$\begin{array}{cccccccc} \widehat{S}_i & = & 118,914 & - & 7,90785 & P_i + & 1,86258 & A_i \quad i = 1, \ldots, 75 \\ (\hat{\sigma}_{\hat{\beta}_k}) & & (6{,}35164) & & (1{,}09599\ ) & & (0{,}683195) & \end{array}$$

$$N = 75 \quad R^2 = 0,448258 \qquad \bar{R}^2 = 0,432932$$

Como puede apreciarse en la ecuación anterior, se indica que bajo cada coeficiente estimado aparece su correspondiente desviación típica estimada[6].
Contrastes de significatividad individual, contrastamos:

$$\left.\begin{array}{l} H_0 : \beta_k = 0 \\ H_a : \beta_k \neq 0 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{(75-3)}$$

- 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:

$$\left.\begin{array}{l} H_0 : \beta_2 = \beta_3 = 0 \\ H_a : \beta_2 \neq 0 \quad y/o \quad \beta_3 \neq 0 \end{array}\right\} \quad \text{con} \quad \frac{R^2/K - 1}{1 - R^2/N - K} \overset{H_0}{\sim} \mathcal{F}_{(K-1,N-K)}$$

Evaluado 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:

[6] 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.

- $\beta_2 \geq 0$, 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:

$$\left.\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 < 0 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t_{(75-3)}$$

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[7]:

$$\frac{-7,908}{1,096} = -7,215 < -1,666 = t_{(72)\,|0,05}$$

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:

$$H_0 : \beta_3 \leq 1 \qquad H_1 : \beta_3 > 1$$

Luego contrastamos:

$$\left.\begin{array}{l} H_0 : \beta_3 = 1 \\ H_a : \beta_3 > 1 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_3 - 1}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t_{(75-3)}$$

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.

[7] 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 $\alpha = 5\,\%$.

Ejemplo 5.16

Utilizamos la función de salarios especificada para el año 2002 que se propuso en el Ejemplo 2.7:

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

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:

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

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:

$$\left.\begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t_{(N-3)}$$

- 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.

$$\left.\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array}\right\} \quad \text{con el estadístico y distribución} \quad \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t_{(N-3)}$$

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:

$$H_0 : w = c \qquad\qquad H_a : w \neq c$$

$$\frac{\hat{w} - c}{\hat{\sigma}_{\hat{w}}} \overset{H_0}{\sim} t_{(N-K)} \quad \text{si H}_0 \text{ es cierta}$$

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
$\mu_w = E(\hat{w}) = E(\hat{\beta}_2 + \hat{\beta}_3) = \beta_2 + \beta_3$
$\sigma_w^2 = V(\hat{w}) = E[\hat{w} - E(\hat{w})]^2 = E[(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)]^2 = V(\hat{\beta}_2) + V(\hat{\beta}_3) + 2Cov(\hat{\beta}_2, \hat{\beta}_3)$
$= \sigma^2(a_{22} + a_{33} + 2a_{23})$

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:

$$\frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\sqrt{\hat{V}(\hat{\beta}_2) + \hat{V}(\hat{\beta}_3) + 2\hat{Cov}(\hat{\beta}_2, \hat{\beta}_3)}} \overset{H_0}{\sim} t_{(N-K)}$$

o lo que es igual:

$$\frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\hat{\sigma}\sqrt{a_{22} + a_{33} + 2a_{23}}} \overset{H_0}{\sim} t_{(N-K)}$$

Con la regla de decisión habitual.

Ejemplo 5.17

Para contrastar:

$$H_0 : \beta_2 = \beta_3 \qquad H_a : \beta_2 \neq \beta_3$$

es equivalente a escribir:

$$H_0 : \beta_2 - \beta_3 = 0 \qquad H_a : \beta_2 - \beta_3 \neq 0$$

que podemos contrastar con el estadístico y distribución:

$$\frac{\hat{\beta}_2 - \hat{\beta}_3}{\hat{\sigma}\sqrt{a_{22} + a_{33} - 2a_{23}}} \overset{H_0}{\sim} t_{(N-K)}$$

Con la regla de decisión habitual.

5.7.3. Estimación mínimo-cuadrática sujeta a restricciones

Un aspecto básico de la inferencia estadística que se lleva a cabo en Economía es que el investigador sólo contrasta hipótesis en cuya validez está dispuesto a creer a priori, de modo que si su contraste no las rechaza, entonces pasa a imponerlas en la representación estructural que está considerando. Si la hipótesis nula no se rechaza, entonces sería muy interesante disponer de un procedimiento para estimar de nuevo el modelo, pero esta vez imponiendo ese conjunto de hipótesis que hemos

contrastado y no rechazado. La idea de eficiencia está ligada a la utilización óptima de toda la información disponible. Si se cree que los coeficientes del modelo satisfacen ciertas restricciones, entonces se ganaría eficiencia introduciendo dichas restricciones en el proceso de información.

En este caso vamos a encontrar el estimador que minimice la suma de cuadrados de los residuos, pero esta vez imponiendo las restricciones, es decir, se trata esta vez de resolver un problema de optimización sujeto a restricciones lineales.

Sea $\hat{\beta}_r$ el estimador resultante de resolver el lagrangiano de tal problema. A $\hat{\beta}_r$, se le llama **estimador de Mínimos Cuadrados Restringidos (MCR)** y es tal que:

$$\hat{\beta}_r = \hat{\beta}_{MCO} + \text{expresión matricial A}$$

donde $\hat{\beta}_{MCO}$ es el estimador Mínimo Cuadrático Ordinario sin restringir.

La matriz de varianzas y covarianzas de este estimador es:

$$V(\hat{\beta}_r) = \sigma^2(X'X)^{-1} - \sigma^2\text{expresión matricial B}$$

Resultados:

1. $\hat{\beta}_r$ es lineal en u.
2. **Si las restricciones que hemos impuesto son ciertas el estimador $\hat{\beta}_r$ es insesgado**. Si la restricción no se cumple el estimador restringido será sesgado, por lo tanto para comparar los estimadores MCR y MCO habrá, en general, que utilizar el criterio del error cuadrático medio.
3. Comparando las matrices de varianzas y covarianzas de los estimadores de mínimos cuadrados ordinarios y mínimos cuadrados restringidos se puede demostrar que

   $$V(\hat{\beta}) - V(\hat{\beta}_r)$$

   es una matriz semidefinida positiva aunque la restricción no se cumpla.

Estimar sujeto a restricciones mediante el estimador $\hat{\beta}_r$ es equivalente a estimar por MCO el modelo que cumple la restricción. A este modelo se le llama modelo restringido. Se puede demostrar que es posible utilizar la suma de cuadrados del modelo restringido ($\hat{u}_r'\hat{u}_r$) para hacer contraste de hipótesis mediante el estadístico siguiente:

$$\frac{\hat{u}_r'\hat{u}_r - \hat{u}'\hat{u}/q}{\hat{u}'\hat{u}/(N-K)} \overset{H_0}{\sim} \mathcal{F}_{(q,N-K)}$$

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.
- $\hat{u}'\hat{u}$ es la suma de cuadrados residual del modelo no restringido o lo que es igual el modelo de interés estimado por MCO.
- q es el número de restricciones que se contrastan.

A este estadístico se le conoce con el nombre de **estadístico de diferencias en las sumas residuales de cuadrados**. Es un estadístico de tipo general que puede ser utilizado para contrastar hipótesis lineales con solo especificar correctamente los modelos restringido y no restringido. Para su aplicación sólo es necesario obtener la SCR del modelo restringido y no restringido. El modelo restringido es aquel que cumple la hipótesis nula mientras que el modelo no restringido es el modelo de interés.

Vamos a estudiarlo en detalle en el ejemplo siguiente.

Ejemplo 5.18

Contraste de un subconjunto de coeficientes.
Supongamos el siguiente modelo de regresión:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \ldots + \alpha_r Z_{ri} + u_i \qquad i = 1, 2, \ldots, N$$

y queremos contrastar si el subconjunto de regresores $Z_{1i}, Z_{2i}, \ldots, Z_{ri}$ son conjuntamente significativos para explicar el comportamiento de la variable endógena. La hipótesis de contraste es:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$

$$H_a : \text{alguna igualdad no se de}$$

El estadístico de contraste y distribución son:

$$\frac{\hat{u}_r'\hat{u}_r - \hat{u}'\hat{u}/r}{\hat{u}'\hat{u}/(N-K)} \overset{H_0}{\sim} \mathcal{F}_{(r,N-K)} \tag{5.23}$$

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:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_{ri} \qquad i = 1, 2, \ldots, N$$

- $\hat{u}'\hat{u}$ es la suma de cuadrados residual del modelo no restringido o lo que es igual el modelo de interés estimado por MCO:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \ldots + \alpha_r Z_{ri} + u_i \qquad i = 1, 2, \ldots, N$$

- r es el número de restricciones que se contrastan, en este caso el número de coeficientes α_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 este 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.

Ejemplo 5.19

Cómo estimar el modelo restringido: Sea el MRLG,

$$\text{MNR:} \quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

donde queremos contrastar la hipótesis nula $H_0 : \beta_2 + \beta_3 = 1$ sustituyendo la restricción en el modelo encontramos el modelo restringido:

$$\text{MR:} \quad Y_i = \beta_1 + \beta_2 X_{2i} + (1-\beta_2) X_{3i} + u_{ri}$$

$$\underbrace{Y_i - X_{3i}}_{=Y_i^\star} = \beta_1^r + \beta_2^r (\underbrace{X_{2i} - X_{3i}}_{=X_i^\star}) + u_{ri}$$

$$Y_i^\star = \beta_1^r + \beta_2^r X_i^\star + u_{ri}$$

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

$$Y^\star = X^\star \beta^r + u_r$$

entonces

$$\hat{u}_r'\hat{u}_r = Y^{\star\prime}Y^\star - \hat{\beta}^{r\prime} X^{\star\prime} Y^\star$$

donde $Y^\star$ y $X^\star$ son las variables que quedan en el modelo restringido y

$$\begin{bmatrix} \hat{\beta}_1^r \\ \hat{\beta}_2^r \end{bmatrix} = \begin{bmatrix} N & \sum X_i^\star \\ \sum X_i^\star & \sum X_i^{\star 2} \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i^\star \\ \sum Y_i^\star X_i^\star \end{bmatrix}$$

$$= \begin{bmatrix} N & \sum(X_{2i} - X_{3i}) \\ \sum(X_{2i} - X_{3i}) & \sum(X_{2i} - X_{3i})^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum(Y_i - X_{3i}) \\ \sum(Y_i - X_{3i})(X_{2i} - X_{3i}) \end{bmatrix}$$

Ejemplo 5.20

El estadístico de diferencias en las sumas residuales de cuadrados puede ser utilizado para contrastar cualquier hipótesis lineal incluidas la significatividad individual y conjunta. Veamos que ocurre si hacemos el contraste de significatividad conjunta con este estadístico: $H_0 : \beta_2 = \beta_3 = \ldots = \beta_K = 0$. Para esta hipótesis el modelo restringido es

$$Y_i = \beta_1 + u_i$$

si estimamos el MR por MCO obtenemos:

$$Min_{\hat{\beta}_1} \sum \hat{u}_i^2 = Min_{\hat{\beta}_1} \sum (Y_i - \hat{\beta}_1)^2$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1) = 0 \longrightarrow \hat{\beta}_1^r = \bar{Y}$$

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.

$$F = \frac{(\hat{u}_r' \hat{u}_r - \hat{u}'\hat{u})/q}{\hat{u}'\hat{u}/N - K} = \frac{R^2/K - 1}{(1 - R^2)/N - K} \overset{H_0}{\sim} \mathcal{F}_{(K-1, N-K)}$$

estadístico que coincide con el obtenido para el contraste de significatividad conjunta.

5.8. 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.

5.8.1. Multicolinealidad exacta

Para verlo más claramente vamos a seguir un ejemplo. Sea el modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = 1, \ldots, N \tag{5.24}$$

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:

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i}^2 \\
2\left(\sum Y_i X_{2i}\right) &= 2\left(\hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 2\hat{\beta}_3) \sum X_{2i}^2\right)
\end{aligned}$$

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 sustituyendo $X_{3i} = 2X_{2i}$ en el modelo (5.24).

$$Y_i = \beta_1 + (\beta_2 + 2\beta_3)X_{2i} + u_i \qquad i = 1, 2, \ldots, N \tag{5.25}$$

donde podemos estimar de forma separada y única el coeficiente β_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 $|X'X| = 0$.

5.8.2. 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[8].

- 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[9] 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:

$$VIF_j = \frac{1}{\left(1 - R_j^2\right)} \qquad TOL_j = \frac{1}{VIF_j}$$

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:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum \left(X_{ji} - \bar{X}_j\right)^2} \frac{1}{\left(1 - R_j^2\right)} = \frac{\sigma^2}{\sum \left(X_{ji} - \bar{X}_j\right)^2} VIF_j$$

[8]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.

[9]En cada regresión se incluye el término constante como regresor pero no como variable dependiente.

donde β_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, σ^2 y $\sum \left(X_{ji} - \bar{X}_j\right)^2$, por lo que un alto VIF_j no es condición suficiente ni necesaria para que dichas varianzas sean elevadas ya que es posible que σ^2 sea pequeño o $\sum \left(X_{ji} - \bar{X}_j\right)^2$ 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.

5.9. 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.

5.9.1. Omisión de variables relevantes

Suponemos que el modelo correctamente especificado es:

$$Y = X\beta + u = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u = X_1\beta_1 + X_2\beta_2 + u \tag{5.26}$$

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 β_1 es un subvector de orden $(K_1 \times 1)$ y β_2 es un subvector de orden $(K_2 \times 1)$. Pero nosotros estimamos el siguiente modelo incorrectamente especificado:

$$Y = X_1\beta_1 + v \qquad \text{donde} \qquad v = X_2\beta_2 + u \tag{5.27}$$

El modelo (5.27) 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 β_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:

$$\hat{\sigma}^2 = \frac{\hat{v}'\hat{v}}{N - K_1} \longrightarrow E(\hat{\sigma}^2) = \frac{E(\hat{v}'\hat{v})}{N - K_1} \neq \sigma^2$$

5.9.2. Inclusión de variables irrelevantes

Este caso formalmente es justo el inverso del anterior. El modelo correctamente especificado es:

$$Y = X_1\beta_1 + u \qquad u \sim N(0, \sigma^2 I) \tag{5.28}$$

y el modelo estimado es:

$$Y = X_1\beta_1 + X_2\beta_2 + v \tag{5.29}$$

donde aparecen las variables irrelevantes en la matriz X_2 de orden $(N \times K_2)$ con unos coeficientes, β_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:

$$Y = X_1\beta_1 + X_2\,0 + u \tag{5.30}$$

$$E\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = E\left(\begin{bmatrix}\beta_1\\ 0\end{bmatrix}+\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1}\begin{bmatrix}X_1'u\\ X_2'u\end{bmatrix}\right) =$$

$$= \begin{bmatrix}\beta_1\\ 0\end{bmatrix}+\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1}\underbrace{\begin{bmatrix}X_1'E(u)\\ X_2'E(u)\end{bmatrix}}_{0} = \begin{bmatrix}\beta_1\\ 0\end{bmatrix}$$

ya que X es fija y $E(u) = 0$. Por lo tanto, el estimador de (5.29) sigue siendo **insesgado** aunque se incluyan variables irrelevantes.

- Las matriz de varianzas y covarianzas es $\mathrm{V}(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- El estimador de la varianza de las perturbaciones del modelo (5.29) es un estimador **insesgado** de σ^2

$$\hat{\sigma}^2 = \frac{\hat{v}'\hat{v}}{N-(K_1+K_2)}$$

5.10. 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:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + u_i.$$

Dada una nueva observación de las variables explicativas,

$$X_p' = \begin{bmatrix}1 & X_{2p} & \cdots & X_{Kp}\end{bmatrix} \qquad p \notin \{1, 2, \ldots, N\}$$

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:

$$Y_p = \beta_1 + \beta_2 X_{2p} + \ldots + \beta_K X_{Kp} + u_p$$

Para abreviar, utilizaremos la expresión vectorial:

$$Y_p = X_p'\beta + u_p$$

Dada la información muestral disponible (no conocemos β 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:

$$\hat{Y_p} = \hat{\beta}_1 + \hat{\beta}_2 X_{2p} + \ldots + \hat{\beta}_K X_{Kp}.$$

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 β 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,

$$e_p \sim N(0, \sigma^2(\,1 + X_p'\left(\,X'X\,\right)^{-1} X_p\,))$$

Demostración:

$$\begin{aligned} e_p &= Y_p - \hat{Y}_p = X_p'\,\beta + u_p - X_p'\,\hat{\beta} = \\ &= u_p - X_p'\,(\hat{\beta} - \beta) \end{aligned} \tag{5.31}$$

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:

$$E(e_p) = E\left[u_p - X_p'\,(\hat{\beta} - \beta)\right] = 0 - X_p'\,(\beta - \beta) = 0$$

$$\begin{aligned}
V(e_p) &= E\left[e_p - E(e_p)\right]\left[e_p - E(e_p)\right]' = \\
&= E\left(e_p\, e_p'\right) = \\
&= E\left[\left(u_p - X_p'\left(\hat{\beta}-\beta\right)\right)\left(u_p - X_p'\left(\hat{\beta}-\beta\right)\right)'\right] = \\
&= E\left[u_p\, u_p'\right] + E\left[X_p'\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)' X_p\right] - 2X_p'\, E\left[\left(\hat{\beta}-\beta\right) u_p'\right] = \\
&= E\left(u_p^2\right) + X_p'\, E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right] X_p - 2X_p'\, E\left[\left(X'X\right)^{-1} X'\, u\, u_p\right] = \\
&= \sigma^2 + \sigma^2 X_p'\left(X'X\right)^{-1} X_p - 0 = \\
&= \sigma^2\left(1 + X_p'\left(X'X\right)^{-1} X_p\right)
\end{aligned}$$

Por tanto:

$$e_p \sim N(0, \sigma^2\left(1 + X_p'\left(X'X\right)^{-1} X_p\right))$$

Tipificando el error de predicción queda:

$$\frac{e_p - 0}{\sigma\sqrt{1 + X_p'\left(X'X\right)^{-1} X_p}} \sim N(0,1)$$

El problema es que σ^2 es desconocida. Utilizando que e_p y $\hat{\sigma}^2$ obtenemos

$$\frac{e_p}{\hat{\sigma}\sqrt{1 + X_p'\left(X'X\right)^{-1} X_p}} \sim t_{(N-K)}$$

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:

$$Pr\left[-t_{\frac{\alpha}{2}(N-K)} \leq \frac{Y_p - \hat{Y}_p}{\hat{\sigma}_{e_p}} \leq t_{\frac{\alpha}{2}(N-K)}\right] = 1-\alpha$$

$$Pr\left[\hat{Y}_p - t_{\frac{\alpha}{2}(N-K)} \cdot \hat{\sigma}_{e_p} \leq Y_p \leq \hat{Y}_p + t_{\frac{\alpha}{2}(N-K)} \cdot \hat{\sigma}_{e_p}\right] = 1-\alpha$$

$$\boxed{IC_{1-\alpha}(Y_p) = \left(\hat{Y}_p - t_{\frac{\alpha}{2}(N-K)}\,\hat{\sigma}_{e_p}\ ,\ \hat{Y}_p + t_{\frac{\alpha}{2}(N-K)}\,\hat{\sigma}_{e_p}\right)}$$

5.11. Estimación, contraste de hipótesis y predicción en el 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.

Vamos a utilizar como ejemplo la estimación realizada con el fichero de datos *data4-1.gdt*:

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

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

Modelo 1: MCO, usando las observaciones 1–14
Variable dependiente: price

	Coeficiente	Desv. Típica	Estadístico t	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 *Coeficiente* aparece la estimación del coeficientre 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.

- Por ejemplo para contrastar:

$$H_0 : \beta_3 = \beta_4 \qquad \text{versus} \qquad H_a : \beta_3 \neq \beta_4$$

cuyo estadístico de contraste y distribución asociada son:

$$\frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_3} + \hat{\sigma}^2_{\hat{\beta}_4} - 2 \times \widehat{Cov}(\hat{\beta}_3, \hat{\beta}_4)}} \sim t_{N-4}$$

en la pestaña *Contrastes* seleccionamos *Restricciones lineales* y escribimos b3-b4=0 y *gretl* nos devuelve el siguiente resultado[10]
Restricción:
b[bedrms] - b[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:

	Coeficiente	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:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3(BEDRMS_i + BATHS) + u_i \quad i = 1, \ldots, 14$$

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

[10] Notar que Gretl realiza todos los contrastes con el estadístico de diferencias en las sumas residuales de cuadrados. Además cuando $q = 1 \quad t^2 = F$. Luego $t_c = \sqrt{0,0266334}$

- 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.

En la pestaña *Análisis* encontramos 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	Coeficiente	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

5.11.1. 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.

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[11]. 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 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.

[11] Para restablecer el tamaño muestral inicial pinchar en *Muestra → Recuperar el rango completo.*

Para comenzar, consideramos el siguiente modelo:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + u_i \qquad i = 1, \ldots, 14 \tag{5.32}$$

donde

- α_1: precio medio de una vivienda sin piscina.
- $\alpha_1 + \alpha_2$: precio medio de una vivienda con piscina.
- α_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	Coeficiente	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

$$\underset{\substack{\\ (10,13)}}{\widehat{PRICE}_i = 306,111} + \underset{(0,63)}{31,869\, POOL_i} \qquad i = 1, \ldots, 14$$

Para contrastar en el modelo (5.32) 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 (5.32) el modelo no restringido mientras que el modelo restringido es $PRICE_i = \alpha_1 + u_i \quad i = 1, \ldots, 14$.

Supongamos que ampliamos el modelo (5.32) 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:

$$FIREPL_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene chimenea} \\ 0 & \text{en caso contrario} \end{cases}$$

$$FAMROOM_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene sala de estar} \\ 0 & \text{en caso contrario} \end{cases}$$

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:

$$PRICE_i = \gamma_1 + \gamma_2 POOL_i + \gamma_3 FAMROOM_i + \gamma_4 FIREPL_i$$

$$+\beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \ldots, 14 \tag{5.33}$$

Donde lo primero a notar es que en el modelo (5.33), 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	Coeficiente	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.

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 (5.32) y (5.33), ninguna de las variables añadidas en el último es significativa individualmente[12]. 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 (5.32)) y no restringido (modelo (5.33)). 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 (**??**) 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.

5.11.2. El p-valor y conclusiones del contraste

Otra forma de llevar a cabo el contraste es utilizar el **valor-p**. Este valor es una probabilidad e indica cuál sería el menor nivel de significación que se tendría que elegir para rechazar la hipótesis nula, dada la realización muestral del estadístico. Si el contraste es a dos colas, el *valor-p* es dos veces el área a la derecha de la realización muestral del estadístico en valor absoluto, en la distribución de éste bajo la hipótesis nula, esto es

$$\textit{valor-p} = 2\,\mathrm{P}(\mathrm{t_j} > \mathrm{t_j^m}|\mathrm{H_0})$$

Si el contraste es a una cola, el *valor-p* sería el área a la derecha de la realización muestral del estadístico en valor absoluto, en la distribución de éste bajo la hipótesis nula, esto es *valor-p* = $\mathrm{P}(\mathrm{t_j} > \mathrm{t_j^m}|\mathrm{H_0})$. A mayor *valor-p*, mayor sería la probabilidad de error de tipo I si elegimos rechazar la hipótesis nula. Luego a mayor *valor-p* menor evidencia contra la hipótesis nula y por el contrario a menor *valor-p* mayor evidencia contra la hipótesis nula. El cálculo del valor-p es más complicado que elegir el nivel de significatividad a priori por lo que generalmente se realiza en el ordenador.

En la práctica se compara el *valor-p* con el valor 0,05 y si *valor-p* $< 0,05$ se rechaza la H_0 mientras que si *valor-p* $> 0,05$ no se rechaza la H_0.

[12]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.

5.11.3. Predicción en gretl

Para hacer predicción con *gretl* debemos incorporar los nuevos datos (X_p) a la base de datos mediante

$$Datos \rightarrow Seleccionar\ todos$$

A continuación, pincharemos la opción

$$Datos \rightarrow Añadir\ Observaciones$$

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

$$Muestra \rightarrow Establecer\ rango$$

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

$$Análisis \rightarrow Predicciones$$

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 prediccion.

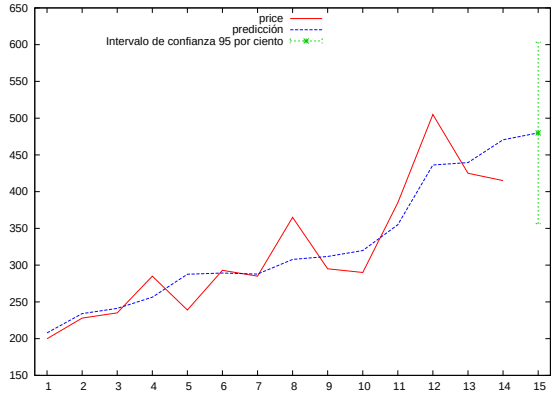
Utilizando los resultados obtenidos en el Ejemplo 5.10 se va a predecir la variable $PRICE$. Los resultados que muestra Gretl son los siguientes:

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
U 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 $95\,\%$.

5.12. Bibliografía del tema

Referencias bibliográficas básicas:

• Teórica:

[1] Stock, James H. y Mark Watson (2012). Introducción a la Econometría. Pearson.

[2] Wooldridge, J.M. (2006). Introducción a la Econometría. Ed. Thomson Learning, 2ª edición.

• 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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] Fernández, A., P. González, M. Regúlez, P. Moral, V. Esteban (2005). Ejercicios de Econometría. Editorial McGraw-Hill.

[4] Gujarati, D. y Porter, D.C. (2010). Econometría. Editorial McGraw-Hill, Madrid. 5ª edición.

[5] Ramanathan, R. (2002), Introductory Econometrics with applications, Ed. South-Western, 5th. edition.

Tema 6

Heterocedasticidad. Implicaciones

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 5 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. Explicar que se entiende por un modelo de regresión lineal con heterocedasticidad.
2. Analizar gráficamente la posible existencia de heterocedasticidad y saber contrastarla utilizando el estadístico de White.
3. Describir las propiedades del estimador MCO bajo heterocedasticidad.

4. Realizar contraste de hipótesis cuando la perturbación del modelo es heterocedástica.
5. Utilizar el software gretl para contrastar la existencia de heterocedasticidad en las perturbaciones y realizar contraste de hipótesis en los coeficientes de un modelo con perturbación heterocedástica.

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:

- Stock and Watson, J. M. (2012). Cap. 5.
- Wooldridge, J.M. (2006). Cap. 8.

6.1. Sobre las perturbaciones: contrastes de heterocedasticidad

6.1.1. 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 σ^2. Llamábamos a este supuesto *homocedasticidad* y lo denotábamos: $V(u_i) = \sigma^2$ ó lo que es igual $E(u_i^2|X) = \sigma^2 \quad \forall i$. La varianza σ^2 es una medida de dispersión de u_i alrededor de su media , $E(u_i|X) = 0$, o equivalentemente, una medida de dispersión de la variable dependiente Y_i alrededor de su media $\beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$. 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.

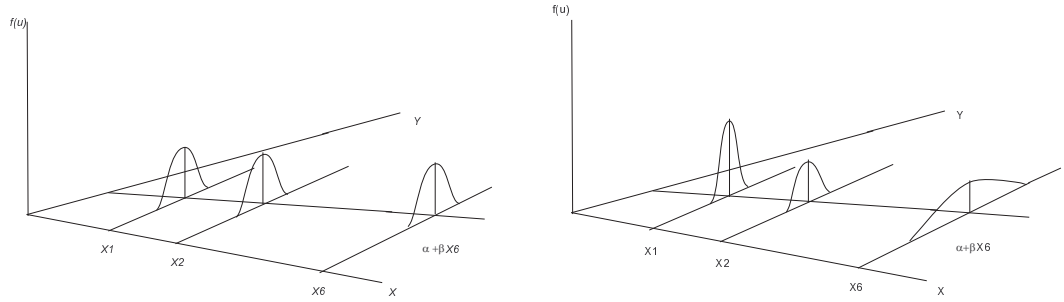


Figura 6.1: Perturbaciones homocedásticas versus heterocedásticas

En la Figura 6.1 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:

$$E(u_i^2|X) = \sigma_i^2$$

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 matriz de varianzas y covarianzas de la perturbación será:

$$E(uu'|X) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_N^2 \end{bmatrix} = \sum$$

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 6.1 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:

  $$GS_i = \beta_1 + \beta_2 R_i + \beta_3 SEN_i + \beta_4 POP_i + u_i \qquad i = 1, \dots, N \tag{6.1}$$

  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 6.2 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:

  $$C_i = \beta_1 + \beta_2 R_i + u_i \qquad i = 1, \dots, N \tag{6.2}$$

  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 6.3 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:

$$D_i = \beta_1 + \beta_2 B_i + u_i \qquad i = 1, \ldots, N \tag{6.3}$$

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 6.4 Supongamos un investigador que desea estimar los coeficientes del siguiente modelo:

$$Y_j = \beta_1 + \beta_2 X_j + u_j \qquad j = 1, \ldots, N \tag{6.4}$$

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:

$$\bar{Y}_i = \beta_1 + \beta_2 \bar{X}_i + \bar{u}_i \qquad i = 1, \ldots, m \tag{6.5}$$

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,

$$Var(\bar{u}_i) = \frac{\sigma^2}{n_i} \qquad i = 1, \ldots, m.$$

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 6.5 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:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad i = 1, \ldots, 20 \tag{6.6}$$

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:

$$\begin{array}{ll} Var(u_i) = \alpha_1 + \alpha_2 & \text{para las observaciones correspondientes a las mujeres} \\ Var(u_i) = \alpha_1 & \text{para las observaciones correspondientes a los hombres} \end{array}$$

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 $E(u|X) = 0$. 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 5 es mínima bajo las hipótesis básicas, es decir bajo $E(u'u|X) = \sigma^2 I_N$. Ahora, sin embargo, éstas no se cumplen: estamos considerando el supuesto de heterocedasticidad por tanto $E(u_i^2) \neq \sigma^2$, $(E(uu'|X) = \sum)$ 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 σ^2 propuesto $\hat{\sigma}^2 = \frac{SCR}{N-K}$ 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 σ^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.

6.1.2. 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 (6.9) recogido en el Ejemplo 6.1:

$$GS_i = \beta_1 + \beta_2 R_i + \beta_3 SEN_i + \beta_4 POP_i + u_i \qquad i = 1, \ldots, N$$

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.

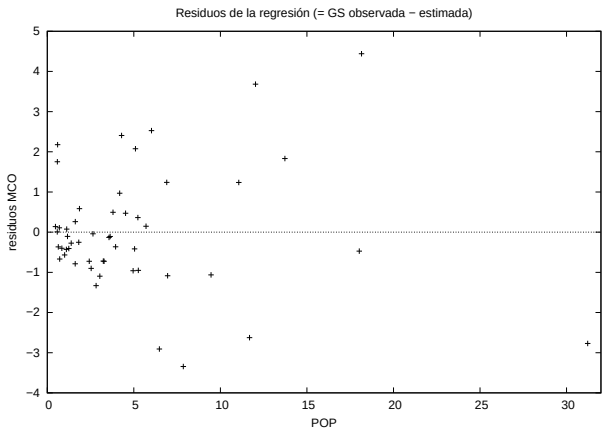


Figura 6.2: Residuos MCO versus POP

Si el gráfico es como el recogido en la Figura 6.2 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 proponer, por ejemplo:

$$E(u_i^2) = \sigma^2 POP_i \qquad i = 1, 2, \ldots, N$$

Si el gráfico de los residuos MCO frente a POP *hubiera* sido como el recogido en la Figura 6.3 supondríamos que el aumento en la varianza de u_i es inversamente proporcional a POP_i y propondríamos:

$$E(u_i^2) = \sigma^2 POP_i^{-1} \qquad i = 1, 2, \ldots, N$$

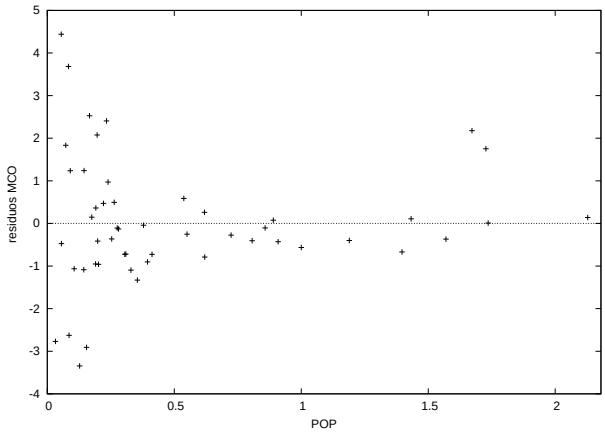


Figura 6.3: Residuos MCO versus POP

También podemos optar por dibujar la serie de los residuos al cuadrados MCO frente a la variable que creemos causa la heterocedasticidad como se muestra en la Figura 6.4. En el gráfico de la izquierda se muestran los pares $(SEN_i, \hat{u}_{MCO,i})$, en el gráfico de la derecha se muestran los pares $(SEN_i, \hat{u}^2_{MCO,i})$. Ambos gráficos muestran la misma información, muestran que la variabilidad de los residuos se incrementa con SEN y podríamos proponer, por ejemplo $Var(u_i) = E(u_i^2) = \sigma^2 SEN_i$.

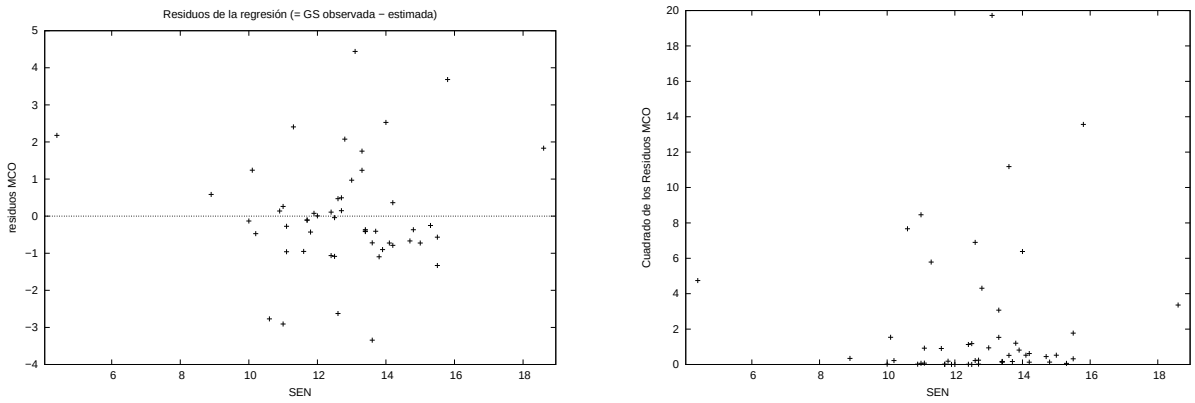


Figura 6.4: Residuos MCO y sus cuadrados versus SEN

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, $\sigma^2 = 1$.

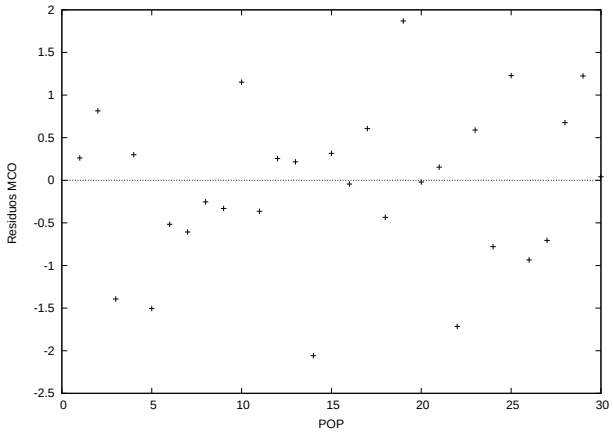


Figura 6.5: Perturbaciones homocedásticas

Si la gráfica entre $\hat{u}_{MCO,i}$ y POP_i *hubiera* resultado como la de la Figura 6.5, 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 6.6 donde claramente la dispersión de los residuos para las mujeres es mucho mayor que para los hombres.

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 σ_i^2. El análisis gráfico no es una pérdida de tiempo ya que la relación entre X_{ki} y $\hat{u}_{MCO,i}$ nos

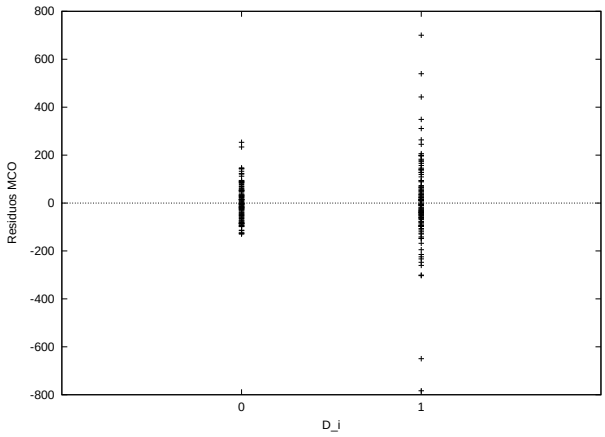


Figura 6.6: Residuos MCO frente a una variable ficticia

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 σ_i^2. Además *gretl* lo proporciona directamente.

6.1.3. 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:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{6.7}$$

$$\begin{aligned} H_0 &: E(u_i^2|X) = \sigma^2 \quad \forall i \\ H_a &: E(u_i^2|X) = \sigma_i^2 \end{aligned}$$

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:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2t} X_{3i} + \omega_i \tag{6.8}$$

   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:

$$H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_6 = 0$$

3. El estadístico de contraste es $\lambda = NR^2$ donde R^2 es el coeficiente de determinación de la regresión auxiliar (6.10). Rechazamos H_0 si $NR^2 > \chi_{(p)|\alpha}$ 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.

6.1.4. Estimador robusto de la matriz de varianzas y covarianzas del estimador MCO bajo heterocedasticidad. Contraste de hipótesis

- En presencia de heterocedasticidad los estimadores de MCO son lineales e insesgados pero ineficientes. Su matriz de varianzas y covarianzas se define $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$.

- El estimador de la matriz de varianzas y covarianzas del estimador MCO cuando no tenemos en cuenta la existencia de heterocedasticidad es:

$$\widehat{Var}(\hat{\beta}_{MCO}) = \hat{\sigma}^2(X'X)^{-1} \qquad donde \qquad \hat{\sigma^2} = \frac{\hat{u}'_{MCO}\hat{u}_{MCO}}{N-k}$$

  utilizar este estimador para hacer inferencia no es adecuado.

- Los estadísticos t y F habituales para hacer inferencia sobre β definidos en base a este estimador de la matriz de varianzas y covarianzas del estimador MCO **son inapropiados** ya que:

  - $\hat{\sigma}^2$ es un estimador sesgado

- y además $(X'X)^{-1} \neq (X'X)^{-1}X'\Omega X(X'X)^{-1}$.

- Para encontrar estas varianzas y covarianzas es necesario conocer Ω.

- La dificultad que entraña el conocimiento de Ω hace interesante el poder contar con una estimación consistente, y robusta a la posible existencia de heterocedasticidad, de $Var(\hat{\beta}_{MCO})$ y de esta forma derivar estadísticos válidos, al menos asintóticamente, para contrastar hipótesis sobre el vector de coeficientes β.

- White (1980) demuestra que un estimador consistente de la matriz de varianzas y covarianzas asintótica de $\hat{\beta}_{MCO}$ en presencia de heterocedasticidad es:

$$(X'X)^{-1}(X'SX)(X'X)^{-1} = \widehat{Var}(\hat{\beta}_{MCO})_{White}$$

  donde $S = diag(\hat{u}_1^2, \hat{u}_2^2, \ldots, \hat{u}_N^2)$ esta matriz de varianzas y covarianzas consistente asintóticamente puede ser utilizada para hacer inferencia válida al menos asintóticamente utilizando $\hat{\beta}_{MCO}$ sin tener que especificar a priori la estructura de heterocedasticidad.

- Así un estadístico válido para contrastar cuando existe heterocedasticidad es:

$$\begin{array}{ll} H_0 : \beta_j = c & \\ H_a : \beta_j \neq c & \dfrac{\hat{\beta}_{j,MCO} - c}{\widehat{desv}(\hat{\beta}_{j,MCO})_{White}} \xrightarrow{H_0} N(0,1) \end{array}$$

  Donde $\widehat{desv}(\hat{\beta}_{j,MCO})_{White}$ se busca apropiadamente en la matrix $\widehat{Var}(\hat{\beta}_{MCO})_{White}$ La regla de decisión es la habitual.

6.2. Heterocedasticidad en gretl

Ejemplo

El Departamento de Sanidad de E.E.U.U. quiere estudiar la relación entre el gasto sanitario agregado en billones de dólares (*exphlth*), la renta personal disponible agregada también en billones de dólares (*income*), el porcentaje de población que supera los 65 años en el año 2005 (seniors) y la población en millones (*pop*). Para ello encarga un estudio a dos becarios de la facultad de Económicas de Harvard poniendo a su disposición datos de 2005 para dichas variables sobre 51 estados americanos[1].

Puedes acceder a estos datos ejecutando GRETL → En Archivo → Abrir datos → Archivo de muestra → Elige Ramanathan, fichero **data8-3.gdt**.

1. Escribe el modelo que te permita analizar la influencia de las variables explicativas income, seniors y pop sobre la variable exphlth. Estímalo por MCO. Interpreta los resultados de la estimación en términos de significatividad y bondad del ajuste.

[1]Fuente: Ramanathan, Ramu (2002): *Introductory Econometrics with Applications*, fichero data8-3.gdt.

El modelo a estimar es:

$$EXPHLTH_i = \beta_1 + \beta_2 INCOME_i + \beta_3 POP_i + \beta_4 SEN_i + u_i \qquad i = 1, \ldots, N \tag{6.9}$$

Los resultados de la estimación por Mínimos Cuadrados Ordinarios son los siguientes:

Modelo 1: MCO, usando las observaciones 1–51
Variable dependiente: exphlth

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	−3.93356	1.34384	−2.9271	0.0053
income	0.106889	0.0141020	7.5797	0.0000
pop	0.784397	0.312314	2.5116	0.0155
seniors	0.314650	0.102968	3.0558	0.0037

Media de la vble. dep.	15.26494	D.T. de la vble. dep.	17.88771
Suma de cuad. residuos	112.4706	D.T. de la regresión	1.546929
R^2	0.992970	R^2 corregido	0.992521
$F(3, 47)$	2212.858	Valor p (de F)	1.40e–50
Log-verosimilitud	−92.53295	Criterio de Akaike	193.0659
Criterio de Schwarz	200.7932	Hannan–Quinn	196.0187

Los resultados de la estimación muestran un buen ajuste, explicamos el $99,3\,\%$ de la variabilidad del gasto sanitario con la variación de las variables exógenas. Además las variables son significativas a nivel individual y conjunto.

2. Obtén los siguientes gráficos y comenta la información que te proporcionan

   a) Gráfico de la serie de residuos MCO.

   b) Gráfico de residuos MCO sobre la variable *income*.

   c) Gráfico de residuos MCO sobre la variable *pop*.

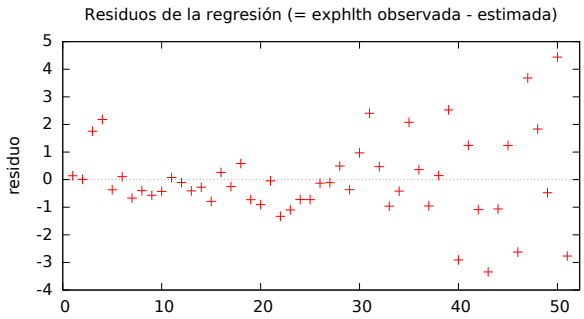


Figura 6.7: Residuos MCO

La Figura 6.7 muestra los residuos MCO, $\hat{u}_{MCO,i}$ por observación. Los residuos aparecen centrados en torno al valor cero como corresponde a su media. Para las 25 primeras observaciones la dispersión de los residuos permanece más o menos constante salvo en dos observaciones. En adelante la observación 25 aumenta la dispersión en los residuos.

3. Gráfico de la serie de residuos MCO sobre la variable INCOME.

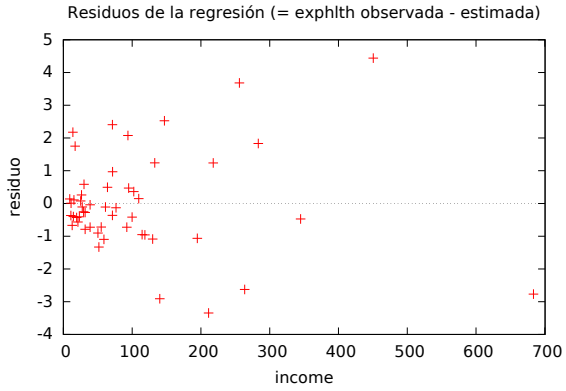


Figura 6.8: Residuos MCO versus INCOME

La Figura 6.8 muestra los pares $(INCOME_i, \hat{u}_{MCO,i})$. Para valores de $INCOME$ en el intervalo $(0, 100)$ vemos una alta concentración de observaciones donde la dispersión de los residuos permanece más o menos constante salvo en dos observaciones. En adelante al valor 100 y a medida que $INCOME$ toma valores mayores aumenta la dispersión en los residuos y la concentración desaparece.

4. Gráfico de residuos MCO sobre la variable POP.

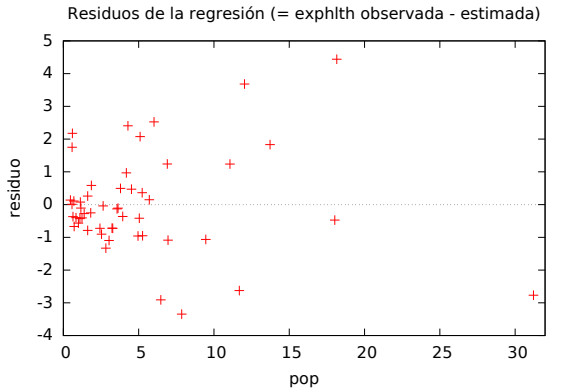


Figura 6.9: Residuos MCO versus POP

La Figura 6.9 muestra los pares $(POP_i, \hat{u}_{MCO,i})$. Para valores de POP en el intervalo $(0, 5)$ vemos una alta concentración de observaciones donde la dispersión de los residuos permanece más o menos constante salvo en dos observaciones. En adelante al valor 5 y a medida que POP toma valores mayores aumenta la dispersión en los residuos y la concentración desaparece. Este gráfico replica la forma del comentado anteriormente.

5. Contrasta la existencia de heterocedasticidad.

   Regresión auxiliar:

   $$\hat{u}_i^2 = \alpha_1 + \alpha_2 INCOME_i + \alpha_3 POP_i + \alpha_4 SEN_i + \alpha_5 INCOME_i^2 + \alpha_6 POP_i^2$$

   $$+\alpha_7 SEN_i^2 + \alpha_8 INCOME_i POP_i + \alpha_9 INCOME_i SEN_i$$

   $$+\alpha_{11} POP_i SEN_i + \omega_i \tag{6.10}$$

   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:

   $$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \ldots = \alpha_{10} = 0$$

   El estadístico de contraste es $\lambda = NR^2$ donde R^2 es el coeficiente de determinación de la regresión auxiliar (6.10). Rechazamos H_0 si $NR^2 > \chi_{(p)|\alpha}$ siendo p el número de coeficientes en la regresión auxiliar sin incluir el término independiente, en el ejemplo $p = 9$.

   Encontramos este contraste en Gretl en la pantalla de resultados de la estimación MCO pinchando en la pestaña Contrates y seleccionando:

   Heterocedasticidad → *Contraste de White*

   Gretl nos devuelve el siguiente resultado:

   Contraste de heterocedasticidad de White MCO, usando las observaciones 1-51
   Variable dependiente: $uhat^2$

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	10.8361	4.89514	2.214	0.0325 **
income	-0.712618	0.348653	-2.044	0.0474 **
pop	15.7074	7.42431	2.116	0.0405 **
seniors	-2.00213	0.965046	-2.075	0.0443 **
sq_income	-0.000884586	0.00102650	-0.861	0.3938
$X2\_X3$	0.0515366	0.0467038	1.103	0.2763
$X2\_X4$	0.0561182	0.0258280	2.173	0.0356 **
sq_pop	-0.715606	0.534564	-1.339	0.1881
$X3\_X4$	-1.17973	0.547742	-2.154	0.0372 **
sq_seniors	0.0860328	0.0461497	1.864	0.0695 *

R-cuadrado = 0.778511
Estadístico de contraste: $TR^2 = 39{,}704042$,
con valor p= $P(Chi - cuadrado(9) > 39{,}704042) = 0{,}000009$

$TR^2 = 39{,}704042 > \chi^2_{(9)0,05} = 16,919$ luego rechazamos la hipótesis nula para $\alpha = 5\,\%$ y existe heterocedasticidad

6. A la vista de lo resultados del contraste contrasta apropiadamente la significatividad individual de la variables POP.

   Dado que existe heterocedasticidad el estimador de MCO es lineal e insesgado pero no es de varianza mínima. Además la inferencia en base a los estadístico t y F habituales no es válida. Para poder realizar inferencia válida con el estimador MCO debemos estimar su matriz de varianzas y covarianzas de forma robusta con el estimador de White.

   Encontramos esta estimación en Gretl en la pestaña *Modelo* pinchamos en *Mínimos Cuadrados Ordinarios*, seleccionamos apropiadamente las variables y *Clikcamos* en **Desviaciones típicas Robustas** eligiendo la opción **HC0**

   Gretl nos devuelve los siguientes resultados:

   Modelo 2: MCO, usando las observaciones 1–51
   Variable dependiente: exphlth
   Desviaciones típicas robustas ante heterocedasticidad, variante HC0

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	−3.93356	1.54437	−2.5470	0.0142
income	0.106889	0.0259509	4.1189	0.0002
pop	0.784397	0.540137	1.4522	0.1531
seniors	0.314650	0.118378	2.6580	0.0107

Media de la vble. dep.	15.26494	D.T. de la vble. dep.	17.88771
Suma de cuad. residuos	112.4706	D.T. de la regresión	1.546929
R^2	0.992970	R^2 corregido	0.992521
$F(3, 47)$	1026.139	Valor p (de F)	8.05e–43
Log-verosimilitud	−92.53295	Criterio de Akaike	193.0659
Criterio de Schwarz	200.7932	Hannan–Quinn	196.0187

   realizamos el contraste pedido, contrastamos:

   $$\begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array} \qquad \frac{\hat{\beta}_{3,MCO}}{\widehat{desv}(\hat{\beta}_{3,MCO})_{White}} \xrightarrow{H_0} N(0,1)$$

   El valor muestral del estadístico que nos proporciona *gretl* es $1,4522 < 1,96 = N(0,1)_{0,025}$ luego no rechazamos la hipótesis nula para un nivel de significatividad del 5 % luego la variable Población no es significativa para explicar el gasto en sanidad.

6.3. 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. ($http://ocw.ehu.es/ciencias-sociales-y-juridicas/analisis-de-regresion-con-greti/Course_listing$).

[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.