Universidad del País Vasco

Euskal Herriko Unibertsitatea

Análisis de Regresión con Gretl

Autores:

M. Victoria Esteban
M. Paz Moral
Susan Orbe
Marta Regúlez
Ainhoa Zarraga
Marian Zubia

Departamento de Economía Aplicada III
Econometría y Estadística
Facultad de Ciencias Económicas y Empresariales
UPV/EHU

Contenido

Figuras

Tablas

Tema 1

Gretl y la Econometría

Contenido

1.1. Introducción

Este curso se dirige a aquellas personas interesadas en aprender a interpretar información estadística sobre la realidad económica. 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 en este curso nos centramos en el modelo más sencillo, y que da nombre a la asignatura, el **modelo de regresión lineal general**. Este modelo explica el comportamiento de una única variable económica o de otra índole más general.

Por otro lado, este curso tiene un carácter totalmente aplicado, en el que los ejemplos prácticos sirven para introducir los conceptos estadístico-econométricos. Así, una parte importante del curso se dedica a estudiar casos prácticos, en los que el estudiante aprenderá a manejar un software econométrico y a interpretar adecuadamente los resultados obtenidos. 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 introducción al análisis econométrico.

Este primer tema se organiza de la siguiente forma: la sección 2 presenta la disciplina que nos ocupa en este curso, la Econometría. La sección 3 describe un ejemplo de estudio econométrico, destacando cuáles son los elementos que integran un modelo econométrico. La sección 4 se ocupa de los datos económicos, sus características, las principales fuentes de obtención de datos y los programas informáticos que sirven para almacenar y procesar los datos. El software Gretl se introduce en el apartado 5, en el que se incluye el esquema de una primera sesión práctica de uso de Gretl. Los dos últimos apartados son un repaso a los conceptos de probabilidad e inferencia estadística que se aplicarán posteriormente, y que se acompaña de una sesión de práctica en Gretl.

1.2. ¿Qué es la Econometría?

En la toma de decisiones de carácter económico suele ser muy útil disponer de información en forma de datos cuantitativos. Por ejemplo, a la hora de elegir unos estudios universitarios podemos guiarnos por nuestras preferencias personales, pero también por factores como las expectativas de salario en la rama elegida o la facilidad con la que esperamos conseguir un empleo. Si se trata de la compra-venta de un piso, nos interesa conocer la situación del mercado inmobiliario. Para ello podemos recopilar datos de precios y de algunas características de los pisos que puedan influir en el precio como, por ejemplo, su tamaño o si es una vivienda usada que necesita reforma. Supongamos que en la sección de anuncios de un periódico local aparecen los siguientes datos sobre 50 pisos en venta en el centro de una ciudad:

- Precio del piso, en miles de euros.
- Tamaño del piso, en metros cuadrados hábiles.
- Estado del piso: si necesita reforma o está para entrar a vivir.

Indicador	Tamaño	Precio	A reformar	Indicador	Tamaño	Precio	A reformar
1	55	210,354	no	26	110	476,600	no
2	59	309,520	no	27	110	456,769	no
3	60	366,617	no	28	115	500,643	no
4	60	299,304	si	29	125	619,000	no
5	60	369,650	no	30	135	645,253	no
6	65	273,460	si	31	135	625,000	no
7	65	155,000	si	32	140	522,800	si
8	70	228,384	no	33	150	390,660	no
9	70	246,415	no	34	150	504,850	si
10	70	255,000	si	35	150	715,204	no
11	75	150,253	si	36	150	570,000	si
12	77	352,800	no	37	160	751,265	no
13	80	366,000	si	38	180	583,000	si
14	80	298,000	si	39	180	738,000	no
15	80	312,530	no	40	180	552,931	si
16	83	240,400	no	41	190	691,200	no
17	85	278,569	si	42	195	811,400	no
18	91	390,658	no	43	200	691,000	si
19	92	216,364	si	44	200	1110,000	no
20	100	402,600	no	45	230	961,620	no
21	100	272,300	si	46	230	661,000	no
22	100	360,607	no	47	240	841,417	no
23	100	570,000	no	48	240	588,992	si
24	100	480,809	no	49	245	841,400	si
25	100	186,314	si	50	250	1051,000	no

Tabla 1.1: Datos sobre precio de vivienda ocupada

Estos datos aparecen en la Tabla 1.1. En base a esta información, si nos ofrecen un piso de 100 m^2 reformado a un precio de 525000€, diríamos que el piso parece caro ya que su precio supera el promedio de precios de los pisos de estas características incluidos en la muestra:

$$\frac{402,6 + 360,607 + 570 + 480,809}{4} = 453,504 \text{ miles de euros}$$

Sin embargo, ¿qué podemos decir si se tratara de un piso de 90 m^2 a reformar? ¿O de un piso de 50 m^2 reformado? No tenemos datos para replicar el procedimiento anterior. Un económetra podría ayudar a dar respuesta a estas cuestiones. En el Gráfico 1.1, que representa conjuntamente el precio y el tamaño de cada piso, se ve un patrón o *relación estable* entre tamaño de un piso y su precio. Esta relación se puede trasladar a un *modelo* útil para responder a las preguntas que planteamos. Las técnicas econométricas nos permiten cuantificar, a partir del modelo y los datos, la influencia que tiene el tamaño del piso o su estado en el precio del mismo. La respuesta podría ser, por ejemplo: *La estimación del precio medio de un piso a reformar de 90 m^2 es de 297350 euros, aunque el precio puede oscilar entre 152711 y 441989 euros a un nivel de confianza del 90 %. Además, si se trata de un piso reformado, la estimación del precio medio se incrementa en más de 100000 euros, siendo factibles precios entre 210521 y 556639 euros.*

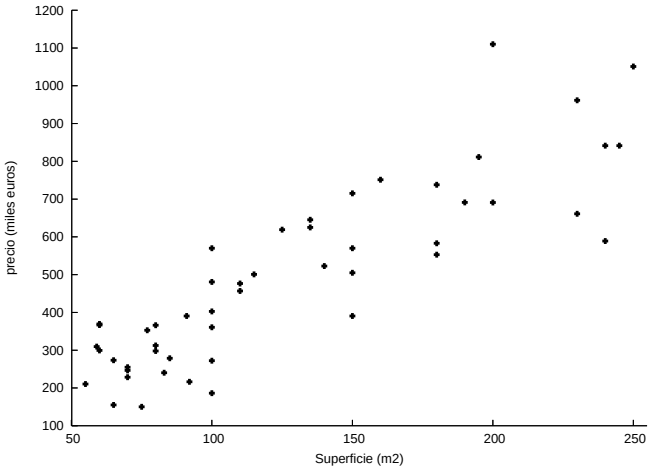


Gráfico 1.1: Diagrama de dispersión superficie-precio de pisos

La Econometría es una rama de la Economía que utiliza la estadística para medir o cuantificar las relaciones existentes entre variables económicas. Es una materia interdisciplinar que utiliza la teoría económica, la matemática, la estadística y los métodos computacionales. En palabras de Ramanathan (2002):

> *En términos sencillos, la* ***econometría*** *se ocupa de la aplicación de métodos estadísticos a la economía. A diferencia de la* estadística económica, *que es principalmente datos estadísticos, la econometría se distingue por la* unificación *de teoría económica, instrumentos matemáticos y metodología estadística. En términos más generales, la econometría se ocupa de (1) estimar relaciones económicas, (2) confrontar la teoría económica con los datos y contrastar hipótesis relativas al comportamiento económico, y (3) predecir el comportamiento de variables económicas.*

1.2.1. ¿Para qué sirve la Econometría?

El objetivo de un estudio econométrico es comprender mejor un fenómeno económico y, como resultado, poder realizar predicciones de la evolución futura del fenómeno de interés. El instrumento básico es el **modelo**, que ayuda a entender las relaciones entre variables económicas y sirve para evaluar los efectos de distintas medidas o políticas económicas. Algunos ejemplos en los que la Econometría es de utilidad son:

- Un analista del mercado de activos puede estar interesado en analizar y cuantificar la relación entre el precio de un activo y distintas características de la empresa que ofrece ese activo así como del estado general de la economía.
- Los directivos de Iberdrola pueden estar interesados en analizar los factores que afectan a la demanda de electricidad.
- El grupo Eroski puede estar interesado en cuantificar el efecto de distintos niveles de publicidad sobre sus ventas y sus beneficios.

- El servicio de estudios del Ministerio de Economía y del Banco de España o del Banco Central Europeo quiere analizar el impacto de las políticas monetarias y fiscales sobre el desempleo, la inflación, las exportaciones e importaciones, los tipos de interés, etc.
- Si un organismo quiere implementar políticas para corregir, por ejemplo, la discriminación salarial por sexo, en primer lugar debe conocer cuáles son los principales factores determinantes del problema y, en segundo lugar, analizar las posibles medidas a tomar, estudiando cuáles pueden ser los efectos de dichas medidas.
- Un gobierno regional puede necesitar previsiones sobre la evolución de la población para planificar la necesidad de servicios sociales y las necesidades de financiación que conllevan. También debe tener información precisa sobre su capacidad de financiación, por lo que le interesa disponer de predicciones relativas a la recaudación impositiva.
- Si una persona quiere contratar un préstamo, le interesa conocer cuál va a ser la evolución de los tipos de interés.

En los últimos años hemos asistido a una mayor difusión y utilización de los métodos econométricos gracias, entre otras razones, a la mayor disponibilidad y calidad de los datos y al desarrollo de los métodos de computación. Además, la aplicación de la Econometría no se restringe al ámbito estrictamente económico, sino que proporciona procedimientos de estudio de datos que pueden aplicarse al campo de las Ciencias Sociales. Por ejemplo, para:

- Analizar si el endurecimiento de las penas, como la introducción de la pena de muerte, tiene como consecuencia la disminución de la tasa de criminalidad.
- Analizar la efectividad de las medidas de seguridad vial, como el carnet por puntos, en la reducción del número de muertes en accidentes de tráfico.
- Predecir los resultados de una competición deportiva como, por ejemplo, el número de goles que marcará la selección de Inglaterra en un mundial de fútbol.
- Analizar cuál puede ser el efecto sobre los votantes en las próximas elecciones de una determinada medida, por ejemplo, prohibir fumar en lugares públicos, legalizar los matrimonios entre personas del mismo sexo, etc.
- Estudiar si hay diferencias en el voto dependiendo de si se trata de elecciones locales, regionales o europeas.
- Analizar si las medidas restrictivas sobre la publicidad de tabaco y alcohol reducen el consumo de estos productos.

Los comienzos de la Econometría pueden situarse en la década de los treinta del siglo pasado. Su coincidencia en el tiempo con la Gran Depresión no es casual: como consecuencia de ésta, los economistas de la época estaban interesados en poder predecir los ciclos económicos que observaban. Entre ellos destaca Keynes, que defendía la intervención del gobierno en la actividad económica para mitigar estas crisis. Así, los primeros económetras se ocuparon de dar respuesta a problemas macroeconómicos con objeto de asesorar a los gobiernos en la implantación de políticas económicas.

En un comienzo, se aplicaron a los datos económicos métodos estadísticos que ya habían sido utilizados en ciencias naturales. Sin embargo, estos métodos no podían reproducirse miméticamente en el ámbito económico, sino que había que adaptarlos o desarrollar nuevos métodos de acuerdo a las características propias que poseen las variables socioeconómicas. Así, en la econometría se han desarrollado dos grandes áreas: la *econometría teórica*, cuyo objetivo es desarrollar métodos de estudio y análisis de datos y determinar sus propiedades, y la *econometría aplicada*, que se ocupa de utilizar estos métodos para responder a los problemas de interés en la práctica. En este curso ponemos mayor énfasis en la parte aplicada. Se trata de proporcionar al alumno las herramientas necesarias para que sea capaz de llevar a cabo un proyecto aplicado. Para ello, es indispensable dedicar tiempo al conocimiento de los métodos e instrumentos básicos del análisis econométrico, ya que son el requisito previo para una buena aplicación práctica.

1.3. Un estudio econométrico

Uno de nuestros objetivos específicos es que, al final del curso, el estudiante debe ser capaz de estructurar y desarrollar un trabajo de investigación. Hoy día, una persona que disponga de un ordenador en su casa puede llevar a cabo un pequeño proyecto econométrico. Así, un estudio econométrico consta de las siguientes etapas, Heij, de Boer, Franses, Kloek & Dijk (2004):

- *Formulación del problema.* Se trata de determinar la cuestión de interés. Debemos plantear de forma precisa las preguntas que nos interesa responder. Por ejemplo, si se trata de conocer la situación del mercado inmobiliario en una ciudad, podemos plantearnos la siguiente pregunta: ¿cuál es el precio de los pisos en esa ciudad y qué factores lo determinan? 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 ejemplo anterior, es fácil recolectar datos sobre el precio de pisos, su tamaño y otras características que pueden influir en su precio (ver Tabla 1.1). 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.* De la unión de las teorías y cuestiones planteadas en la primera etapa con los datos se llega a un **modelo econométrico**. Por ejemplo, podemos plantear que, en media, el precio de un piso, Y, depende de su tamaño, X. Un posible modelo econométrico que recoge esta teoría es:

$$Y|X \sim N(\alpha + \beta X,\ \sigma^2)$$

  Es decir, el precio de los pisos dado un tamaño, por ejemplo 100 m^2, se distribuye alrededor de su media $\alpha + \beta 100$ según una normal de varianza σ^2. Al formular el

modelo hemos elegido la forma funcional de la relación entre las variables y la naturaleza estocástica de la variable de interés o endógena, Y. El objetivo es obtener un modelo relevante y útil para dar respuesta a nuestros objetivos.

El siguiente paso es la estimación de los parámetros desconocidos de la distribución y que son de interés para el análisis. En el ejemplo del precio de los pisos, interesan los parámetros de su media, α y β. 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. Por ejemplo, si es correcto asumir que el tamaño del piso influye en su precio, si la relación lineal entre ambas variables es correcta, etc. 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.

Dado que para la realización de un proyecto econométrico es necesario conocer dónde obtener los datos y manejar un software específico de análisis econométrico, vamos a extendernos un poco en estos dos puntos.

1.4. Los datos y su manejo

¿Cómo se obtienen datos económicos? No proceden de experimentos controlados sino que los economistas, al igual que otros investigadores del campo de las Ciencias Sociales, obtienen los datos de la observación de la realidad. En un experimento controlado, como los realizados en laboratorios, el investigador tiene control sobre las condiciones del estudio. Por ejemplo, para analizar el efecto de un fertilizante, podemos aplicar distintas dosis de fertilizante sobre un conjunto de sembrados, controlando también el grado de humedad o la luz que recibe cada planta. Además, se puede repetir el experimento, manteniendo las mismas condiciones o alterando algunas como las dosis o el grado de humedad. Obviamente, aunque las cantidades elegidas sean exactamente las mismas, no esperamos que el resultado, por ejemplo, el crecimiento de las plantas, sea idéntico entre experimentos porque las semillas utilizadas

son distintas o porque hay pequeños errores de medida. Estas diferencias naturales en los resultados de los experimentos se conocen como *variaciones muestrales*.

Los datos obtenidos de experimentos controlados son típicos de las Ciencias Naturales y se conocen como *datos experimentales*. Los datos que son resultado de un proceso que tiene lugar en la sociedad, y que no es controlable por una o varias personas, se conocen como *datos no experimentales*. Esta característica ha sido un factor importante en el desarrollo de las técnicas econométricas y debemos tenerlo en cuenta en la interpretación de los resultados.

Clasificación de los datos económicos. Los datos económicos pueden ser de diferentes tipos, lo que va a determinar el análisis que realicemos. Una primera clasificación distingue entre datos *cuantitativos*, aquéllos que toman valores numéricos dentro de un rango de valores, como precio o tamaño de un piso, y datos *cualitativos*, que aparecen como categorías o atributos, como por ejemplo el sexo, la profesión o el estado de un piso. Los seis primeros temas de este curso se centran en el análisis de datos cuantitativos. El tema siete considera situaciones en las que algún factor explicativo es cualitativo.

Una segunda clasificación distingue entre *datos de series temporales* y *datos de sección cruzada*. Los primeros se refieren a observaciones recogidas en sucesivos momentos de tiempo, normalmente regulares, como años, trimestres o meses. Ejemplos de datos temporales son el Producto Interior Bruto (PIB) de la Contabilidad Nacional trimestral, el número mensual de afiliaciones a la Seguridad Social o el valor diario del IBEX35. Los segundos se refieren a valores que toman diferentes agentes en un momento del tiempo, por ejemplo, la población desempleada en el año 2005 en cada uno de los países de la Unión Europea (UE), el salario medio en cada sector industrial en el 2006 o el gasto realizado en libros de texto por un conjunto de familias en septiembre pasado. También es posible tener una combinación de datos de sección cruzada y series temporales, por ejemplo, las puntuaciones obtenidas por los estudiantes de Econometría en los cursos 2004-05, 2005-06 y 2006-07. Cuando se encuesta a los mismos individuos a lo largo del tiempo, como la tasa de paro y el crecimiento del PIB desde 1990 hasta 2006 para los 25 países de la UE, se conocen con el nombre de *datos de panel* o *datos longitudinales*. En este curso nos centraremos en el análisis de datos de sección cruzada. Las técnicas que utilicemos también se pueden aplicar en series temporales, aunque en ocasiones su estudio es más complejo.

Una tercera clasificación se establece en función del nivel de agregación. Se conocen como *datos microeconómicos o microdatos* los referidos al comportamiento de agentes económicos como individuos, familias o empresas. Un ejemplo es la Encuesta de Población Activa, elaborada por el INE y publicada en *http://www.ine.es/prodyser/micro_epa.htm*. Los *datos macroeconómicos o macrodatos* son los datos referidos a ciudades, regiones o naciones que son resultantes de la agregación sobre agentes individuales, como son los resultados de la Contabilidad Nacional. Por ejemplo, la Contabilidad Nacional Trimestral de España, elaborada también por el INE y publicada en *http://www.ine.es/inebmenu/mnu_cuentas.htm*.

1.4.1. 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 working 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), que están accesibles como archivos de muestra en Gretl.

1.4.2. El software econométrico

El desarrollo de los ordenadores ha permitido almacenar una gran cantidad de datos, a la vez que ha facilitado su manejo. Existen en la actualidad un amplio conjunto de paquetes para el análisis econométrico que realizan complejas operaciones mediante unas instrucciones muy sencillas. Si los datos están disponibles en papel, las hojas de cálculo, como EXCEL, son un instrumento sencillo para introducir y preparar los datos y realizar operaciones sencillas. Sin embargo, en general es conveniente utilizar programas econométricos específicos. Algunos de los más populares en los cursos de Econometría son:

- **EViews**, desarrollado por Quantitative Micro Software, contiene una amplia gama de técnicas de análisis econométrico. Muchos manuales de Econometría contienen un CD con ejemplos prácticos en Eviews. Su página web con la información del programa es *http*: *//www.eviews.com*.

- **SHAZAM**, elaborado en la Universidad British of Columbia (Canadá), incluye técnicas para estimar muchos tipos de modelos econométricos. Más información se puede obtener en *http*: *//shazam.econ.ubc.ca*\, donde se puede ejecutar el programa remotamente.

- **Gretl**, acrónimo de *Gnu Regression, Econometric and Time Series* (Biblioteca Gnu de Regresión Econometría y Series Temporales), elaborado por Allin Cottrell (Universidad Wake Forest). Es software libre, muy fácil de utilizar. 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.

- **RATS**, acrónimo de *Regression Analysis of Time Series*. Contiene una amplia gama de técnicas de análisis econométrico con especial dedicación al Análisis de Series Temporales. Su web es: *http*: *//www.estima.com*

- **R**, software libre para cómputo estadístico y gráficos. Consiste en un lenguaje, un entorno de ejecución, un debugger y la habilidad de correr programas guardados en archivos de tipo script. Su diseño fue influenciado por dos lenguajes existentes: S y Scheme. Página web: *http*: *//www.r − project.org*

Un objetivo de este curso es que el estudiante se familiarice con el uso de programas econométricos. Por su sencillez y accesibilidad, en este curso introductorio se utiliza el programa Gretl para estudiar casos prácticos. En la página

http: *//gretl.sourceforge.net/gretl_espanol.html*

se encuentra toda la información en castellano relativa a la instalación y manejo del programa. El manual, en inglés, se encuentra en la carpeta *en/*.

Junto con el programa se pueden cargar los datos utilizados como ejemplos de aplicaciones econométricas en los siguientes libros de texto Davidson & Mackinnon (2004), Greene (2008), Gujarati (1997), Ramanathan (2002), Stock & 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. La unión de teoría y práctica permiten al alumno un autoaprendizaje tanto de los contenidos básicos del curso de Análisis de Regresión como de la utilización del software Gretl.

1.5. Introducción a Gretl

La primera sesión con el programa Gretl consiste en una práctica guiada en la que se aprenderá a crear un fichero, introducir los datos de la Tabla 1.1 y realizar un análisis descriptivo.

Preparación del fichero. Al ejecutar Gretl, aparece la siguiente ventana principal:

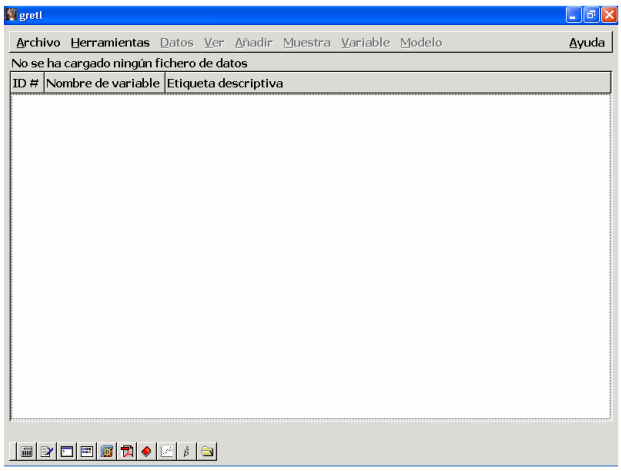


Gráfico 1.2: Pantalla inicial de Gretl

Como todavía no se ha cargado ningún fichero, varias opciones del menú principal, en gris claro, no están disponibles. Los datos a analizar no están incluidos en la base de Gretl, por lo que vamos a la opción *Archivo →Nuevo conjunto de datos Control+N*. Completamos la información que va solicitando el programa:

- *número de observaciones*, en la Tabla 1.1 se incluyen 50 pisos. Pinchar en *Aceptar*.
- El tipo de datos que utilizamos. En este caso, marcamos *de sección cruzada* y *Adelante*.
- Si el paso anterior se ha realizado correctamente, confirmamos la estructura del conjunto de datos pinchando en *Aceptar*. Al pinchar en *Atrás* se recupera sólo la ventana de tipo de datos, por lo que esta opción no permite corregir un error en el número de observaciones.
- En la última ventana marcaremos *Sí* queremos empezar a introducir los datos.
- En la siguiente ventana escribimos el *Nombre de la primera variable*, por ejemplo *m2*. No se pueden utilizar la letra *ñ*, acentos ni más de 15 caracteres para nombrar a las variables. Tras *Aceptar*, se abre una hoja de cálculo, de modo que en la pantalla aparece:

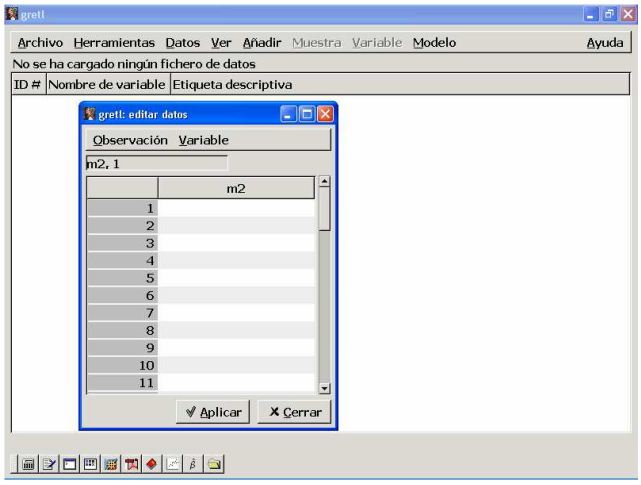


Gráfico 1.3: Añadir datos: hoja de cálculo de Gretl

Para incluir los datos de la variable $m2$, vamos a la celda correspondiente, por ejemplo la primera, y pinchamos sobre ella con la tecla izquierda del ratón; tras teclear la cifra, 55, damos a la tecla *Entrar*. Si por error no tecleamos algún dato, por ejemplo, la segunda observación de 59 m^2, nos situaremos en la fila posterior, en este caso en el primer dato de 60 m^2, y vamos a *observación* →*insertar obs*. Se crea una nueva fila en blanco por encima de la anterior. Para guardar las modificaciones en la sesión de trabajo hay que pinchar en *Aplicar*.

Podemos añadir más variables con la opción *Variable* →*Añadir* del menú de la hoja de cálculo. Por ejemplo, creamos una nueva variable que denominamos *Reforma*. Esta variable es cualitativa, por lo que asociamos a la situación *a reformar = sí* el valor 0 y a la otra opción, *a reformar = no* el valor 1. Una vez que se han incluido todos los datos, vamos a *Aplicar* y *Cerrar* la hoja de cálculo. Si no habíamos guardado los últimos cambios realizados, al cerrar la hoja de cálculo aparece un cuadro que nos pide confirmar los cambios. Las series creadas deben aparecer así en la pantalla:

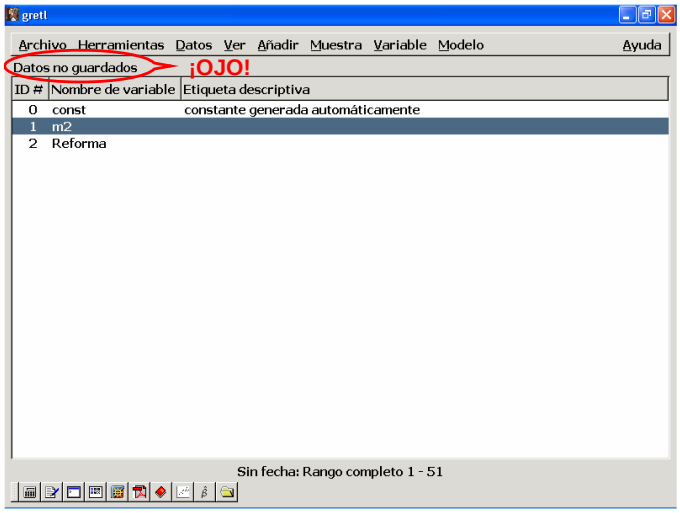


Gráfico 1.4: Fin de carga de datos con hoja de cálculo

Es recomendable guardar los datos ya incorporados en un fichero de datos Gretl mediante la opción del menú principal *Archivo* →*Guardar datos*. En el siguiente cuadro añadimos el directorio y el nombre del fichero de datos, por ejemplo, *pisos*. Por defecto, grabará los datos con la extensión *gdt*. Para usar estos datos en una sesión posterior, sólo hay que pinchar dos veces sobre el fichero.

Con frecuencia, los datos están almacenados en otra hoja de cálculo, como EXCEL. Por ejemplo, en el fichero EXCEL *pisos.xls* se encuentran las variables $m2$ y *precio* de la Tabla 1.1. Añadir los datos de *precio* al fichero de Gretl es muy sencillo. Una vez abierto el fichero *pisos.gdt*, hay que:

- Utilizar la opción del menú principal *Archivo* →*Añadir datos* →*EXCEL* ....
- Dar el nombre y ubicación del fichero EXCEL, *pisos.xls*.
- Dar la celda a partir de la cual hay que empezar a importar los datos. En este caso la variable *precio* empieza en la celda B1, donde está su nombre, e importaremos los datos desde *columna 2, fila 1*. Para añadir las dos variables, *m2* y *precio*, comenzaríamos a importar datos en *columna 1, fila 1*. Finalmente, hay que pinchar en *Aceptar*.

Para comprobar si no hay errores en los datos vamos a *Datos →seleccionar todos* y luego activamos la hoja de cálculo mediante *Datos →Editar valores* o bien mostramos los datos en pantalla con *Datos →Mostrar valores →Todas las variables*. Debe aparecer la siguiente ventana:

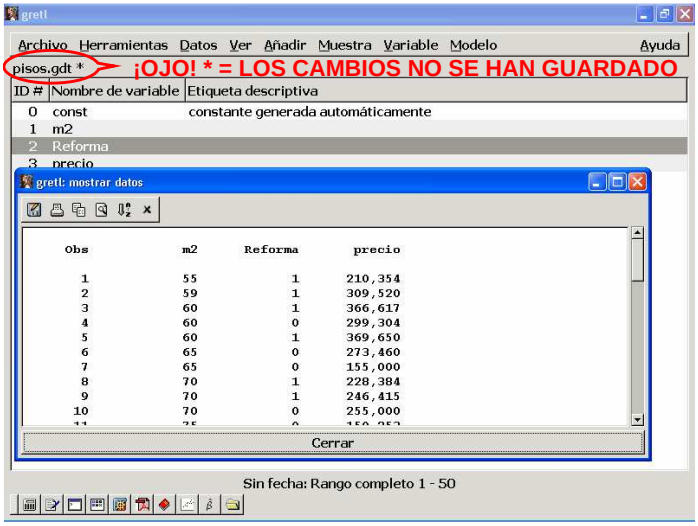


Gráfico 1.5: Fichero con datos de tres variables

Una vez que los datos se han cargado correctamente, los almacenamos en el mismo fichero *pisos.gdt* pinchando en *Archivo →Guardar datos*. Una vez guardadas las modificaciones, en la pantalla de Gretl aparece el nombre del fichero sin el asterisco *.

Notas explicativas. Al crear un fichero, nos interesa incluir notas explicativas del trabajo ya realizado. En Gretl es posible añadir esta información en dos apartados, uno general y otro específico de cada variable. Es posible añadir una breve descripción de cada variable y que aparezca como *etiqueta descriptiva* junto con el nombre de la variable. Por ejemplo, añadiremos la nota informativa sobre la interpretación de la variable *Reforma*:

Valor 0 si el piso está para reformar, valor 1 si está reformado

Marcamos con el ratón la variable y vamos a *Variable→editar atributos*. El cuadro siguiente en el apartado *descripción* escribimos el texto y pinchamos en *Aceptar* (ver Gráfico 1.6).

Las etiquetas descriptivas son útiles para saber la fuente de datos o las unidades de medida. Por ejemplo, para la variable *precio* y *m2* añadiremos las siguientes etiquetas descriptivas:

Variable	Etiqueta descriptiva	Nombre a mostrar en gráficos
precio	Precio de pisos en miles de euros	Precio (miles euros)
m2	Tamaño de pisos en metros cuadrados	Superficie (m2)

La opción *Datos →Editar información* da lugar a un cuadro que permite añadir texto informativo, por ejemplo,

Datos utilizados en el tema 1 de Análisis de regresión con Gretl

Finalmente, la opción *Datos →Ver descripción* permite visualizar la información de la estructura del conjunto de datos junto con las notas explicativas añadidas. Si todo el proceso se ha realizado correctamente, en pantalla debe aparecer el siguiente cuadro:

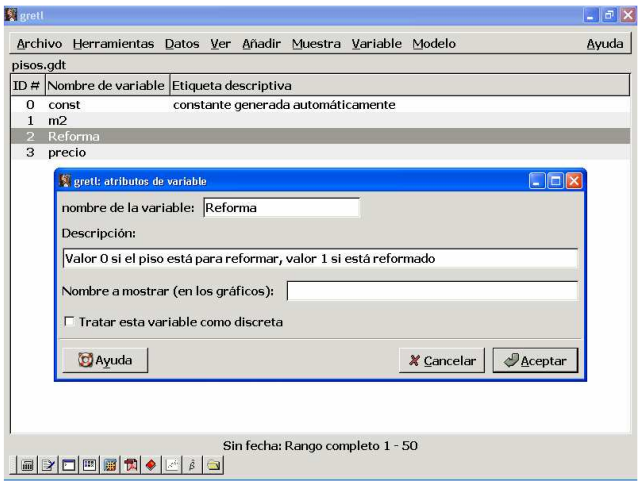


Gráfico 1.6: Cuadro de descripción de variables

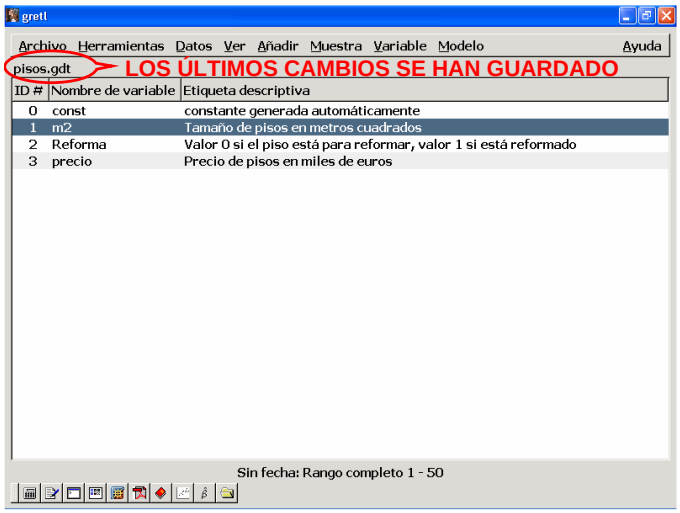


Gráfico 1.7: Fichero con descripción de variables

1.5.1. Análisis descriptivo de una variable

Una vez incorporados los datos, vamos a obtener una visión general de los mismos. El objetivo del análisis descriptivo es resumir un conjunto de datos, extrayendo las características e información más relevante para el estudio. En primer lugar, sintetizaremos la información de cada una de las variables y en una segunda etapa, obtendremos una primera idea de las relaciones existentes entre las variables. Para ello se utilizan *gráficos* y números-resumen conocidos como *estadísticos descriptivos*[1]. El análisis descriptivo de una única variable que proporciona Gretl se encuentra en la opción *variable* del menú principal; un resumen de este análisis se obtiene en el *menú auxiliar* que aparece al pinchar con la tecla derecha del ratón sobre la variable.

El gráfico más utilizado para resumir datos de sección cruzada de una única variable económica es el **histograma**, que aparece con la opción del menú auxiliar *Gráfico de frecuencias*. Se trata de un diagrama de barras que en el eje horizontal o abscisa representa los va-

[1]Este apartado es un resumen de los conceptos mínimos relevantes. Explicaciones más detalladas se encuentran en manuales como Peña & Romo (1997).

lores de la variable divididos en intervalos. Sobre cada intervalo se dibuja una barra, cuya superficie refleja el número de observaciones que pertenecen a dicho intervalo. Si, por ejemplo, pinchamos con la tecla derecha del ratón sobre la variable *precios* y vamos a *Gráfico de frecuencias*, aparece el cuadro de opciones del histograma en la que fijamos:

- *Número de intervalos*: Por defecto aparecen 7 intervalos, que es un número entero próximo a $\sqrt{N}$, siendo N el número de observaciones, en este caso 50.
- *Valor mínimo intervalo izquierdo* y *grosor del intervalo*: todos los intervalos deben tener la misma amplitud. Por defecto, se eligen de manera que el punto central o marca de clase de los intervalos primero y último sean, respectivamente, los valores mínimo y máximo que toma la variable en el conjunto de datos.

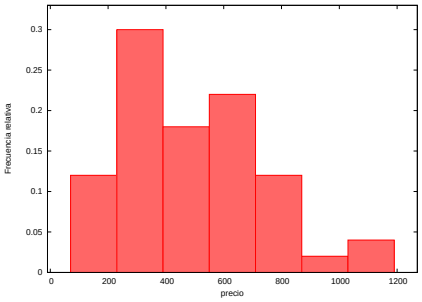


Gráfico 1.8: Histograma de frecuencias relativas

Usando las opciones estándar de Gretl obtenemos el Gráfico 1.8. Si pinchamos sobre el gráfico, se despliega un menú auxiliar que permite hacer cambios en el gráfico (*editar*) o guardarlo en diversos formatos (portapapeles, postcript, etc). La opción *guardar a sesión como icono* guarda el gráfico a lo largo de la sesión de Gretl. Es decir, una vez cerrada la ventana del gráfico, se recupera pinchando en el cuarto símbolo de la *barra de herramientas* situada en parte inferior derecha de la ventana principal (*vista iconos de sesión*) y, a continuación, pinchando dos veces en el icono *gráfico 1*.

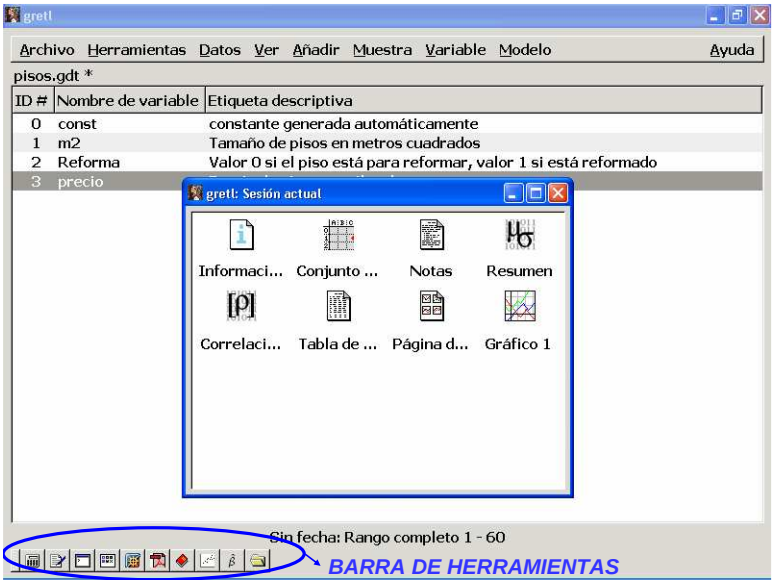


Gráfico 1.9: Iconos de la sesión

Para ver la tabla con la distribución de frecuencias representada en el histograma, hay que marcar la variable correspondiente e ir a la opción *Variable →Distribución de frecuencias*. Por ejemplo, la tabla de distribución de frecuencias de la variable *precio* es:

```
Distribución de frecuencias para precio, observaciones 1-50 número
de cajas = 7, media = 489,858, desv.típ.=237,416

      intervalo     punto medio   frecuencia  rel     acum.
          <  230,23     150,25         6     12,00%   12,00% ****
    230,23 - 390,19     310,21        15     30,00%   42,00% **********
    390,19 - 550,15     470,17         9     18,00%   60,00% ******
    550,15 - 710,11     630,13        11     22,00%   82,00% *******
    710,11 - 870,06     790,08         6     12,00%   94,00% ****
    870,06 - 1030,0     950,04         1      2,00%   96,00%
          >= 1030,0     1110,0         2      4,00%  100,00% *
```

Tabla 1.2: Distribución de frecuencias del precio de 50 pisos

En la primera columna aparecen los intervalos en que se han dividido los valores que toma la variable *precio* y la segunda incluye el punto medio o **marca de clase** del intervalo. La columna *frecuencia* es lo que se conoce como **frecuencia absoluta** de un intervalo, es decir, el número de pisos con precio en ese intervalo. Por ejemplo, en la Tabla 1.1 hay 15 pisos cuyo precio se encuentra entre 230232€ y 390190€. La columna, *rel*, contiene la **frecuencia relativa** de cada intervalo, es decir, la fracción de observaciones que hay en cada tramo. Con estas frecuencias se ha construido el histograma anterior. Por ejemplo, los 15 pisos con precio en el intervalo [230,232; 390,190) constituyen el 30 % del total de los 50 pisos. Y, como todos los intervalos son de igual amplitud, la altura de la segunda barra del histograma es la frecuencia relativa asociada en tanto por uno, es decir, 0,3. Si a la frecuencia relativa de un intervalo se le suman las frecuencias relativas de los anteriores se obtiene la **frecuencia relativa acumulada** hasta cada intervalo, que aparece en la columna *acum.* Por ejemplo, en el conjunto de pisos que estudiamos, un 42 % de ellos tiene un precio inferior a 390190€.

La descripción numérica de una variable se encuentra en la opción del mismo menú auxiliar *Estadísticos descriptivos* o en el menú principal, *Variable →Estadísticos principales*. El resultado para la variable *precio* es la Tabla 1.3:

```
Estadísticos principales, usando las observaciones 1 - 50
  para la variable 'precio' (50 observaciones válidas)

Media             489,86      Desviación típica     237,42
Mediana           466,68      C.V.                    0,48466
Mínimo            150,25      Asimetría               0,68052
Máximo           1110,0       Exc. de curtosis       -0,19251
```

Tabla 1.3: Estadísticos descriptivos del precio de 50 pisos

Esta ventana tiene un nuevo menú. La opción *Copiar* permite importar la tabla a un fichero MS Word, Latex o simplemente, como aparece en pantalla (*Texto plano*). Estos estadísticos

descriptivos reflejan algunas características de la distribución recogidas en el histograma. La media y la mediana son medidas de posición, la desviación típica y el coeficiente de variación son medidas de dispersión, mientras que la asimetría y exceso de curtosis son medidas de forma de la distribución.

Las **medidas de posición** dan una idea de la situación o centro del conjunto de puntos. La *media* es el valor promedio. Si disponemos de N datos de una variable $x_1, x_2, \ldots, x_N$, la media, o también momento muestral de primer orden, se define como:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

La media es un estadístico poco robusto frente a la presencia de valores extremos: observaciones anómalas van a tener una gran influencia en el valor que tome. Por ejemplo, si el piso número 50 tuviera un precio *muy alto*, por ejemplo, 1350 miles de euros en lugar de 1051, entonces el precio medio aumentaría en casi 6000 euros, situándose en 495,84 miles de euros.

En general, interesan estadísticos cuyo valor no varíe mucho ante cambios en los valores de unas pocas observaciones, por muy grandes que sean esas variaciones. La *mediana*, que es el valor central de la distribución, posee esta propiedad. Así, la mediana del *precio* es $466,68$ miles de euros.

Las medidas de posición proporcionan un valor representativo del conjunto de datos que debe complementarse con una medida del error asociado. Para valorar la representatividad de este único valor se utilizan las **medidas de dispersión**, que informan de si las observaciones están poco concentradas (o muy dispersas) alrededor de su centro. Una medida sencilla es la diferencia entre los valores máximo y mínimo que toman los datos en la muestra, lo que se conoce como *recorrido*. Es decir,

$$\text{Recorrido} \quad = \quad \text{Máximo} \quad - \text{ Mínimo}$$

En el ejemplo, tenemos que el recorrido de los precios es 1110-150,25 = 959,75 miles de euros. Esta medida sólo tiene en cuenta dos valores, los extremos. Otras medidas se elaboran con todos los datos, por ejemplo, la desviación típica, que es la raíz cuadrada positiva de la varianza. La varianza de un conjunto de datos se define como un promedio de los cuadrados de las desviaciones de los datos a la media. Gretl calcula la varianza, S^{*2} o S_x^{*2}, como:

$$S_x^{*2} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Por tanto, la *desviación típica*, S_x^*, se calcula según:

$$S_x^* = +\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Varianza y desviación típica son medidas de la dispersión de los datos alrededor de la media. Tiene el valor mínimo cero cuando todos los datos de la variable toman el mismo valor. La ventaja de la desviación típica es que tiene las mismas unidades de medida que la variable original. En general, cuanto más próxima a cero esté S_x^*, más concentrados estarán los datos

alrededor de la media y ésta será más representativa del conjunto de observaciones. Sin embargo, al depender S_x^* de las unidades de medida, no es fácil comparar su representatividad en dos conjuntos de datos. Para solucionar este problema se utiliza el *coeficiente de variación*, *C.V.*, que es una medida adimensional de la dispersión, y se define como:

$$C.V. = \frac{S_x^*}{|\bar{x}|} \qquad \text{si } \bar{x} \neq 0$$

En el ejemplo de precios tenemos que $C.V. = 0,485 < 1$, la dispersión de los datos es pequeña en relación a su nivel, por lo que consideramos que la media sí es bastante representativa del conjunto de datos.

Media y desviación típica son los estadísticos-resumen más conocidos. Se acompañan de las **medidas de forma**, que reflejan otras características del histograma. La asimetría de una distribución se refiere a si los datos se distribuyen de forma simétrica alrededor de la media o no. El *coeficiente de asimetría* se define como:

$$\text{Coeficiente de asimetría } = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{S_x}\right)^3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{S_x^3}$$

con $S_x = \sqrt{(N-1)/N} \times S_x^* = \sqrt{\sum_i (x_i - \bar{x})^2/N}$. El coeficiente de asimetría es cero cuando los datos se distribuyen simétricamente alrededor de la media, es positivo cuando la cola derecha (asociada a valores por encima de la media) es más larga que la izquierda siendo negativa en caso contrario. En el ejemplo de los precios de los pisos, observamos que la asimetría es positiva, lo que se corresponde con una media mayor que la mediana, es decir, $\bar{x} > Mediana(X)$.

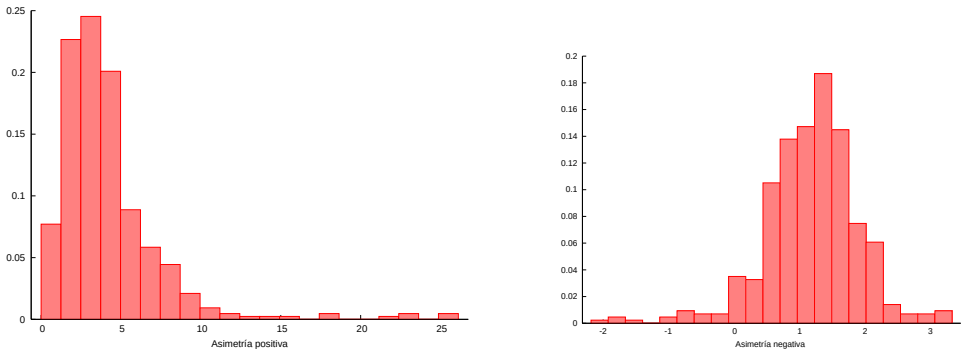


Gráfico 1.10: Tipos de asimetría

El coeficiente de curtosis es una medida del apuntamiento de la distribución y se define:

$$\text{Curtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{S_x}\right)^4 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{S_x^4}$$

Este coeficiente mide la cantidad de observaciones que se encuentran en las colas en relación con las situadas alrededor de la media. El nivel de referencia es tres, que es el valor de la

curtosis de la distribución *normal*. Así, se define el *exceso de curtosis* como:

$$\text{Exc. de curtosis } = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{S_x^4} - 3 \tag{1.1}$$

Un exceso de curtosis positivo indica mayor peso de observaciones en la cola y mayor apuntamiento que la distribución normal, mientras que si es negativo indica menor número de observaciones en la cola y menor apuntamiento.

Cuando tenemos un conjunto de variables, Gretl permite recoger en una única tabla los estadísticos descriptivos de todas las variables. El proceso es el siguiente:

1. Seleccionar las variables de interés pinchando simultáneamente la tecla izquierda del ratón y la tecla *Control*.
2. Ir a *Ver* →*Estadísticos principales* o utilizar *Estadísticos descriptivos* en el menú auxiliar que aparece al pinchar la tecla derecha del ratón sobre las variables seleccionadas.

Así, con los datos de la Tabla 1.1 se obtiene la siguiente tabla de estadísticos descriptivos:

Estadísticos principales, usando las observaciones 1 - 50

Variable	MEDIA	MEDIANA	MIN	MAX
m2	127,34	105,00	55,000	250,00
Reforma	0,62000	1,0000	0,00000	1,0000
precio	489,86	466,68	150,25	1110,0

Variable	D.T.	C.V.	ASIMETRÍA	EXC.CURTOSIS
m2	59,048	0,46370	0,67091	-0,77954
Reforma	0,49031	0,79083	-0,49445	-1,7555
precio	237,42	0,48466	0,68052	-0,19251

Tabla 1.4: Estadísticos descriptivos del conjunto de datos

donde D.T. indica desviación típica, MIN es mínimo y MAX denota el máximo. Al interpretar estos resultados, hay que tener en cuenta que la variable *Reforma* no es una variable cuantitativa continua, sino una variable cualitativa discreta, que sólo toma valores 1 ó 0.

1.5.2. Relaciones entre variables

Cuando el conjunto de datos contiene, por ejemplo, dos variables cuantitativas nos interesa estudiar la relación o asociación que existe entre ellas. En general, al analizar dos (o más) variables, podemos establecer una relación de causalidad entre ellas. Por ejemplo, podemos pensar que el precio de un piso puede ser consecuencia del tamaño de la vivienda, pero no al revés. Se llama variable independiente o exógena, x, a la que causa el efecto y variable dependiente o endógena, y, a la que lo recibe. La relación entre estas variables puede estudiarse con gráficos o expresarse numéricamente mediante, por ejemplo, el coeficiente de correlación. Todos estos elementos del análisis descriptivo de un conjunto de variables se realiza con el menú que se despliega en la opción *Ver* de Gretl.

Representación gráfica. El diagrama de dispersión o *scatterplot* da una primera idea de la relación entre dos variables. Es el gráfico que representa cada punto (x_i, y_i), $i = 1, \ldots N$ en el plano: la variable x aparece en el eje de abscisas y la variable y en el eje de ordenadas. Por ejemplo, para obtener con Gretl el Gráfico 1.11, precio sobre superficie, podemos seguir uno de los siguientes pasos:

- *Ver* →*Gráficos* →*Gráfico X-Y (scatter)* y en el cuadro *Definir el gráfico* marcar:
  Variable de eje X Elegir −> *m2*
  Variables de eje Y Añadir −> *precio*
- O bien seleccionar las variables *precio* y *m2* pinchando simultáneamente la tecla izquierda del ratón y la tecla *Control* e ir al menú auxiliar, *Gráfico de dos variables XY*. En el siguiente cuadro, se selecciona la variable de la abscisa, $m2$.

Al pinchar en *Aceptar* aparece el Gráfico 1.11 que, además de la nube de puntos, incluye una recta-síntesis de la relación, la recta de regresión mínimo cuadrática que veremos más adelante.

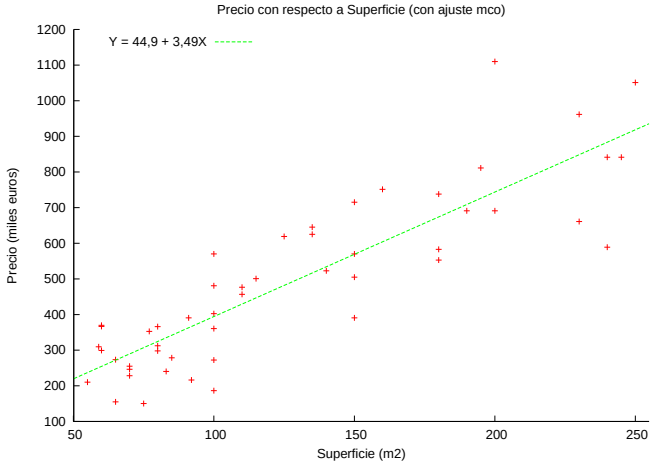


Gráfico 1.11: Diagrama de dispersión superficie-precios (2)

Al pinchar sobre el gráfico aparece un menú auxiliar que sirve para:

- Exportar el gráfico a ficheros en diferentes formatos en *Guardar como Windows metafile (EMF)...*, *PNG...*, *postscript (EPS)...*, *PDF...*.
- Copiar/exportar el gráfico a otros ficheros con *Copiar al portapapeles*.
- Guardar el fichero en la sesión de Gretl en *Guardar la sesión como icono*.
- Realizar cambios en el fichero con *Editar*. En la pestaña *Principal* se controla el título del gráfico, el tamaño y tipo de letra, el color de las líneas/puntos, el dibujo del marco completo, la situación de texto explicativo de las variables representadas (*posición de la clave*) o la eliminación de la recta-resumen. La escala y la explicación de los ejes se modifica en *Eje X* y *Eje Y*. En *líneas* se controla la representación de los datos, tipo de línea o punto, y el texto explicativo de las variables. *Etiquetas* permite añadir texto en el gráfico y *salida a fichero* incluye varios formatos para guardar el gráfico.

El gráfico de dispersión permite distinguir la posible relación, lineal o no, que existe entre las variables. Se dice que hay una **relación lineal positiva** entre ambas variables cuando al aumentar x, aumenta en promedio el valor de y (figura b en el Gráfico 1.12). Diremos que hay una **relación lineal negativa** entre ambas variables cuando observamos que al aumentar x, disminuye en promedio el valor de y (figura c). En el ejemplo, se observa una clara relación lineal positiva entre precio y tamaño del piso.

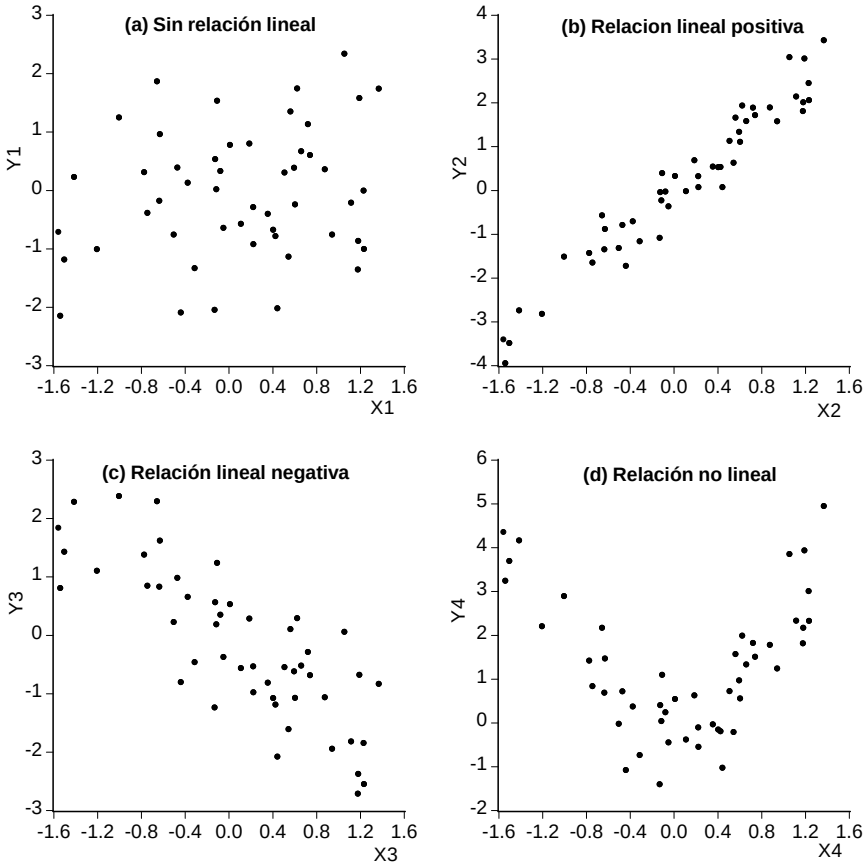


Gráfico 1.12: Diagramas de dispersión

Covarianza y correlación. La **covarianza** es una medida del grado de asociación lineal entre dos variables. Si se tienen N pares de datos de dos variables, $(x_1, y_1) \ldots (x_N, y_N)$, la covarianza se denota por S_{xy} y se define:

$$S_{xy} = cov(x, y) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

siendo $\bar{x}$ e $\bar{y}$ las medias aritméticas de las variables. La covarianza depende de las unidades de medida de las variables, lo que no permite comparar la relación entre distintos pares de variables medidas en unidades diferentes. En estos casos se utiliza el **coeficiente de correlación lineal** entre x e y, que se define:

$$r_{xy} = corr(x, y) \quad = \quad \frac{S_{xy}}{S_x \, S_y} \quad = \quad \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

El coeficiente de correlación lineal y la covarianza tienen el mismo signo: son positivos si existe relación lineal directa o positiva (figura b en el Gráfico 1.12), son negativos si existe relación lineal inversa o negativa (figura c) y toma valor cero si x e y son independientes (figura a) o cuando la relación, si existe, es no lineal (figura d). Además, su valor no depende del orden en que se consideren las variables, es decir, $S_{xy} = S_{yx}$ y $r_{xy} = r_{yx}$. A diferencia de la covarianza, el coeficiente de correlación es una medida adimensional de la relación que toma valores entre -1 y 1, $-1 \leq r_{xy} \leq 1$: un coeficiente de correlación igual a uno en valor absoluto indica que las variables están relacionadas linealmente de forma exacta y los datos se sitúan sobre una línea.

En Gretl, si se marcan las variables que interesan y se va a *Ver →Matriz de correlación* se obtiene una tabla (matriz) con los coeficientes de correlación para cada par de variables consideradas. El resultado para los datos de precios, tamaño y reforma de los pisos es:

```
Coeficientes de correlación, usando las observaciones 1 - 50
   valor crítico al 5% (a dos colas) = 0,2787 para n = 50
    m2       Reforma        precio
1,0000        0,0440        0,8690  m2
              1,0000        0,2983  Reforma
                            1,0000  precio
```

Tabla 1.5: Matriz de coeficientes de correlación

Por ejemplo, el coeficiente de correlación entre el precio y el tamaño de los pisos se encuentra en la primera fila, columna tercera, (precio-m2). Es decir, $r_{precio,m2} = 0,869$, lo que indica que hay una fuerte relación lineal positiva entre estas variables. Hay que tener en cuenta que este coeficiente se define para variables cuantitativas, por lo que no lo aplicamos a la variable *Reforma*.

Bibliografía

Davidson, D. y J. Mackinnon (2004), *Econometric Theory and Methods*, Oxford University Press.

Greene, W. (2008), *Econometric Analysis*, 6ª edn., Prentice-Hall.

Gujarati, D. (1997), *Econometría básica*, 4ª edn., McGraw-Hill.

Heij, C., de Boer, P., Frances, P., Kloek, T. y H. Van Dijk (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press.

Peña, D. y J. Romo (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.

Stock, J. y M. Watson (2003), *Introduction to Econometrics*, Addison-Wesley.

Verbeek, M. (2004), *A Guide to Modern Econometrics*, 2ª edn., John Wiley.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2ª edn., South-Western.

Tema 2

Modelo de Regresión Lineal Simple

Contenido

2.1. Introducción. Un ejemplo

Supongamos que nos interesa conocer la relación que hay entre el precio de una vivienda y determinadas características de la misma. Empezaremos considerando el caso más sencillo, una única característica, la 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*:

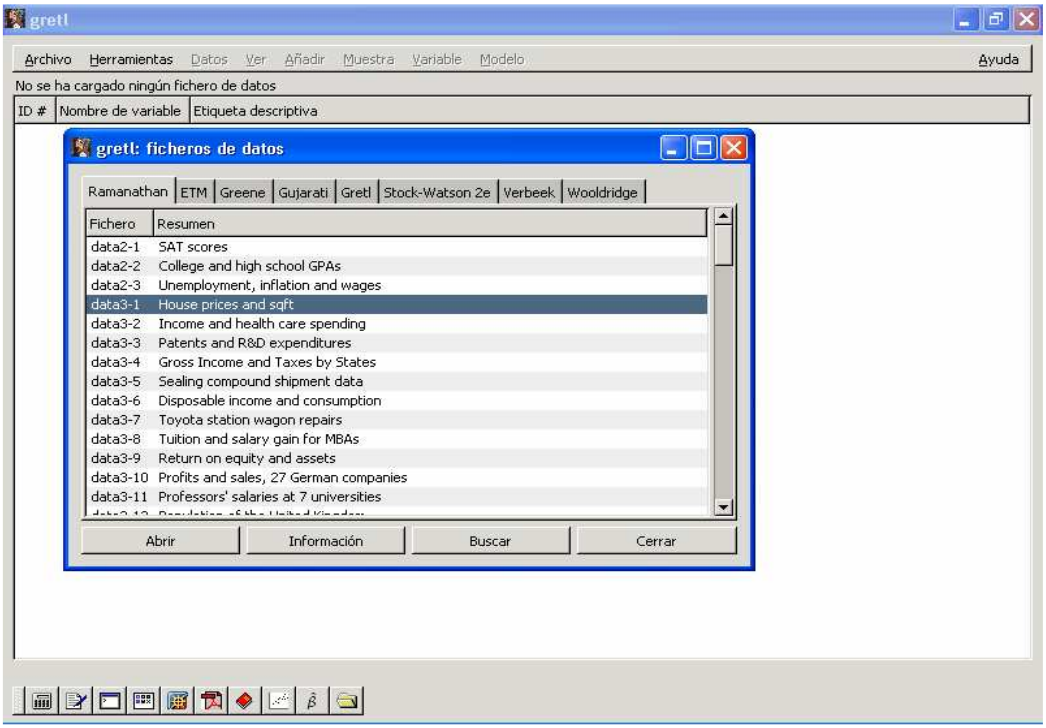


Gráfico 2.1: Selección de un fichero de muestra

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)
```

i	P_i	F2	i	P	F2
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*

3. Seguidamente en *Variable →Editar atributos* cambiamos los nombres a las variables (*P* y *F2*), la descripción (*Precio de venta en miles de dólares* y *Pies cuadrados hábiles*) y el nombre a mostrar (*Precio, P* y *Superficie, F2*)

4. Guardamos los cambios en un fichero llamado *datos-cap3.gdt* con *Archivo →Guardar datos*.

5. Abrimos el diagrama de dispersión entre las dos variables (ver el Gráfico 2.2). En él observamos una relación lineal positiva entre P y $F2$.

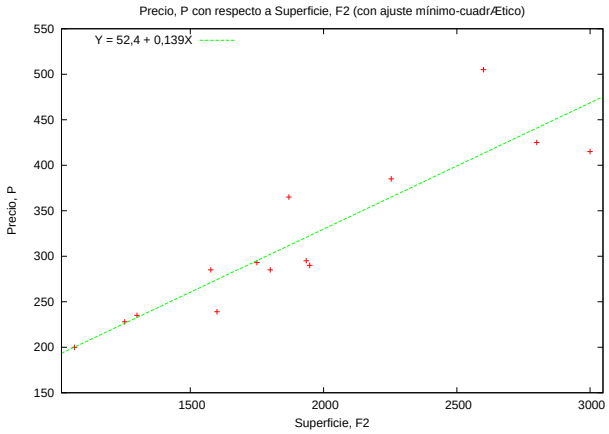


Gráfico 2.2: Diagrama de dispersión precio-superficie de viviendas

Un modelo sencillo que recoge una relación lineal causa-efecto entre superficie y precio es $P_i = \alpha + \beta F2_i$. 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. Debemos, por tanto, especificar un modelo econométrico que recoge esta característica: el modelo de regresión lineal simple.

2.2. Elementos del modelo de regresión simple

El modelo simple relaciona dos variables de forma lineal,

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

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 α y la pendiente β del modelo son los ***coeficientes de la regresión***. 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[1].

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:

$$P_i = \alpha + \beta\ F2_i + u_i \qquad i = 1, \ldots, N \tag{2.2}$$

donde

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

[1] En este capítulo y los siguientes, por simplicidad, no reservaremos la letra mayúscula para variables aleatorias X y las minúsculas para realizaciones (x) sino que utilizaremos mayúsculas tanto para una variable aleatoria como como para su realización, es decir, para los datos.

- $F2_i$ es la observación i de la variable independiente (exógena o explicativa) *área habitable* en pies cuadrados.
- Los dos coeficientes a estimar son α y β, y sospechamos que al menos β 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 α y β, que son los parámetros de la relación entre P y $F2$. 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.

2.3. Hipótesis básicas

El modelo (2.1) 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.

• *Sobre la forma funcional*

1. El modelo es lineal en los coeficientes. Los modelos a estimar a lo largo del curso son lineales en los coeficientes, $Y_i = \alpha + \beta X_i + u_i$. Sin embargo, podemos permitir no linealidades en las variables explicativas como puede ser la especificación:

$$P_i = \alpha + \beta \left(F2_i\right)^2 + u_i$$

en la que la superficie habitable de los pisos no influye de forma lineal sobre el precio, sino de forma cuadrática.

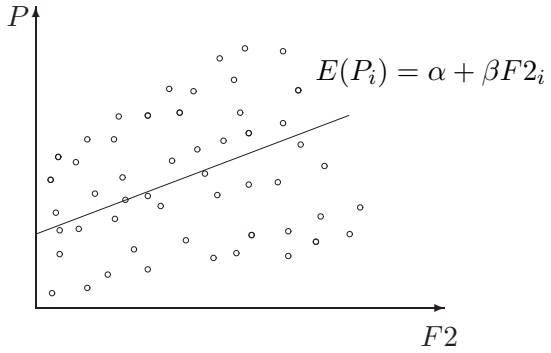
• *Sobre los coeficientes*

2. Los coeficientes α y β se mantienen constantes a lo largo de la muestra. Vamos a considerar que la influencia de las variables explicativas es estable a lo largo de la muestra. Supongamos que estamos interesados en analizar, en términos medios, el precio de los

pisos de Bilbao (P) en función de la superficie habitable en metros cuadrados ($F2$). En este caso interesaría estimar la *recta central* representada en el caso 1 del Gráfico 2.3.

No obstante, supongamos que algunos de estos pisos están localizados en el centro de Bilbao (representados en azul) y que otros están localizados en la periferia (en rojo). El caso 2 del Gráfico 2.3 muestra esta hipotética situación: en general, para una determinada superficie, los pisos del centro tienen mayor precio. Así, en el gráfico es posible distinguir dos nubes de puntos, cada una asociada a pisos de una determinada zona. Si este fuera el caso, estaríamos dispuestos a creer que existen (y debemos estimar) *dos rectas centrales* (la azul y la roja) permitiendo que tanto la ordenada como la pendiente cambien a lo largo de la muestra, dependiendo de la zona en la que se localice el piso.

Caso 1: Sin discriminar por localización



Caso 2: Discriminando por localización

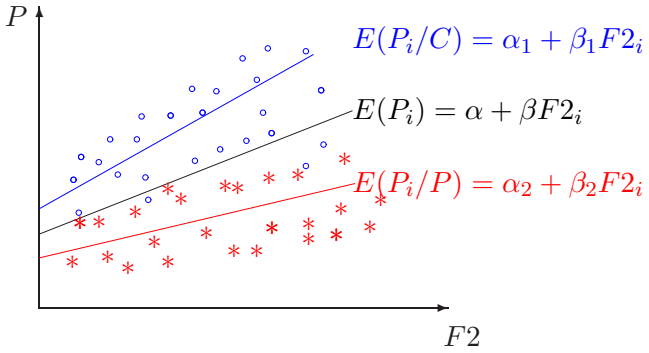


Gráfico 2.3: Precio pisos de Bilbao *vesus* superficie habitable

- *Sobre la variable endógena*

3. La variable endógena es cuantitativa. A lo largo de este curso básico vamos a suponer que la variable a explicar es cuantitativa. Lo contrario, una variable endógena cualitativa, requiere métodos de estimación alternativos al método que se analiza en este curso.

- *Sobre la variable explicativa*

4. La variable explicativa X tiene varianza muestral S_X^2 no nula y además $N \geq K = 2$. Estas hipótesis son necesarias para poder identificar los coeficientes (ordenada y pendiente). En primer lugar, si el número de coeficientes a estimar fuera mayor que el número de observaciones disponibles en la muestra, no tenemos suficiente información para poder llevar a cabo la estimación. Más adelante veremos que esta condición debe hacerse más estricta, $N > 2$, si además de estimar los dos parámetros α y β que determinan el valor medio de Y, nos interesa estimar su variabilidad.

   Por otra parte, si la variable explicativa tuviera varianza muestral nula ($S_X^2 = 0$), es decir, si la variable explicativa tomase un valor constante, por ejemplo, $X_i = 5\ \forall i$, la pendiente y la ordenada no podrían ser identificadas. Esto se debe a que la variable X es una combinación lineal del término constante, $X = 5 \times$ término constante $= 5 \times 1 =$

5. De hecho, tal y como se puede observar en el Gráfico 2.4, una situación de estas características no puede explicar las variaciones de la variable de interés Y.

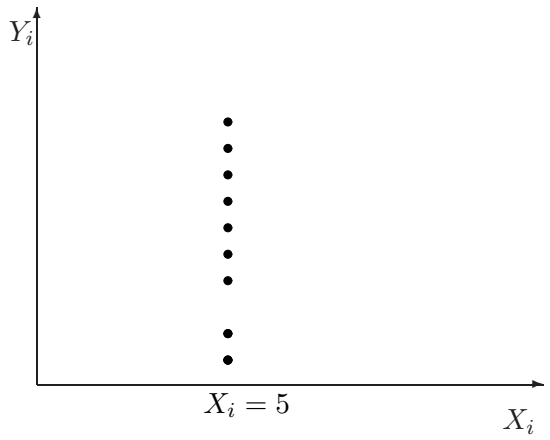


Gráfico 2.4: Modelo $Y_i = \alpha + \beta \times 5 + u_i$, con $S_X^2 = 0$

5. La variable exógena X es fija, no aleatoria. Las observaciones del regresor $X_1, \ldots X_N$ son valores fijos en muestras repetidas, es decir, suponemos que trabajamos en un contexto de experimento controlado. Esta condición implica que la variable explicativa X no podrá estar medida con error. En el caso práctico que estamos considerando, esto significa que los metros cuadrados habitables están medidos con exactitud. En muchos casos es un supuesto poco realista, pero lo utilizamos como punto de partida. El contexto en el que la variable explicativa X tiene carácter aleatorio se estudia en textos más avanzados, por ejemplo, Wooldridge (2003) o Alonso, Fernández & Gallastegui (2005).

6. El modelo está bien especificado. En general, esta hipótesis requiere que en el modelo no se incluyan variables irrelevantes ni que se omitan variables relevantes para explicar Y. En el contexto del modelo de regresión simple, esto significa que la variable explicativa X es la única variable relevante para explicar y predecir la variable de interés Y.

• *Sobre la perturbación*

El término de error recoge aquellos elementos que afectan a la variable de interés y que no observamos. Podemos hacer conjeturas sobre los valores que puede tomar, cuáles son más probables y cuáles menos. Así, consideramos que u_i es aleatorio y tiene las siguientes propiedades.

7. La perturbación tiene media cero. El error impredecible, la parte aleatoria del modelo, tiene media cero. Esto implica que la parte sistemática del modelo $(\alpha + \beta X_i)$ puede ser interpretada como el comportamiento medio a analizar, es decir, $E(Y_i) = \alpha + \beta X_i$.

8. La perturbación tiene varianza constante. Suponemos que la variabilidad del error se mantiene constante, $var(u_i) = \sigma^2$, $\forall i$ (ver caso 1 del Gráfico 2.5). De este modo, como puede verse en la distribución de la figura izquierda del Gráfico 2.6, dados unos valores específicos de la variable explicativa, el rango de posibles valores que puede tomar la variable endógena tiene la misma amplitud y la probabilidad de observar elementos alejados de la media no depende del valor que tome la variable explicativa X.

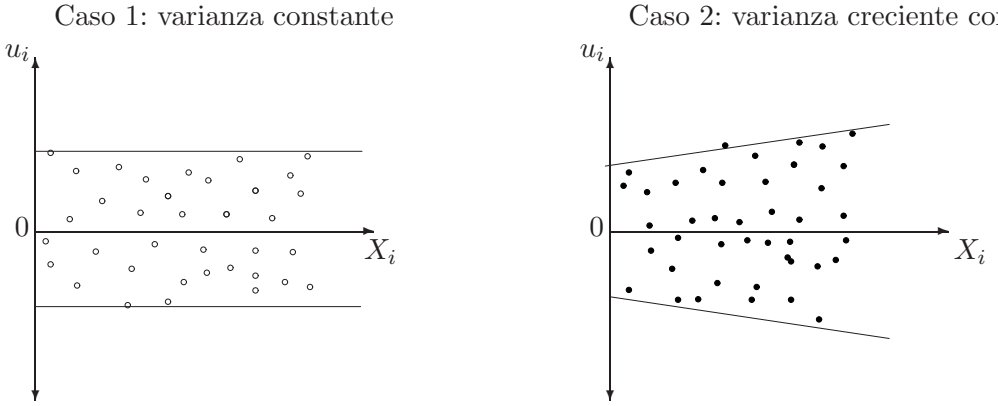


Gráfico 2.5: Ejemplos de realizaciones de u

En el caso contrario, estaríamos hablando de perturbaciones heterocedásticas, cuya dispersión puede variar a lo largo de la muestra (ver caso 2 del Gráfico 2.5). En el caso de los pisos, significaría, por ejemplo, que el rango de los precios de los pisos con menor superficie es más pequeño que el de los pisos con mayor superficie habitable (ver la figura derecha en el Gráfico 2.6). En otras palabras, los pisos pequeños y con la misma superficie tienen los precios bastante parecidos. Sin embargo, a medida que aumenta la superficie, la holgura crece y podemos encontrar pisos grandes de igual tamaño a diversos precios; es decir, $var(u_i)$ es una función creciente en X.

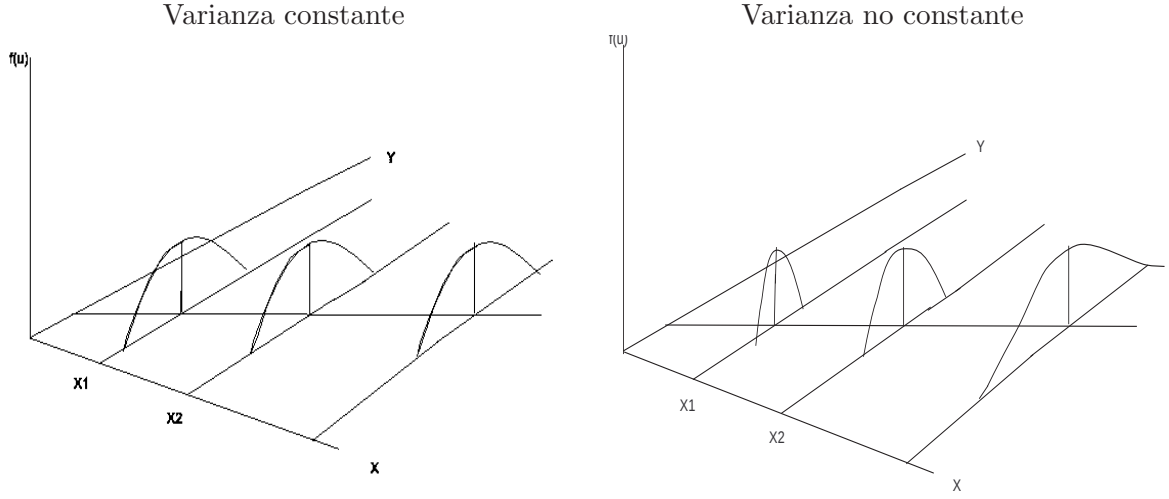


Gráfico 2.6: Ejemplos de distribución de Y

9. La perturbación no está autocorrelacionada. Por el momento vamos a suponer que la correlación entre dos observaciones distintas cualesquiera de la perturbación es cero, $corr(u_i, u_j) = r_{u_i,u_j} = 0;\ \forall i \neq j$. Esto implica que las covarianzas entre dos perturbaciones también es cero: $cov(u_i, u_j) = 0, \forall i \neq j$.

10. La perturbación sigue una distribución normal. Este último supuesto, como veremos más adelante, no se necesita para la estimación ni para la obtención de propiedades del estimador[2]. Sin embargo es necesario para poder realizar contraste de hipótesis o calcular intervalos de confianza.

2.3.1. Resumen: modelo de regresión lineal simple con hipótesis básicas

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

$$Y_i = \alpha + \beta X_i + u_i, \quad X_i \text{ fija y } u_i \sim NID(0, \sigma^2) \quad \forall i$$

Es decir, $Y_i \sim NID(\alpha + \beta X_i, \sigma^2)$, siendo α, β y σ^2 parámetros desconocidos. En particular, nos interesamos por los parámetros de la media y su interpretación en este modelo es:

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

- $\beta = \dfrac{\Delta E(Y_i)}{\Delta X_i} = \dfrac{\partial E(Y_i)}{\partial X_i}$: un aumento unitario en la variable explicativa conlleva un aumento medio de β unidades en la variable endógena. La pendiente mide el efecto de un aumento marginal en la variable explicativa sobre $E(Y_i)$.

→ Así, volviendo a nuestro ejemplo tenemos que:

$\alpha = E(P_i|F2_i = 0)$ 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.

$\beta = \dfrac{\Delta E(P_i)}{\Delta F2_i}$ indica que, cuando un piso aumenta su superficie hábil en un pie cuadrado, su precio medio aumenta en β miles \$.

2.4. 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. Esto significa que tenemos N igualdades con una misma estructura:

[2] Esto es así porque el método de estimación que se va a derivar es el de Mínimos Cuadrados Ordinarios. Sin embargo, si se estimase por máxima verosimilitud el supuesto de normalidad sobre la distribución de Y sí es necesario para la obtención del estimador.

$$\begin{aligned} Y_1 &= \alpha + \beta X_1 + u_1 \\ &\vdots \\ Y_i &= \alpha + \beta X_i + u_i \\ &\vdots \\ Y_N &= \alpha + \beta X_N + u_N \end{aligned}$$

El Gráfico 2.7 representa gráficamente una posible muestra. Los puntos (Y_i, X_i) se sitúan o distribuyen alrededor de la recta $\alpha + \beta X_i$. 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 el Gráfico 2.7, 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.

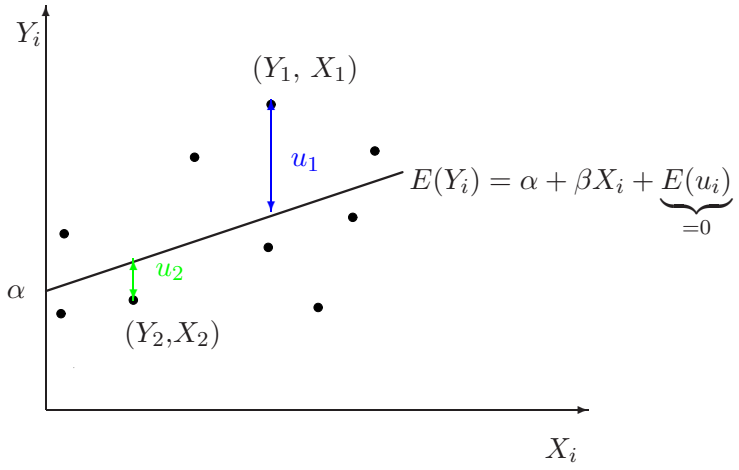


Gráfico 2.7: Modelo de regresión simple

Así, la recta central sería aquella recta que se obtiene cuando el valor de la perturbación es cero. Teniendo en cuenta que suponemos que la perturbación tiene media cero, es decir, que no tiene efectos sistemáticos sobre Y, la *recta central* recoge el comportamiento medio de la variable de interés. La ***estimación*** de un modelo de regresión pretende obtener una aproximación a esta recta central no observable. En términos econométricos, queremos calcular el comportamiento medio de la variable de interés, $\alpha + \beta X_i$, a partir de observaciones provenientes de una muestra $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_N, X_N)$. Gráficamente, la estimación consiste en calcular la pendiente y la ordenada que mejor se ajusta a la nube de puntos.

Antes de proceder a la estimación del modelo es preciso definir algunos nuevos conceptos. La recta central objeto de estimación se denomina ***Función de Regresión Poblacional (FRP)*** y depende de los coeficientes poblacionales desconocidos α y β. Se trata de la parte sistemática o predecible del modelo y corresponde al comportamiento medio o esperado de la variable a explicar:

$$E(Y_i) = E(\alpha + \beta X_i + u_i) \ = \ \alpha + \beta X_i + \underbrace{E(u_i)}_{=0} \ = \ \alpha + \beta X_i$$

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

$$u_i = Y_i - \alpha - \beta X_i$$

El resultado final obtenido a partir de la información que ofrece una muestra dada se define como la ***Función de Regresión Muestral (FRM)***. Se obtiene una vez que los coeficientes de la regresión hayan sido estimados $(\hat{\alpha}, \hat{\beta})$ y también se conoce como ***modelo estimado***:

$$\hat{Y}_i = \widehat{E(Y_i)} = \hat{\alpha} + \hat{\beta} X_i$$

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:

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i = \alpha + \beta X_i + u_i - \hat{\alpha} - \hat{\beta} X_i \\ &= (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) X_i + u_i \end{aligned} \tag{2.3}$$

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 (α, β) introduce un error adicional. Es importante, por tanto, diferenciar y no confundir el residuo con la perturbación.

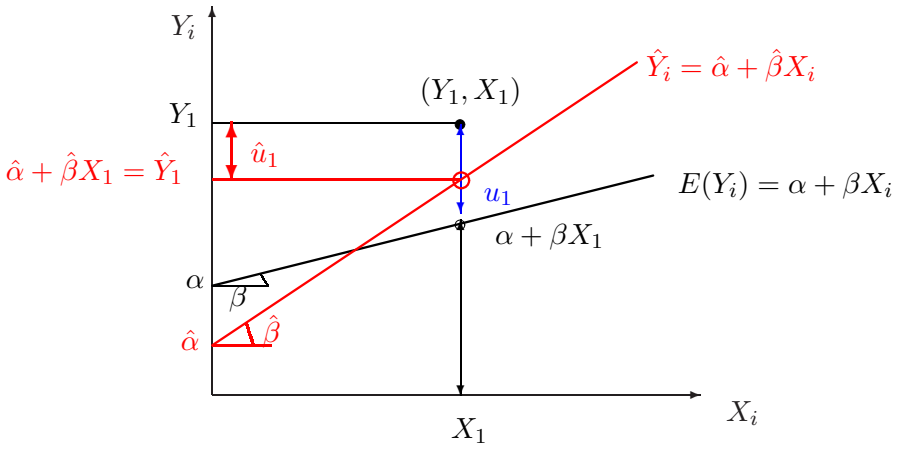


Gráfico 2.8: Función de regresión poblacional y función de regresión muestral

En el Gráfico 2.8 la función de regresión poblacional está trazada en color negro así como los coeficientes poblacionales, la ordenada (α) y la pendiente (β). Podemos ver que el valor Y_i se obtiene como la suma del valor que toma la parte sistemática $\alpha + \beta X_i$ (situada sobre la FRP) y del valor que toma la perturbación u_i, esto es, $Y_i = \alpha + \beta X_i + u_i$.

La función de regresión muestral y los coeficientes estimados ($\hat{\alpha}$ y $\hat{\beta}$) 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 $(\hat{\alpha} \neq \alpha, \hat{\beta} \neq \beta)$. Basándonos en la FRM podemos obtener el valor del punto Y_i como la suma del valor estimado de la parte sistemática $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ (situado sobre la FRM) y del valor que toma el residuo $\hat{u}_i$, esto es, $Y_i = \hat{Y}_i + \hat{u}_i$.

2.4.1. 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 $\hat{\alpha}$ y $\hat{\beta}$ 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 α y β se obtiene de minimizar la suma de los residuos al cuadrado:

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} \hat{u}_i^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \tag{2.4}$$

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

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0$$

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

$$\sum_{i=1}^{N} \underbrace{(Y_i - \hat{\alpha} - \hat{\beta} X_i)}_{\widehat{u}_i} = 0 \tag{2.5}$$

$$\sum_{i=1}^{N} \underbrace{(Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i}_{\widehat{u}_i X_i} = 0 \tag{2.6}$$

Las expresiones de los estimadores MCO para los coeficientes poblacionales α y β se obtienen de resolver las ecuaciones para $\hat{\alpha}$ y $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2} \tag{2.7}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{2.8}$$

2.4.2. Propiedades de los estimadores MCO

Necesitamos saber cuáles son las propiedades que justifican el uso de los estimadores MCO en el modelo de regresión simple bajo las hipótesis básicas. Los estimadores $\widehat{\alpha}$ y $\widehat{\beta}$ son **lineales** en la perturbación, es decir, pueden expresarse como una combinación lineal de las perturbaciones $u_1, \ldots, u_N$. En segundo lugar, los estimadores MCO son variables aleatorias cuya distribución está centrada alrededor del valor poblacional, esto es

$$E(\hat{\alpha}) = \alpha \qquad E(\hat{\beta}) = \beta$$

y, por tanto, son estimadores **insesgados**. Y en cuanto a la precisión, el Teorema de Gauss-Markov prueba que los estimadores MCO tienen **mínima varianza** dentro del conjunto de los estimadores lineales (en u) e insesgados. Las varianzas y covarianza para los estimadores son las siguientes:

$$var(\hat{\alpha}) = \sigma^2 \left(\frac{\sum_{i=1}^{N} X_i^2}{N \sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{X}^2}{N\, S_X^2} \right) \tag{2.9}$$

$$var(\hat{\beta}) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = \frac{\sigma^2}{N} \frac{1}{S_X^2} \tag{2.10}$$

$$cov(\hat{\alpha}, \hat{\beta}) = \sigma^2 \left(-\frac{\bar{X}}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = -\frac{\sigma^2}{N} \frac{\bar{X}}{S_X^2} \tag{2.11}$$

Ambas varianzas dependen de la dispersión de la perturbación $var(u_i) = \sigma^2$, del tamaño muestral y de la dispersión del regresor X. En ambos casos, cuanto mayor sea N o la variabilidad de X, S_x^2, menor es la varianza de los estimadores MCO. En cuanto a la covarianza será no nula a no ser que la media aritmética de la variable explicativa sea cero.

2.4.3. La estimación MCO en Gretl

→ Como ejemplo, calcularemos las estimaciones MCO del modelo para el precio de la vivienda, $P_i = \alpha + \beta F2_i + u_i$, con la muestra del fichero *datos-cap3.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 $\hat{\alpha} = 52,4$ y $\hat{\beta} = 0,139$, como se puede ver en el Gráfico 2.2.

Vamos a ver cómo podemos obtener una tabla de resultados detallados. Una vez iniciada la sesión de Gretl y abierto el fichero *datos-cap3.gdt*, vamos a

Modelo →Mínimos cuadrados ordinarios...

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

Gráfico 2.9: Ventana de especificación del modelo lineal

- 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 $F2$ y luego *Añadir* $->$. La ventana de especificación aparece en el Gráfico 2.9.

Tras pinchar en *Aceptar* aparece la ventana de resultados del modelo (ver el Gráfico 2.10).

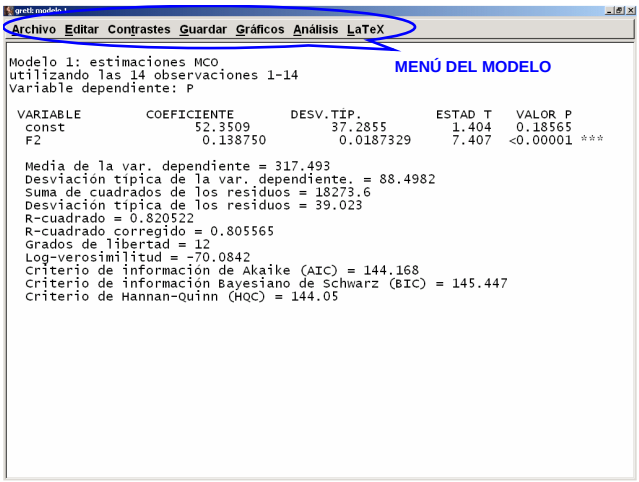


Gráfico 2.10: Ventana de resultados de estimación MCO

En esta ventana aparecen los resultados básicos para el análisis del modelo y que se explican detalladamente a lo largo del curso. La primera columna muestra las variables explicativas que se han incluido en el modelo, la constante (*const*) y la superficie que posee la vivienda ($F2$). 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 $\widehat{\alpha}$= 52,35 miles de dólares y la estimación de la pendiente es $\widehat{\beta} = 0,138750$ miles \$ por pie cuadrado. Así la función de regresión muestral es:

$$\widehat{P}_i = 52,3509 + 0{,}138750\ F2_i \tag{2.12}$$

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

Esta ventana de resultados del modelo tiene un menú con siete opciones, *Archivo, Editar, Contrastes, Guardar, Gráficos, Análisis* y *Latex*, que sirven para mostrar otro tipo de resultados de estimación o guardarlos. Veamos algunas de estas utilidades.

Guardar resultados. Si en el menú de resultados del modelo vamos a *Archivo* $\rightarrow$*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 (ver el Gráfico 2.11), 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

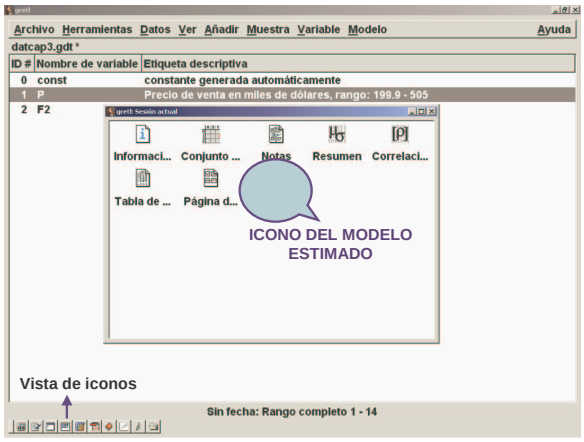


Gráfico 2.11: Ventana de iconos: recuperar resultados estimación

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 F2* obtenemos el gráfico de dispersión de las observaciones reales P_i frente a la variable explicativa $F2_i$ junto con la función de regresión muestral (2.12). El resultado es la figura izquierda del Gráfico 2.12.

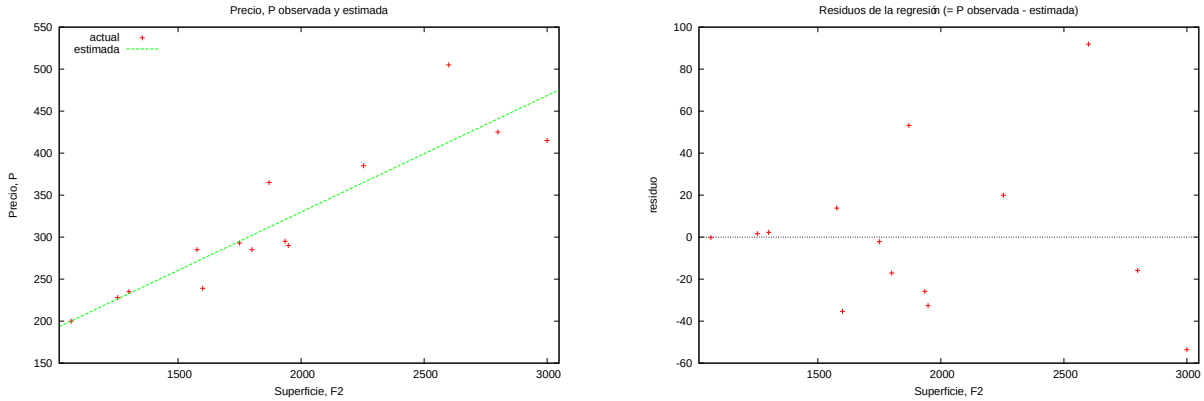


Gráfico 2.12: Gráficos de resultados de regresión MCO

- Si seleccionamos *Gráficos* →*Gráfico de residuos* →*contra F2*, se representan los errores de ajuste $\hat{u}_i$ sobre la variable explicativa $F2_i$, es decir, el diagrama de dispersión de los pares de puntos $(F2_1, \hat{u}_1), \ldots, (F2_{14}, \hat{u}_{14})$, como aparece en la figura derecha del Gráfico 2.12. 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 2.2. Podemos guardar cualquiera de estos valores seleccionando la opción *Guardar* del menú del modelo, tal como muestra el Gráfico 2.13.

```
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 2.2: Residuos de la regresión MCO.

Para almacenar $\hat{P}_i$ hay que elegir *Guardar* →*Valores estimados*. Sale una ventanilla 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 ventanilla 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 $F2$.

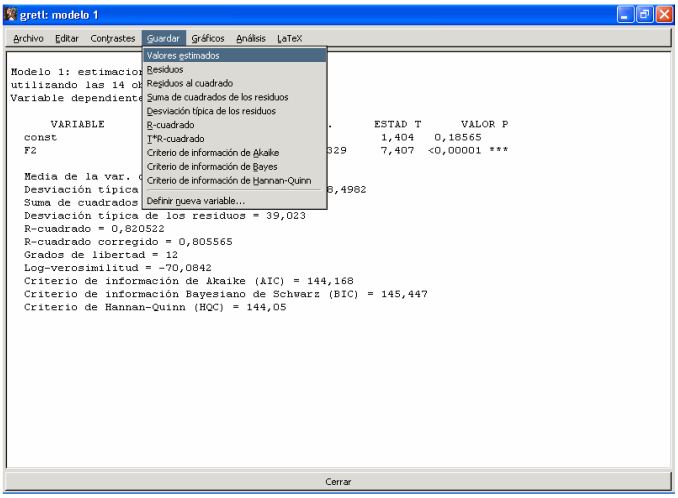


Gráfico 2.13: Residuos MCO

2.4.4. Propiedades de la recta mínimo-cuadrática

Vamos a realizar un pequeño análisis de las variables que intervienen en la regresión mínimo-cuadrática, con objeto de estudiar las similitudes y relaciones que pueden existir entre ellas. Finalmente, generalizaremos estos resultados, comprobando que estas propiedades se cumplen en cualquier regresión lineal mínimo-cuadrática.

Comenzaremos obteniendo los estadísticos descriptivos del regresor $F2$, la variable endógena P, su ajuste $\hat{P}$ y su residuo $\hat{u}$ en *Ver* →*Estadísticos principales* de la ventana inicial de Gretl:

Estadísticos principales, usando las observaciones 1 - 14

Variable	Media	Mediana	Mínimo	Máximo
P	317,493	291,500	199,900	505,000
F2	1910,93	1835,00	1065,00	3000,00
phat1	317,493	306,958	200,120	468,602
uhat1	0,000000	−1,1919	−53,601	91,8983

Variable	Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
precio	88,4982	0,278741	0,653457	−0,529833
F2	577,757	0,302344	0,485258	−0,672125
phat1	80,1640	0,252491	0,485258	−0,672125
uhat1	37,4921	6,15597e+15	1,02687	0,817927

Tabla 2.3: Estadísticos descriptivos de variables de la FRM

Analizando esta tabla-resumen de los datos comprobamos que:

i) La media de los residuos (*uhat1*) es cero, $\bar{\hat{u}} = 0$.

ii) Las medias de la variable dependiente P_i y la estimada (*phat1*) coinciden, $\bar{P} = \bar{\widehat{P}}$.

iii) Los coeficientes de asimetría y curtosis de la variable dependiente ajustada $\widehat{P}_i$ coinciden con las de la variable independiente $F2_i$.

A continuación, vamos a analizar las relaciones lineales existentes entre estas variables. Mediante *Ver →Matriz de correlación* obtenemos la siguiente matriz de correlaciones:

Coeficientes de correlación, usando las observaciones 1 - 14
valor crítico al 5\% (a dos colas) = 0,5324 para n = 14

P	F2	uhat1	phat1	
1,0000	0,9058	0,4236	0,9058	P
	1,0000	−0,0000	1,0000	F2
		1,0000	−0,0000	uhat1
			1,0000	phat1

Tabla 2.4: Matriz de correlaciones

Podemos ver que:

iv) Los valores ajustados $\widehat{P}_i$ y el regresor $F2_i$ están perfectamente correlacionados, $r_{\widehat{P}F2} = 1$.

v) La correlación entre los valores observados P_i con los valores ajustados $\widehat{P}_i$ y la variable explicativa $F2_i$ es la misma, $r_{P\widehat{P}} = r_{PF2}$.

vi) Los residuos $\widehat{u}_i$ y la variable explicativa $F2_i$ están incorrelacionados, $r_{\widehat{u}F2} = 0$.

vii) Los residuos $\widehat{u}_i$ y la variable ajustada $\widehat{P}_i$ están incorrelacionados, $r_{\widehat{u}\widehat{P}} = 0$.

Justificación de estos resultados: La propiedad *i)* se deriva de la primera ecuación normal (2.5), que nos indica que la suma de los residuos ha de ser cero, por lo que $\bar{\hat{u}} = 0$. Notar que la primera ecuación normal existe sólo si el modelo tiene término independiente y no en otro caso. Por lo tanto, los resultados que se obtienen derivados de ella solo se cumplen en el caso

de que el término independiente exista. De $\bar{\hat{u}} = 0$ y como $\bar{Y} = \bar{\hat{Y}} + \bar{\hat{u}}$, se obtiene la propiedad *ii)*.

Las propiedades *iii)*, *iv)* y *v)* se deben a que los valores de $\widehat{P}$ se obtienen de un cambio de origen y escala de la variable $F2$, $\widehat{P} = \hat{\alpha} + \hat{\beta}F2$. Esta relación implica que sus distribuciones de frecuencias tienen las mismas las medidas de forma, están perfectamente correlacionadas entre sí y tienen la misma correlación lineal frente a terceras variables.

La propiedad *vi)* se deriva de las ecuaciones normales (2.5), que indica que $\bar{\hat{u}} = 0$, y (2.6), que implica que los residuos son ortogonales a la variable explicativa X, $\sum_i X_i \hat{u}_i = 0$. Como consecuencia, la covarianza muestral entre residuo y variable explicativa es cero:

$$S_{X\hat{u}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}}) = \frac{1}{N}\sum_{i=1}^{N} X_i\hat{u}_i - \bar{X}\bar{\hat{u}} = 0$$

y, por tanto, la correlación entre ambas variables es: $r_{\hat{u}X} = S_{\hat{u}X}/S_{\hat{u}}S_X = 0$. Esto nos viene a decir que en la parte del modelo que queda sin explicar, el residuo $\hat{u}$, ya no queda nada que la variable exógena X pueda explicar o aportar en términos lineales. Finalmente, basándonos en que $r_{\hat{u}X} = 0$ y que el ajuste $\hat{Y}$ es una transformación lineal de X, se demuestra la propiedad *vii)*, $r_{\widehat{u}\widehat{Y}} = 0$. De esta condición y dado que $Y_i = \hat{Y}_i + \hat{u}_i$, se deriva una última propiedad:

viii) La varianza muestral de Y puede descomponerse en dos términos: la varianza explicada por X y la varianza residual, es decir,

$$S_Y^2 = S_{\hat{Y}}^2 + S_{\hat{u}}^2$$

2.4.5. La precisión de la estimación y la bondad del ajuste

Una vez realizada las estimaciones de los coeficientes del modelo, la siguiente etapa del análisis consiste en el análisis y evaluación de los resultados. Por ejemplo,

1. Obtener una medida de la precisión en la estimación de α y β.
2. Evaluar la calidad del ajuste a los datos, es decir, si la función de regresión muestral, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, resume bien el comportamiento observado de la variable endógena.
3. Evaluar si el modelo propuesto es *correcto* o si hay algún error en la especificación del modelo, en las hipótesis planteadas.

Este apartado desarrolla los puntos 1 y 2. La respuesta al punto 3 es más compleja, de modo que el siguiente apartado introduce algunos aspectos de la evaluación del modelo.

La precisión de la estimación

En el apartado 7 del tema 1 vimos que la desviación típica de la distribución muestral de los estimadores era un buen indicador de la precisión. Sin embargo, habitualmente la desviación típica de los estimadores tiene algún elemento desconocido. Esto sucede en este caso, como puede comprobarse en la expresión de las varianzas (2.9) y (2.10), que dependen de la varianza

de la perturbación $var(u_i) = \sigma^2$. Podemos obtener una estimación de la desviación típica sustituyendo el parámetro poblacional σ por un estimador insesgado, $\widehat{\sigma}$. El resultado se conoce como **errores típicos de los coeficientes de la regresión**, es decir,

$$\begin{aligned} \textit{Error típico } (\hat{\alpha}) &= \widehat{des}(\hat{\alpha}) = \frac{\hat{\sigma}}{\sqrt{N}}\sqrt{1+\frac{\bar{X}^2}{N\,S_X^2}} \\ \textit{Error típico } (\hat{\beta}) &= \widehat{des}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{N}}\frac{1}{S_X} \end{aligned}$$

Un estimador insesgado de la varianza σ^2 es:

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

donde $\sum_i \hat{u}_i^2$ es la **suma de cuadrados residual**, (o SCR), y $N-2$ son los grados de libertad que tenemos tras estimar α y β. Su raíz cuadrada $\hat{\sigma}$ se conoce como **error típico** de los perturbaciones o **error típico** de la regresión. Por tanto, la precisión de las estimaciones de los coeficientes aumenta con el número de observaciones N y la dispersión del regresor S_X y disminuye cuando crece el error típico $\hat{\sigma}$.

De forma similar, se construye el siguiente estimador insesgado de la matriz de las varianzas y la covarianza de los estimadores MCO:

$$\widehat{V}\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} = \begin{pmatrix}\widehat{var}(\hat{\alpha}) & \widehat{cov}(\hat{\alpha},\hat{\beta})\\ \widehat{cov}(\hat{\alpha},\hat{\beta}) & \widehat{var}(\hat{\beta})\end{pmatrix} = \hat{\sigma}^2\begin{pmatrix}\left(\frac{1}{N}+\frac{\bar{X}^2}{\sum_i(X_i-\bar{X})^2}\right) & \frac{-\bar{X}}{\sum_i(X_i-\bar{X})^2}\\ & \frac{1}{\sum_i(X_i-\bar{X})^2}\end{pmatrix}$$

→ **Errores típicos de estimación y estimación de las varianzas en Gretl.** En los resultados de estimación del caso práctico aparecen los siguientes valores relacionados con la precisión:

```
Modelo 1: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: P

  VARIABLE       COEFICIENTE     DESV.TÍP.       ESTAD T     VALOR P
  const           52,3509        37,2855         1,404      0,18565
  F2               0,138750       0,0187329      7,407     <0,00001 ***

Suma de cuadrados de los residuos = 18273,6
Desviación típica de los residuos = 39,023
```

La columna encabezada por *DESV. TÍP.* proporciona los errores típicos de estimación, es decir, $\widehat{des}(\hat{\alpha})$ y $\widehat{des}(\hat{\beta})$. Se observa que es más precisa la estimación del efecto marginal de la superficie del piso β que la de la ordenada α ya que su varianza estimada es menor. La *desviación típica de los residuos* es el error típico $\hat{\sigma}$ y *Suma de cuadrados de los residuos* es $SCR = \sum_i \hat{u}_i^2$.

En esta tabla no aparece la estimación de la varianza de la perturbación, pero se puede calcular:

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

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

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 2.5: Estimación de varianzas y covarianza de $\hat{\alpha}$ y $\hat{\beta}$.

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

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. Por otro lado, interesa tener una medida que nos indique, en la medida de lo posible, si estamos ante unos buenos resultados de ajuste a los datos de la función de regresión muestral.

Bondad del ajuste

La medida de la bondad del ajuste que vamos a utilizar es el coeficiente de determinación, R^2 ó R-cuadrado. Este coeficiente, descrito al final de la primera práctica, tiene la siguiente expresión en el modelo de regresión lineal simple:

$$R^2 = r_{XY}^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{\sum_i (\hat{Y}_i - \bar{\hat{Y}}_i)^2}{\sum_i (Y_i - \bar{Y})^2} \tag{2.13}$$

Este coeficiente mide la ganancia obtenida al pasar de un modelo sin variable explicativa X:

$$Y_i = \alpha + u_i$$

a otro en el que se incluye esta variable: $Y_i = \alpha + \beta X_i + u_i$

Por tanto el R-cuadrado 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. 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, es decir, $Y_i = \beta X_i + u_i$.

→ Si analizamos el caso práctico, vemos que el coeficiente de determinación aparece en la tabla de resultados básicos de estimación, `R-cuadrado = 0,820522`. Podemos decir que este ajuste es bueno, ya que la variabilidad muestral de la superficie de la vivienda ($F2$) ha explicado el 82 % de la variabilidad muestral de los precios de venta de dichas viviendas (P).

2.5. Contrastes de hipótesis e intervalos de confianza

Al proponer un modelo para el precio de los pisos hemos asumido que el tamaño del piso es el factor más relevante en la fijación de su precio. Las conclusiones que obtengamos de la estimación y predicción dependerán del cumplimiento de esta hipótesis. Por tanto, conviene valorar si este supuesto es sensato. Para ello vamos a utilizar los contrastes de hipótesis y los intervalos de confianza sobre la distribución de los estimadores. El planteamiento es el siguiente:

- Si el precio de un piso no se ve afectado por su superficie, entonces su efecto marginal es cero, luego $\beta = 0$, y diremos que la variable explicativa no es significativa o relevante para explicar Y. Si esto es cierto, el modelo propuesto no tiene sentido y debemos reformularlo.
- Por el contrario, si el precio está relacionado con la superficie del piso, entonces $\beta \neq 0$ y decimos que el regresor X es significativo o relevante para explicar (y predecir) Y.

2.5.1. Contrastes de hipótesis sobre β

Contraste de significatividad individual de X. Para verificar si la variable independiente $F2$ es significativa para determinar el precio medio de la vivienda, podemos realizar un contraste. Planteamos las siguientes hipótesis a contrastar:

$$\begin{cases} H_0 \colon \beta = 0 & (X \text{ no es significativa o relevante para explicar } Y) \\ H_a \colon \beta \neq 0 & (X \text{ es significativa o relevante para explicar } Y) \end{cases}$$

Para obtener un estadístico de contraste partimos de la siguiente variable aleatoria:

$$\frac{\widehat{\beta} - \beta}{\widehat{des}(\widehat{\beta})} \sim t_{(N-K)} \tag{2.14}$$

El estadístico del contraste se obtiene sustituyendo en esta variable el valor recogido en H_0:

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

Se trata de un estadístico tipo t similar al visto en el apartado 7.2 del tema 1. Es un contraste bilateral, como se observa en el siguiente gráfico de la distribución del estadístico bajo H_0:

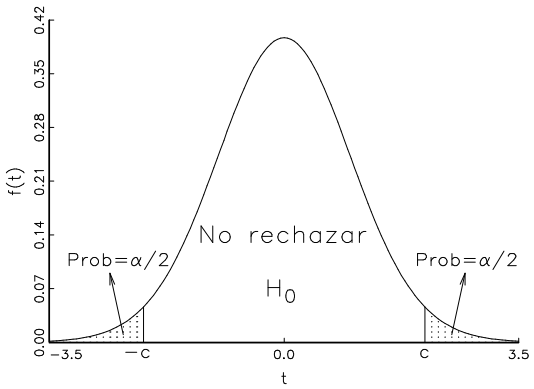


Gráfico 2.14: Criterio de decisión del contraste de significatividad individual

por lo que la regla de decisión es la siguiente: fijado un nivel de significación α,

- Rechazamos H_0 si el valor muestral del estadístico t^m pertenece a la región crítica, es decir, si es menor que $-c = -t_{(N-K)\alpha/2}$ o bien mayor que $c = t_{(N-K)\alpha/2}$ y concluimos que la *variable explicativa es relevante*.
- No rechazamos H_0 en otro caso, es decir, si el valor muestral t^m se sitúa en el intervalo $[-c, c]$ con $c = t_{(N-K)\alpha/2}$. Concluimos que la variable X *no es relevante* o significativa para explicar la variable dependiente Y.

→ Veamos si la superficie de la vivienda es un factor relevante para determinar su precio:

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

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

$$ESTAD\ T = t^m = 7,4068 = \frac{columna\ COEFICIENTE}{columna\ DESV.TIP.} = \frac{0,13875}{0,0187329}$$

El valor crítico del contraste para el nivel de significación del 5 % es $c = t_{(14-2)0,05/2} = 2,179$. Como resultado tenemos que $7,4068 > 2,179$, por lo que t^m pertenece a la región crítica y, en consecuencia, rechazamos H_0 a un nivel de significación del 5 %. Podemos concluir que la variable $F2$ 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.

Otros contrastes sobre β. Como hay evidencia estadística de que β 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 ejem-

plos.

→ **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:

$$\begin{cases} H_0 : \beta = 0,1 \\ H_a : \beta \neq 0,1 \end{cases}$$

Sustituyendo en la variable (2.14) el valor bajo H_0, obtenemos el estadístico de contraste:

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

Hay que tener en cuenta que la columna *ESTAD T* de los resultados de estimación de Gretl, corresponde al valor muestral del estadístico para H_0: $\beta = 0$. Por tanto, tenemos que calcular el valor muestral del estadístico de contraste, que en este caso es:

$$t^m = \frac{0,138750 - 0,1}{0,0187329} = 2,068$$

El valor crítico para $\alpha = 5\,\%$ es $c = t_{(14-2)0,05/2} = 2,179$. Como el valor calculado cae fuera de la región crítica, $-2,179 < 2,068 < 2,179$, 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:

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

El estadístico de contraste en este caso toma el valor

$$t^m = \frac{0,138750 - 0,15}{0,0187329} = -0,6005 \Rightarrow -c = -2,179 < -0,6005 < 2,179 = c$$

con $c = t_{(12)0,025}$. Así, no rechazamos H_0 a un nivel de significación del $5\,\%$ y también es posible que si $\Delta F2 = 1$, entonces el precio medio de la vivienda aumente en 150$.

Si observamos los contrastes anteriores, siempre y cuando el valor del estadístico calculado t^m esté fuera de la región crítica, es decir, en el intervalo $[-2,179; 2,179]$ no rechazaremos la hipótesis nula propuesta.

2.5.2. Intervalos de confianza

Un intervalo de confianza está definido por dos valores entre los cuales se encuentra el valor del parámetro con un determinado nivel de confianza que se denota $(1 - \alpha)$. Para obtener el intervalo de confianza del coeficiente β, definimos el intervalo de valores que tiene una probabilidad $(1 - \alpha)$ en la distribución (2.14) asociada al estimador. Así

$$Prob\left[-t_{(N-2)\alpha/2} \leq \frac{\widehat{\beta}-\beta}{\widehat{des}(\widehat{\beta})} \leq t_{(N-2)\alpha/2}\right] \;=\; 1-\alpha$$

Reordenamos:

$$Prob\left[\widehat{\beta}-t_{(N-2)\alpha/2}\,\widehat{des}(\widehat{\beta}) \leq \beta \leq \widehat{\beta}+t_{(N-2)\alpha/2}\,\widehat{des}(\widehat{\beta})\right] \;=\; 1-\alpha$$

y obtenemos el intervalo de confianza $(1-\alpha)$ para el parámetro β. Observamos que está centrado en la estimación puntual y que se desvía en una cantidad que está dada por $t_{(N-K)\alpha/2}$ veces su error típico de estimación, $\widehat{des}(\widehat{\beta})$. Si estimamos con muy poca precisión, este intervalo será amplio. Esto quiere decir que la variabilidad muestral del estimador acota a β en un intervalo más amplio. En lo que sigue del curso emplearemos la siguiente notación para expresar el intervalo de confianza:

$$IC(\beta)_{1-\alpha} = \left[\hat{\beta} \pm t_{(N-2)\alpha/2}\;\widehat{des}(\hat{\beta})\right]$$

El correspondiente intervalo de confianza para α se obtiene de forma similar:

$$IC(\alpha)_{1-\alpha} = \left[\hat{\alpha} \pm t_{(N-2)\alpha/2}\;\widehat{des}(\hat{\alpha})\right]$$

→ Continuando con la práctica, 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 es:

```
t(12, .025) = 2,179

      VARIABLE      COEFICIENTE      INTERVALO DE CONFIANZA 95%
           const         52,3509         (-28,8872, 133,589)
              F2          0,138750       (0,0979349, 0,179566)
```

Tabla 2.6: Estimación por intervalo

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

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

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

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.

2.6. 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^m de significatividad individual o los grados de libertad. Por ejemplo,

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

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

Bibliografía

Alonso, A., Fernández, F. J. e I. Gallastegui (2005), *Econometría*, Prentice-Hall.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2ª edn., South-Western.

Tema 3

Modelo de Regresión Lineal Múltiple

Contenido

3.1. Introducción. Un ejemplo

En este tema consideramos introducir en el modelo de regresión, además del término constante, más de una variable explicativa por lo que pasamos del llamado modelo de regresión lineal simple al modelo de regresión lineal múltiple.

Comenzamos con el ejemplo que se ha seguido en el tema sobre el Modelo de Regresión Lineal Simple. El precio de una casa, en miles de dólares, (P) era la variable dependiente y las variables explicativas eran el término constante y el tamaño de la casa o el número de pies cuadrados del área habitable (F2). Ampliaremos el modelo incluyendo dos variables explicativas más, el número de habitaciones (BEDRMS) y el número de baños (BATHS) siendo el modelo de regresión lineal múltiple[1]

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \qquad i = 1, 2, \ldots, N \tag{3.1}$$

El modelo de regresión lineal general (MRLG), con K variables explicativas

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + u_i \qquad i = 1, 2, \ldots, N. \tag{3.2}$$

se puede escribir en notación matricial:

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

donde cada uno de los elementos se definen:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Por el momento, seguimos suponiendo las mismas hipótesis básicas sobre el término de perturbación y sobre las variables explicativas o regresores, a saber:

i) $E(u_i) = 0 \quad \forall\, i, \quad E(u_i^2) = \sigma^2 \quad \forall\, i, \quad E(u_i u_j) = 0 \quad \forall i \neq j.$

ii) La perturbación sigue una distribución normal.

iii) Las variables X_2 a X_k no son estocásticas. Esto quiere decir que en muestras repetidas de N observaciones de $Y_i, X_{2i}, \ldots, X_{ki}$, las variables $X_{2i}, \ldots, X_{ki},\ i = 1, \ldots, N$ tomarían siempre los mismos valores. Este supuesto, junto a $E(u_i) = 0$, implica que los regresores y el término de perturbación están incorrelacionados.

iv) Los regresores son linealmente independientes, esto quiere decir que el rango de la matriz de datos de los regresores X es K tal que no tiene columnas repetidas ni unas son combinaciones lineales de otras.

v) Además se supone que se dispone de un número suficiente de observaciones para estimar los parámetros $\beta_j,\ \ j = 1, \ldots, K$, esto es $K < N$.

[1] Dado que seguimos con los mismos datos de sección cruzada utilizamos el subíndice $i = 1, \ldots, N$. La notación para datos de series temporales suele ser $t = 1, \ldots, T$.

Interpretación de cada uno de los coeficientes de regresión:

- Los parámetros β_j, $j = 2, \ldots, K$:

  Manteniendo constante el valor del resto de variables explicativas, si X_{ji} cambia en una unidad, Y_i se espera que cambie en media β_j unidades.

- El parámetro β_1 que acompaña al término constante recoge el valor esperado de la variable dependiente cuando el resto de variables explicativas o regresores incluidos toman el valor cero.

Siguiendo con el ejemplo, el modelo (3.1) se puede escribir en notación matricial:

$$\underset{(N\times 1)}{Y} = \underset{(N\times 4)}{X}\ \underset{(4\times 1)}{\beta} + \underset{(N\times 1)}{u}$$

donde cada uno de los elementos se definen:

$$Y = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & F2_1 & BEDRMS_1 & BATHS_1 \\ 1 & F2_2 & BEDRMS_2 & BATHS_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & F2_N & BEDRMS_N & BATHS_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Interpretación de los coeficientes:

- El coeficiente β_1 es el valor medio esperado de aquellas viviendas que no tienen ningún pie cuadrado de área habitable, ni habitaciones ni baños.

- El coeficiente β_2:

  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 β_3:

  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 β_4:

  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.

El análisis de regresión múltiple nos permite examinar el **efecto marginal** de una variable explicativa en particular, una vez hemos controlado por otras características recogidas en el resto de variables explicativas que mantenemos constantes. Por eso a veces al resto de regresores se les llama variables de control. Veremos más adelante cuándo es importante controlar por otras variables y qué problemas tendremos si las omitimos.

3.2. Estimación de Mínimos Cuadrados Ordinarios utilizando Gretl

Se dispone de una base de datos sobre el precio de venta de una vivienda y distintas características de 14 viviendas vendidas en la comunidad universitaria de San Diego en 1990. Son datos de sección cruzada y las variables que se consideran son:

P:	Precio de venta en miles de dólares (Rango 199.9 - 505)
F2:	Pies cuadrados de área habitable (Rango 1065 - 3000)
BEDRMS:	Número de habitaciones (Rango 3 - 4)
BATHS:	Número de baños (Rango 1,75 - 3)

Los datos para P y F2 son los mismos que los utilizados en el ejemplo del Tema 2 sobre el modelo de regresión lineal simple. Además tenemos información sobre dos nuevas variables que vamos a considerar incluir como explicativas en el modelo para el precio de la vivienda.

Comenzamos una sesión en Gretl para estimar este modelo con la muestra de 14 viviendas:

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \qquad i = 1, \ldots, 14$$

En la parte de arriba de la ventana principal de Gretl tenemos distintas opciones. Si posicionamos el cursor podemos ir eligiendo dentro de ellas.

1. Leemos los datos que están disponibles en Gretl como archivo de muestra:

   Archivo → *Abrir datos* → *Archivo de muestra*

   Elegir de Ramanathan el fichero *data4-1* proporcionados en el cuarto capítulo del libro de Ramanathan (2002). *Abrir*.

2. Podemos ver los datos de todas las variables. Las dos primeras columnas coinciden con los datos utilizados en el Tema 2.

P	F2	BEDRMS	BATHS
199.9	1065	3	1.75
228.0	1254	3	2.00
235.0	1300	3	2.00
285.0	1577	4	2.50
239.0	1600	3	2.00
293.0	1750	4	2.00
285.0	1800	4	2.75
365.0	1870	4	2.00
295.0	1935	4	2.50
290.0	1948	4	2.00
385.0	2254	4	3.00
505.0	2600	3	2.50
425.0	2800	4	3.00
415.0	3000	4	3.00

Tabla 3.1: Modelo (3.1). Datos de características de viviendas

3. Estimación por Mínimos Cuadrados Ordinarios (MCO).

   Modelo → Mínimos Cuadrados Ordinarios

   Se abre una nueva ventana. Utilizando el cursor, seleccionar de la lista de variables de la izquierda:

   - La variable dependiente (P) y pulsar elegir.
   - Las variables independientes o regresores de esta especificación y pulsar añadir cada vez. La variable Const es el término constante o variable que toma siempre valor uno. Por defecto ya está incluida pero si no se quisiera poner se podría excluir. Simplemente habría que seleccionarla con el cursor y dar a *Quitar*.

   Pinchar en *Aceptar*.

   Aparece una nueva ventana con los resultados de la estimación[2]. Iremos comentando los resultados mostrados. Situando el cursor en la parte de arriba de esta ventana podremos ver que hay distintos menús cuyas funciones estarán asociadas a esta regresión.

4. Hay varios formatos para guardar los resultados, como por ejemplo un formato compatible con Microsoft Word mediante:

   Editar → Copiar → RTF(Ms Word)

   Abrir un documento con Microsoft Word. Elegir *Edición → Pegar*. Se pegarán todos los resultados de la ventana anterior. Guardar el documento y minimizar si se quiere volver a utilizar más tarde para pegar y guardar otros resultados.

3.3. Análisis de los resultados mostrados

En esta sección vamos a ir comentando los resultados que nos muestra el programa cuando utilizamos la opción de estimación por Mínimos Cuadrados Ordinarios. Algunos de estos resultados ya han sido comentados en el Tema 2 sobre el modelo de regresión lineal simple, pero nos servirá también de repaso. Una vez especificado el modelo, el programa Gretl muestra en la ventana **gretl:modelo1** la siguiente información sobre la estimación MCO del modelo con los datos del fichero elegido:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

[2]Recordar que esta ventana puede ser minimizada para su posible utilización posterior o el modelo puede guardarse en la sesión como icono. Si la cerramos tendríamos que volver a hacer lo mismo para obtener de nuevo esta ventana y poder elegir dentro de las opciones asociadas a esta regresión.

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16700,1
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657
R^2	0,835976
$\bar{R}^2$ corregido	0,786769
$F(3, 10)$	16,9889
valor p para $F()$	0,000298587
Log-verosimilitud	−69,453
Criterio de información de Akaike	146,908
Criterio de información Bayesiano de Schwarz	149,464
Criterio de Hannan–Quinn	146,671

Algunos Gráficos.
En la ventana de resultados de estimación, Gretl nos ofrece la posibilidad de analizar el gráfico de residuos así como el gráfico de la variable observada y estimada tanto por observación como sobre las distintas variables que hay en la especificación del modelo. Por ejemplo elegimos

Gráficos → *Gráfico de residuos* → *Por número de observación*

y obtenemos el gráfico de los residuos del modelo estimado para el precio de la vivienda a lo largo de las 14 observaciones de la muestra En el gráfico 3.1 se observa que los residuos se

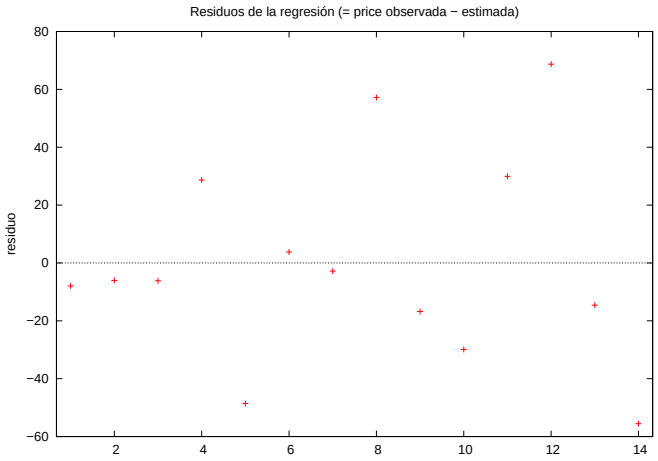


Gráfico 3.1: Gráfico de residuos por número de observación

disponen alrededor del valor cero ya que esta es su media muestral. La dispersión de estos residuos es mayor para las últimas viviendas en la muestra. Si elegimos

Gráficos → *Gráfico de residuos* → *Contra F2*

obtenemos el gráfico de los residuos sobre la variable *F2*. Este gráfico muestra que la dispersión de los residuos alrededor de su media muestral que es cero, aumenta a mayor valor de *F2*. Esto sugiere que la hipótesis básica sobre la varianza de la perturbación pueda no ser adecuada.

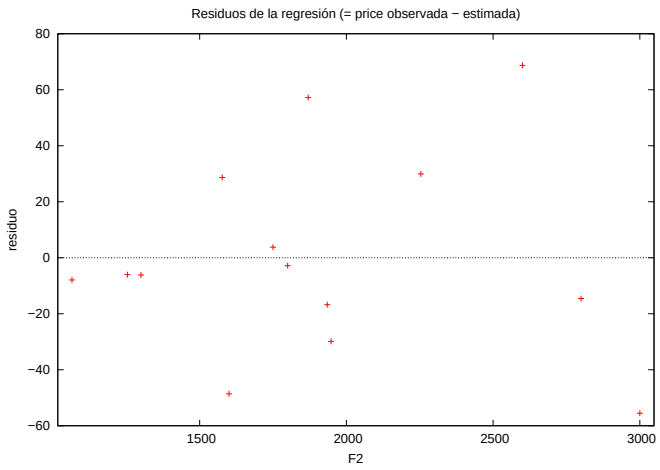


Gráfico 3.2: Gráfico de residuos contra la variable F2

Otro gráfico que ilustra la bondad del ajuste de nuestro modelo relativamente a los datos observados, es el gráfico de la variable estimada y observada por número de observación. Para obtener este gráfico elegimos

Gráficos → Gráfico de variable estimada y observada → por número de observación

De esta forma obtenemos el siguiente gráfico

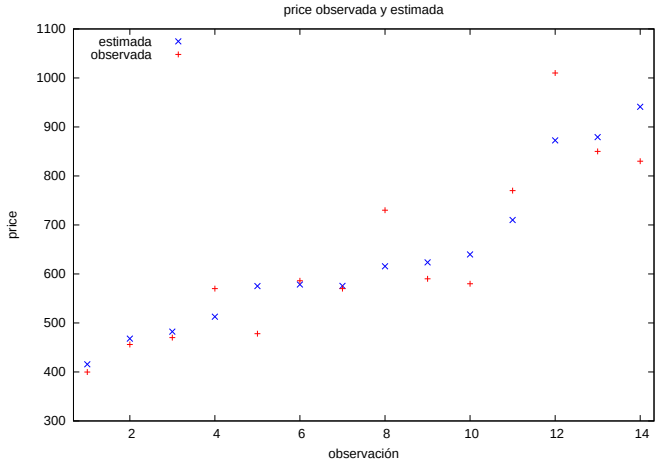


Gráfico 3.3: Gráfico de la variable estimada y observada por número de observación

En este gráfico se puede observar el valor estimado del precio de las viviendas en la muestra, dados los valores observados de las variables explicativas y el modelo estimado, en relación al precio observado. El ajuste parece empeorar para las últimas viviendas en la muestra. Si hacemos el gráfico de la variable estimada y observada contra la variable F2 que recoge el tamaño de las viviendas

Gráficos → Gráfico de variable estimada y observada → Contra F2

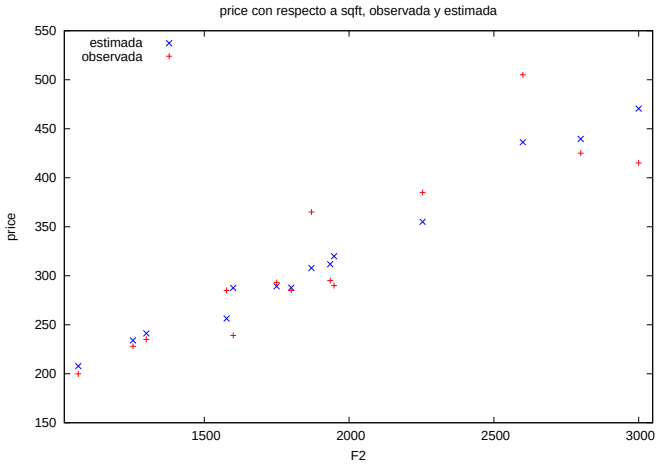


Gráfico 3.4: Gráfico de la variable estimada y observada contra F2

En el gráfico 3.4 se observa que el modelo se ajusta mejor a las observaciones asociadas a las viviendas de menor tamaño, ya que los valores estimados están más concentrados alrededor de los observados para esas viviendas. El ajuste es peor para viviendas de más de 2000 pies cuadrados.

3.3.1. Coeficientes estimados

Las estimaciones obtenidas de los coeficientes que se muestran en la segunda columna están asociados a cada una de las variables explicativas que figuran al lado en la primera columna. Dadas las realizaciones muestrales de la variable dependiente $Y_i \equiv P_i$, y explicativas, $X_{2i} \equiv F2_i$, $X_{3i} \equiv BEDRMS_i$, $X_{4i} \equiv BATHS_i$, las estimaciones se obtienen de minimizar la suma de cuadrados de los residuos con respecto a los coeficientes desconocidos $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$. Estos coeficientes estimados se han obtenido de utilizar el siguiente criterio de estimación por el método de Mínimos Cuadrados Ordinarios

$$\min_{\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \hat{\beta}_4 X_{4i})^2$$

Las condiciones de primer orden de este problema resultan en cuatro ecuaciones con cuatro incógnitas.

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \hat{\beta}_4 \sum X_{4i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} + \hat{\beta}_4 \sum X_{4i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \hat{\beta}_4 \sum X_{4i} X_{3i} \\
\sum Y_i X_{4i} &= \hat{\beta}_1 \sum X_{4i} + \hat{\beta}_2 \sum X_{2i} X_{4i} + \hat{\beta}_3 \sum X_{3i} X_{4i} + \hat{\beta}_4 \sum X_{4i}^2
\end{aligned}$$

Estas ecuaciones se conocen con el nombre de **Ecuaciones Normales**. Al igual que en el modelo de regresión lineal simple, la primera ecuación o primera condición asociada al término constante implica que la suma de los residuos debe de ser cero. El resto de ecuaciones

implican que los residuos tienen que ser ortogonales a cada una de las variables explicativas. En conjunto, estas condiciones implican que los residuos de la estimación MCO están incorrelacionados con los regresores. En términos matriciales se pueden escribir como:

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

Si las cuatro ecuaciones son linealmente independientes, el rango de $(X'X)$ es igual a $K = 4$, y por lo tanto existe una única solución a este sistema de ecuaciones. La solución será el estimador MCO del vector de parámetros β.

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

Sustituyendo los valores muestrales del fichero *data4-1* para Y y X darían lugar a las estimaciones obtenidas de los coeficientes.

Para el modelo especificado en la ecuación (3.1), la relación estimada es

$$\widehat{\mathrm{P}}_i = 129,062 + 0,1548\,\mathrm{SQFT}_i - 21,588\,\mathrm{BEDRMS}_i - 12,193\,\mathrm{BATHS}_i \tag{3.3}$$

Aunque hemos utilizado los mismos datos para P y F2 que en el Tema 2, el incluir las dos nuevas variables explicativas en el modelo ha hecho que las estimaciones de los coeficientes asociados al término constante y a F2 hayan cambiado[3].

Esto ocurre porque las nuevas variables BEDRMS y BATHS están correlacionadas con la ya incluida F2 y su media es distinta de cero[4].

Si esto no ocurriera y $\sum X_{3i} = \sum X_{4i} = \sum X_{2i}X_{3i} = \sum X_{2i}X_{4i} = 0$, las ecuaciones normales quedarían de la siguiente forma

$$\begin{aligned}
\textstyle\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \textstyle\sum X_{2i} && \Leftrightarrow \textstyle\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i}) = 0 \\
\textstyle\sum Y_i X_{2i} &= \hat{\beta}_1 \textstyle\sum X_{2i} + \hat{\beta}_2 \textstyle\sum X_{2i}^2 && \Leftrightarrow \textstyle\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i})X_{2i} = 0 \\
\textstyle\sum Y_i X_{3i} &= \hat{\beta}_3 \textstyle\sum X_{3i}^2 + \hat{\beta}_4 \textstyle\sum X_{4i}X_{3i} \\
\textstyle\sum Y_i X_{4i} &= \hat{\beta}_3 \textstyle\sum X_{3i}X_{4i} + \hat{\beta}_4 \textstyle\sum X_{4i}^2
\end{aligned}$$

[3] En el caso de considerar un MRLS solamente con F2 además de la constante se obtenía

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

$$T = 14 \quad \bar{R}^2 = 0,8056 \quad F(1,12) = 54,861 \quad \hat{\sigma} = 39,023$$

(Desviaciones típicas entre paréntesis)

[4] Usando las observaciones 1 - 14, la matriz de correlaciones entre BEDRMS, BATHS y $F2$ es

F2	BEDRMS	BATHS	
1,0000	0,4647	0,7873	F2
	1,0000	0,5323	BEDRMS
		1,0000	BATHS

y las medias muestrales de BEDRMS y BATHS son:

Variable	Media
BEDRMS	3,64286
BATHS	2,35714

Dadas esas condiciones, las dos últimas ecuaciones no dependen de $\hat{\beta}_1$ ni de $\hat{\beta}_2$ y las dos primeras ecuaciones normales coinciden con las que se obtenían en el Tema 2 para el modelo de regresión lineal simple. Por lo tanto, en ese caso se obtendría la misma solución para $\hat{\beta}_1$ y $\hat{\beta}_2$ que en el MRLS inlcuyendo solamente el término constante y $F2 \equiv X_2$ y entonces las mismas estimaciones de esos coeficientes. Por lo tanto, en general no da lo mismo incluir o no otras variables en el modelo a la hora de estimar el efecto de una variable sobre la variable dependiente.

Interpretación de los coeficientes estimados.
El coeficiente estimado que acompaña a la variable F2, variable que recoge el tamaño total de la vivienda, es positivo y parece ser el signo adecuado. Si consideramos dos viviendas con el mismo número de baños y habitaciones, parece razonable pensar que aquella con mayor área habitable tenga un precio mayor. Esto indica que las habitaciones serán más grandes.

Los signos de los coeficientes asociados a BEDRMS y BATHS son negativos. Podemos pensar que si aumenta el número de habitaciones o el número de baños, esto indicaría una vivenda más lujosa y por lo tanto debería de aumentar el valor de la vivienda. Pero hay que tener en cuenta que a la hora de interpretar un coeficiente de regresión asociado a uno de los regresores estamos manteniendo constante el resto de variables explicativas.

Si la misma superficie habitable se tiene que dividir para poder incluir una nueva habitación, el resultado será que cada habitación será más pequeña. El signo del coeficiente estimado indica que un comprador medio valora negativamente tener más habitaciones a costa de un menor tamaño de éstas. Lo mismo se puede interpretar en el caso del coeficiente que acompaña a BATHS.

Interpretación de los coeficientes estimados:

- El coeficiente estimado $\hat{\beta}_1 = 129,062$ indica el precio medio estimado en miles de euros, de aquellas viviendas que no tienen ningún pie cuadrado de área habitable, ni habitaciones ni baños.

- El coeficiente estimado $\hat{\beta}_2 = 0,154800$:

  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 estima que en media su precio de venta se incremente en 154.800 dólares.

- El coeficiente estimado $\hat{\beta}_3 = -21,5875$:

  Si aumenta el número de habitaciones, manteniendo constante el tamaño de la vivienda y el número de baños, el precio medio se estima disminuirá en 21.588 dólares.

- El coeficiente $\hat{\beta}_4 = -12,1928$:

  Manteniendo el tamaño de la vivienda y el número de habitaciones constante, añadir un baño completo más significa tener habitaciones más pequeñas, por lo que el precio medio se estima disminuirá en 12.193 dólares.

¿Se mantendría el signo del coeficiente que acompaña a BEDRMS si no incluimos la variable F2 ni BATHS?
Pues seguramente no, porque en ese caso no estamos controlando por esa variable en la regresión, y como hemos visto F2 y BEDRMS están correlacionados. Por lo tanto más habitaciones implicaría mayor superficie de piso, y por lo tanto más precio en media. Lo mismo ocurriría si solamente incluimos BATHS. Ahora bien, ¿qué ocurriría si excluimos solamente F2 y dejamos las otras dos variables explicativas? Veremos las implicaciones que tiene omitir o no controlar por variables relevantes en un tema posterior.

Estimación del incremento medio en el precio de la vivienda ante cambios en las variables explicativas.
Utilizando los resultados (3.3) de la estimación del modelo (3.1), si manteniendo el número de baños tenemos dos habitaciones más y aumenta el área habitable en 500 pies cuadrados, el cambio en el precio medio estimado de una vivienda será de 34.224 dólares, esto es

$$\widehat{\triangle \mathrm{P}}_i = 0,1548 \triangle \mathrm{F2}_i - 21,588 \triangle \mathrm{BEDRMS}_i = (0,1548 \times 500) - (21,588 \times 2) = 34,224$$

3.3.2. Desviaciones típicas e intervalos de confianza

Por el momento nos hemos centrado en la interpretación de las estimaciones puntuales. Pero también tenemos que tener en cuenta que estas estimaciones son realizaciones muestrales de un estimador, que es una variable aleatoria. Por lo tanto, pueden estar sujetas a variación muestral ya que distintas muestras puedan dar lugar a distintas realizaciones muestrales. Estas estimaciones de un mismo vector de parámetros β estarán distribuidas con mayor o menor variación alrededor de su valor poblacional siguiendo cierta distribución de probabilidad.

Bajo las hipótesis básicas que hemos enumerado al principio de este tema, el valor poblacional del vector de parámetros β es la media de la distribución ya que $\hat{\beta}_{MCO}$ es un estimador insesgado. Su distribución es una Normal y la matriz de varianzas y covarianzas viene dada por la expresión $V(\hat{\beta}_{MCO}) = \sigma^2(X'X)^{-1}$. Esto se suele denotar como

$$\hat{\beta}_{MCO} \sim N(\beta, \sigma^2(X'X)^{-1}) \tag{3.4}$$

La varianza de las perturbaciones, σ^2, es un parámetro desconocido. Un estimador insesgado de la misma bajo las hipótesis básicas es

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

donde $\hat{u} = Y - X\hat{\beta}_{MCO}$ es el vector de residuos. El programa, en la ventana gretl:modelo1 muestra las realizaciones muestrales de la suma de cuadrados de los residuos (SCR), $\hat{u}'\hat{u} = 16700,1$ y de la desviación típica de los residuos $\sqrt{\hat{\sigma}^2} = 40,8657$.

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

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

En la ventana de resultados de la estimación del modelo por MCO, **gretl:modelo1**, podemos obtener la realización muestral de este estimador $\hat{V}(\hat{\beta}_{MCO}) = \hat{\sigma}^2(X'X)^{-1}$ eligiendo:

Análisis → Matriz de covarianzas de los coeficientes

Se abre una nueva ventana, **gretl:covarianzas de los coeficientes**, donde se muestra la estimación de las varianzas (elementos de la diagonal principal) y covarianzas (elementos fuera de la diagonal principal) de los coeficientes de regresión $\hat{\beta}$, como se muestra en la Tabla 3.2. Dado que es una matriz simétrica, solamente aparecen los valores por encima de la diagonal principal. La raíz cuadrada de los elementos de la diagonal principal son los mismos

Matriz de covarianzas de los coeficientes

const	F2	BEDRMS	BATHS	
7797,47	0,670891	−1677,1	−1209,3	const
	0,00102019	−0,0754606	−0,995066	F2
		730,585	−356,40	BEDRMS
			1870,56	BATHS

Tabla 3.2: Modelo (3.1). Estimación de la matriz de covarianzas de $\hat{\beta}$

valores que los mostrados en la tercera columna de la ventana **gretl:modelo1**. Por ejemplo, la varianza estimada del coeficiente $\hat{\beta}_2$ asociado a F2 es $\widehat{var}(\hat{\beta}_2) = 0,00102019$ y su raíz cuadrada es su desviación típica estimada $\widehat{des}(\hat{\beta}_2) = 0,0319404$.

También podemos obtener estimaciones de las covarianzas entre los coeficientes estimados. Por ejemplo, la covarianza estimada entre los coeficientes $\hat{\beta}_2$ asociado a F2 y $\hat{\beta}_4$ asociado a $BATHS$ es igual a $\hat{cov}(\hat{\beta}_2, \hat{\beta}_4) = -0,995066$.

Intervalos de confianza:

Seguidamente vamos a ver cómo podemos obtener intervalos de confianza para cada coeficiente individual. ¿Qué nos indican estos intervalos? ¿Cuál es su utilidad?

Bajo las hipótesis básicas, se puede demostrar que la variable aleatoria

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{des}(\hat{\beta}_j)} \sim t(N-K) \tag{3.5}$$

donde $\widehat{des}(\hat{\beta}_j)$ es la desviación típica estimada del estimador $\hat{\beta}_j$ y $t(N-K)$ denota la distribución t de Student de $(N-K)$ grados de libertad. Esto es válido para cualquiera de los coeficientes β_j , $j = 1, \ldots, K$.

Denotamos por $c = t_{(N-K)\alpha/2}$ la ordenada de la distribución t de Student con $N-K$ grados de libertad, tal que deja a la derecha una probabilidad de $\alpha/2$, esto es $P(t > c) = \alpha/2$. Esto implica que:

$$Pr\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\widehat{des}(\hat{\beta}_j)} \leq c\right) = Prob\left(\hat{\beta}_j - c\,\widehat{des}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c\,\widehat{des}(\hat{\beta}_j)\right) = 1-\alpha \tag{3.6}$$

Por lo tanto, un intervalo de confianza del $(1-\alpha)$ por ciento para un coeficiente cualquiera β_j viene dado por

$$IC(\beta_j)_{1-\alpha} = \left[\hat{\beta}_j \pm c\,\widehat{des}(\hat{\beta}_j)\right]$$

El cálculo de los intervalos de confianza para los coeficientes de regresión del modelo se conoce con el nombre de **estimación por intervalo**. Un intervalo de confianza nos dice que, con

probabilidad $(1-\alpha)$ se estima que el parámetro β_j estará dentro de ese rango de valores. Este intervalo puede ser demasiado amplio, y esto dependerá de la precisión con la que estimemos los parámetros recogido en $\widehat{des}(\hat{\beta}_j)$. Es importante tener en cuenta que la validez de estos intervalos de confianza depende de que se satisfagan las hipótesis básicas.

Siguiendo con el ejemplo del modelo (3.1) para el precio de la vivienda, Gretl nos permite obtener directamente los intervalos de confianza del 95 por ciento para los coeficientes. El resultado mostrado en la Tabla 3.3 se obtiene eligiendo en la ventana **gretl:modelo1**

Análisis → Intervalos de confianza para los coeficientes

Variable	Coeficiente	Intervalo de confianza 95 %	
		bajo	alto
const	129,062	−67,690	325,814
F2	0,154800	0,0836321	0,225968
BEDRMS	−21,587	−81,812	38,6376
BATHS	−12,192	−108,56	84,1742

Tabla 3.3: Modelo (3.1): Estimación por intervalo de los coeficientes.

A su vez, utilizando los resultados mostrados en la ventana **gretl:modelo1**

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007***
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

podemos obtener intervalos de confianza para cada uno de los coeficientes, dado un nivel de confianza $(1-\alpha)$, por ejemplo del 95 por ciento[5]. Los intervalos de confianza obtenidos son:

$$\begin{aligned} \beta_1: &\quad 129,0620 \pm (2,228 \times 88,3033) \\ \beta_2: &\quad 0,1548 \pm (2,228 \times 0,0319404) \\ \beta_3: &\quad -21,5875 \pm (2,228 \times 27,0293) \\ \beta_4: &\quad -12,1928 \pm (2,228 \times 43,2500) \end{aligned}$$

El intervalo de confianza además se puede utilizar para contrastar la hipótesis de que el parámetro β_j tome determinado valor. Si el valor del parámetro bajo la hipótesis nula

[5] Al 95 por ciento de confianza, $(\alpha/2 = 0,025)$, el valor en las tablas de la distribución t de Student con 10 grados de libertad es $c = t_{(10)0,025} = 2,228$. Recordar que Gretl permite acceder a algunos valores tabulados de distintas distribuciones, Normal, t-Student, Chi-cuadrado, F de Snedecor. En la ventana principal **gretl** en *Herramientas → Tablas estadísticas*. En el caso de la t de Student hay que introducir los grados de libertad (gl). Los valores mostrados corresponden a los valores de $\alpha/2$ de 0,10-0,05-0,025-0,01-0,001.

está dentro del intervalo de confianza, no podemos rechazar esa hipótesis al nivel de significación α. Dada la muestra y nuestra especificación del modelo, no podemos rechazar con una confianza del 95 por ciento, excepto para el parámetro asociado a F2, que el coeficiente asociado a cada una de estas variables sea igual a cero ya que este valor está dentro del intervalo de confianza. ¿Quiere decir entonces que el valor poblacional de cada uno de esos parámetros es cero? La respuesta es NO, ya que por esa misma regla de tres el parámetro β_j debería de tomar cada uno de los valores en el intervalo.

3.3.3. Significatividad individual y conjunta

Contrastes de significatividad individual

Uno de los principales objetivos de un primer análisis de regresión es la de contrastar si son o no estadísticamente relevantes los factores que hemos considerado como explicativos de la variable dependiente en cuestión, dada la especificación de nuestro modelo. Podemos considerar individualmente cada regresor y contrastar:

$$H_0\colon\ \beta_j = 0$$
$$H_a\colon\ \beta_j \neq 0$$

donde la hipótesis nula implica que, dada la especificación del modelo una vez se ha controlado por el resto de factores incluidos como variables explicativas, el efecto marginal de la variable X_j sobre el valor medio de la variable dependiente es cero.

Dado que en la hipótesis alternativa se contempla la posibilidad de que el coeficiente, de ser distinto de cero, pueda ser indistintamente negativo o positivo, el contraste es a dos colas. Normalmente en estos contrastes, conocidos con el nombre de contrastes de significatividad individual, se considera esta alternativa.

El estadístico de contraste y su distribución bajo la hipótesis nula es:

$$t_j = \frac{\hat{\beta}_j}{\widehat{des}(\hat{\beta}_j)} \overset{H_0}{\sim} t(N-K) \tag{3.7}$$

Una vez obtenido el valor muestral del estadístico, t_j^m, ¿cómo decidimos si rechazar o no la hipótesis nula?

- Se elige un nivel de significación α que indicaría nuestra elección de la probabilidad de error de tipo I (rechazar la hipótesis nula cuando esta fuera cierta) o tamaño del contraste. Obtenemos el valor crítico o umbral $c = t_{(N-K)\alpha/2}$ tal que $Pr(t_j > c) = \alpha/2$.
- Rechazamos la hipótesis nula a un nivel de significación α, si en valor absoluto la realización muestral del estadístico es mayor que el valor crítico $|t_j^m| > c$. No rechazamos la hipótesis nula en caso contrario.

Si no se rechaza la hipótesis nula, en el lenguaje econométrico se dice que la variable que acompaña al coeficiente en cuestión no es significativa o que el coeficiente no es significativamente distinto de cero al α por ciento de significación. Si por el contrario se rechaza la hipótesis nula, se dice que la variable es significativa o que el coeficiente es significativamente distinto de cero.

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

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

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 $Pr(t_j > t_j^m|H_0)$. 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.

¿Cuál será la regla de decisión del contraste mirando al *valor-p*?

Rechazar la hipótesis nula si el *valor-p* es menor que el nivel de significación elegido y no rechazarla en caso contrario.

Esta es exactamente la misma regla de decisión que antes. Elegido un nivel de significación, si el valor muestral es mayor en valor absoluto que el valor crítico c, querrá decir que dos veces la probabilidad que deja a la derecha el valor muestral es más pequeño que ese nivel de significación.

Siguiendo con nuestro ejemplo, vamos a comentar qué nos indican la cuarta y quinta columna que aparecían en la ventana de resultados de la estimación por MCO del modelo (3.1) **gretl:modelo1**.

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007***
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

Los valores obtenidos en la cuarta columna se obtienen de dividir los correspondientes valores de la segunda y tercera columnas esto es, la estimación del coeficiente dividida por su desviación típica estimada. Esta sería la realización muestral del estadístico t_j bajo la hipótesis nula de que el valor poblacional del parámetro β_j asociado a esa variable es igual a cero.

La quinta columna es el *valor-p* asociado a cada coeficiente, siendo el contraste de significatividad individual a dos colas. Habitualmente se eligen como niveles de significación el 1 %, 5 % y 10 % siendo el 5 % el más utilizado. Gretl indica con uno, dos o tres asteriscos cuando se rechaza la hipótesis nula al 10 %, al 5 %, o al 1 % respectivamente.

En este caso solamente es significativa la variable F2 al 1 % y se indica con tres asteriscos. El *valor-p* asociado a esta variable es más pequeño que 0,01 y por lo tanto que 0,05 y que 0,1.

Para el resto de coeficientes no se rechazaría la hipótesis nula. Los coeficientes asociados al término constante, BEDRMS y BATHS no serían significativamente distintos de cero ni

siquiera al 10 %. El *valor-p* asociado es mayor que 0,1. Estos valores oscilan entre 0,175 y 0,784 por lo que, si rechazásemos la hipótesis nula de que cada uno de estos coeficientes es cero, habría desde un 17,5 a un 78,4 por ciento de probabilidad de cometer el error de rechazar esa hipótesis siendo cierta.

Si miramos a los valores críticos en cada uno de estos niveles de significación tenemos que:

$$\begin{array}{ll} \alpha = 0,01 & t_{(10)0,005} = 3,169 \\ \alpha = 0,05 & t_{(10)0,025} = 2,228 \\ \alpha = 0,1 & t_{(10)0,05} = 1,812 \end{array}$$

Excepto en el caso de la variable F2, el valor muestral de los estadísticos t_j en valor absoluto es más pequeño que cualquiera de estos valores críticos. Por lo tanto solamente se rechaza la hipótesis nula de que el coeficiente asociado a la variable SQFT sea igual a cero. Esto parece indicar que dado que el número de habitaciones y de baños está ya recogido en el tamaño de la vivienda, una vez incluimos esta variable el tener más o menos habitaciones o baños no tiene un efecto marginal significativo en el precio medio de ésta. Lo normal es tener una vivienda con un número de habitaciones y baños proporcional a su tamaño.

Esto mismo concluimos mirando a los intervalos de confianza, aunque en ese caso el nivel de significación elegido sólo fue del 5 por ciento.

Contraste de significación conjunta

Otro estadístico que se muestra en la ventana de resultados de la estimación es el valor del estadístico F (3, 10) = 16,9889 con *valor-p* = 0, 000299. ¿Cómo se calcula este estadístico? ¿Qué hipótesis nula se está contrastando?

La hipótesis nula que se está contrastando es que conjuntamente todos lo coeficientes, excepto el asociado al término constante, sean cero. En nuestro ejemplo en concreto

$$\begin{array}{l} H_0\colon\ \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a\colon\ \text{alguna de las igualdades no se cumple} \end{array}$$

Este estadístico se puede considerar como un contraste general de bondad de ajuste del modelo. Si la hipótesis nula no se rechaza podemos concluir que ninguna de las variables en conjunto puede explicar la **variación** en el precio de la vivienda. Esto significa que es un modelo muy pobre y que debiera de ser reformulado.

Estamos excluyendo de la hipótesis nula el parámetro que acompaña al término constante. El modelo bajo la hipótesis nula, al que llamaremos Modelo Restringido es:

$$\texttt{Modelo Restringido} \qquad P_i = \beta_1 + u_i \qquad\qquad i = 1, 2, \ldots, N \tag{3.8}$$

Este modelo incluye solamente un término constante como regresor y le compararemos con el Modelo No Restringido (3.1). El estimador MCO del parámetro β_1 en el modelo restringido es aquél que

$$\min_{\hat{\beta}_1} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1)^2$$

En este caso tenemos solamente un parámetro a estimar por lo que sólo hay una ecuación normal,

$$\sum_i Y_i = N\hat{\beta}_1 \tag{3.9}$$

cuya solución es

$$\hat{\beta}_{1,R} = \frac{1}{N}\sum_i Y_i = \bar{Y}$$

El coeficiente estimado que acompaña al término constante nos recoge simplemente la media muestral de la variable dependiente. El residuo correspondiente al modelo restringido es $\hat{u}_{i,R} = Y_i - \hat{\beta}_{1,R} = Y_i - \bar{Y}$, por lo que la suma de cuadrados residual coincide con la suma de cuadrados total o variación total de la variable dependiente. Esto implica que la suma de cuadrados explicada o variación explicada con la estimación de este modelo (3.8) es nula

$$SCR_R = \sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \overline{Y})^2 = SCT \qquad \Rightarrow \qquad SCE_R = 0$$

Por último, y teniendo en cuenta como se define el coeficiente de determinación R^2

$$R^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2}$$

para este modelo el coeficiente de determinación es igual a cero[6]. Dado que en el modelo solamente incluimos un regresor que no varía, éste no puede explicar **variación** o varianza de la variable dependiente. Si estimamos con Gretl el modelo (3.8) obtenemos los siguientes resultados:

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	317,493	23,6521	13,4234	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	101815,
Desviación típica de los residuos ($\hat{\sigma}$)	88,4982
R^2	0,000000
$\bar{R}^2$ corregido	0,000000
Grados de libertad	13
Log-verosimilitud	−82,108
Criterio de información de Akaike	166,216
Criterio de información Bayesiano de Schwarz	166,855
Criterio de Hannan–Quinn	166,157

[6]Esto es así dado que $\sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \bar{Y})^2 \Rightarrow R_R^2 = 1 - \frac{\sum_i \hat{u}_{i,R}^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - 1 = 0$.

Podemos comprobar que la estimación del coeficiente que acompaña al término constante coincide con la media muestral de la variable dependiente ($\bar{P} = 317,493$). La desviación típica de los residuos coincide con la desviación típica de la variable dependiente, ya que la suma de cuadrados residual coincide con la suma de cuadrados total, $SCR_R = \sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \overline{Y})^2 = 101815$, y también los grados de libertad de ambas, $T - K = T - 1 = 13$. Por lo tanto,

$$\sqrt{\frac{\sum_i \hat{u}_{i,R}^2}{13}} = \sqrt{\frac{\sum_i (Y_i - \overline{Y})^2}{13}} = 88,4982$$

Por último, el coeficiente de determinación R^2 es igual a cero.

Un estadístico general de contraste de restricciones lineales es aquél que compara las sumas de cuadrados de residuos de la estimación del modelo restringido y del modelo no restringido, teniendo en cuenta los grados de libertad en la estimación de cada modelo,(gl_R) y (gl_{NR}) respectivamente[7]

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-K)} \overset{H_0}{\sim} \mathcal{F}(q,\, N-K) \tag{3.10}$$

donde $q = (gl_R - gl_{NR})$ es el número de restricciones bajo la hipótesis nula y $N - K = gl_{NR}$. Si dividimos numerador y denominador por la suma de cuadrados total SCT y utilizamos los siguientes resultados:

a) $1 - R^2 = SCR_{NR} \,/\, SCT$ y en este caso $1 - R_R^2 = 1 - 0 = 1$.

b) $gl_R - gl_{NR} = (N-1) - (N-K) = K - 1$ que es el número de restricciones bajo la hipótesis nula.

el estadístico general (3.10) nos queda para este contraste en concreto igual a

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} = \frac{R^2}{(1-R^2)}\frac{(N-K)}{(K-1)} \overset{H_0}{\sim} \mathcal{F}(K-1,\, N-K) \tag{3.11}$$

En nuestro ejemplo sobre el precio de la vivienda, $K - 1 = 3$ que es el número de restricciones bajo la hipótesis nula y $N - K = 14 - 4 = 10$. Dado el resultado mostrado $F(3, 10) = 16,9889$ (valor p $= 0,000299$), si consideramos el valor-p se rechazaría la hipótesis nula a cualquier nivel de significación razonable, en particular al $\alpha = 0,05$ ya que este valor es mayor que el *valor-p* obtenido. Si utilizamos el valor crítico $\mathcal{F}_{(3,10)0,05} = 3,71$ obtenemos el mismo resultado ya que el valor muestral del estadístico es mayor que el valor crítico. Esto indica que al menos uno de los coeficientes, aparte del asociado al término constante, es distinto de cero.

Aunque hemos utilizado en esta sección el coeficiente de determinación en relación al estadístico de significación conjunta, en la siguiente sección vamos a hablar de su utilización junto con el coeficiente de determinación corregido y otros estadísticos para la selección entre distintos modelos.

[7] En temas posteriores veremos la utilización de este estadístico para contrastar otro tipo de restricciones lineales.

3.4. Bondad de ajuste y selección de modelos

En los temas anteriores se ha presentado el coeficiente de determinación como una medida de bondad de ajuste que es invariante a unidades de medida[8]. Este coeficiente se define como la proporción de variación explicada por la regresión del total de variación a explicar en la muestra de la variable dependiente. Si hay término constante en el modelo,

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2} \qquad 0 \leq R^2 \leq 1$$

Este indicador tiene que ser considerado como uno más a tener en cuenta a la hora de valorar si un modelo es adecuado, pero no debemos darle más importancia de la que tiene. Obtener un valor del R^2 cercano a 1 no indica que nuestros resultados puedan ser fiables. Por ejemplo, podemos tener problemas de no satisfacerse alguna hipótesis básica y nuestra inferencia no ser válida.

Por otro lado, obtener un valor más o menos alto del coeficiente de determinación puede estar influido por el tipo de datos que estemos analizando. Normalmente con datos de series temporales, donde las variables pueden presentar tendencias similares en el tiempo, es fácil obtener R^2 altos, mientras que con datos de sección cruzada eso no suele ocurrir ya que normalmente las variables presentan mayor dispersión.

Por otro lado, si queremos utilizar el R^2 para comparar distintos modelos, estos deben de tener la misma variable dependiente ya que así tendrán igual suma de cuadrados total. Aún así, esta medida adolece del problema de aumentar su valor al añadir una nueva variable explicativa, sea cual sea su aportación al modelo. Además no tiene en cuenta que hay que estimar un nuevo parámetro con el mismo número de observaciones.

Para tener en cuenta este problema se suele utilizar el R^2 corregido por grados de libertad. Esta medida tiene en cuenta los grados de libertad tanto de la suma de cuadrados residual, $(N - K)$, como de la suma de cuadrados total, $(N - 1)$. Se define como

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(N-K)}{\sum (Y_i - \bar{Y})^2/(N-1)} = 1 - \frac{N-1}{N-K}(1 - R^2) \qquad -\infty < \bar{R}^2 \leq R^2$$

El $\bar{R}^2$ puede disminuir si el incluir una nueva variable no compensa la pérdida de grados de libertad al tener que estimar un nuevo parámetro[9]. El coeficiente de determinación corregido $\bar{R}^2$ no tomará valores mayores que el R^2 pero sí puede tomar valores negativos. Esto último indicaría que el modelo no describe adecuadamente el proceso que ha generado los datos.

Hasta el momento hemos ido comentado los resultados que normalmente se muestran en la estimación de un modelo. Una forma de presentarlos es la siguiente:

$$\underset{(estad.\,t)}{\widehat{\mathrm{P}}} = \underset{(1{,}462)}{129{,}062} + \underset{(4{,}847)}{0{,}154800}\,\mathrm{F2} - \underset{(-0{,}799)}{21{,}5875}\,\mathrm{BEDRMS} - \underset{(-0{,}282)}{12{,}1928}\,\mathrm{BATHS}$$

$$N = 14 \quad R^2 = 0{,}8359 \quad \bar{R}^2 = 0{,}7868 \quad F(3, 10) = 16{,}989$$

[8] Esto no ocurre con otras medidas como puede ser la desviación típica de los residuos, $\hat{\sigma} = \sqrt{SCR/N-K)}$ ya que la suma de cuadrados de los residuos no es invariante a un cambio de escala en las variables.

[9] Se puede demostrar que si el valor absoluto del estadístico t de significatividad individual asociado a una variable es menor que la unidad, eliminar esta variable del modelo aumentará el $\bar{R}^2$ mientras que si es mayor que la unidad lo reducirá.

Una alternativa a presentar los estadísticos t de significatividad individual, aunque suele ser lo más habitual, es mostrar las desviaciones típicas estimadas de los coeficientes o los valores p correspondientes.

Otros criterios de selección de modelos que muestra Gretl son los criterios de información de Akaike (AIC), 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.

Selección de un modelo para el precio de la vivienda.

Vamos a continuar con nuestro ejemplo sobre el precio de la vivienda y comparar distintas especificaciones, para seleccionar una especificación entre varias propuestas. Para ello, utilizamos distintos indicadores que hemos visto hasta ahora, significatividad individual, conjunta, coeficientes de determinación y criterios de información. Podemos considerar que estos indicadores nos ayudan a valorar la especificación en términos de la contribución de las variables explicativas incluidas en el modelo[10].

Vamos a estimar las siguientes especificaciones o modelos alternativos para explicar el precio de la vivienda:

Modelo 1 $P_i = \beta_1 + \beta_2\, F2_i + u_i$

Modelo 2 $P_i = \beta_1 + \beta_2\, F2_i + \beta_3\, BEDRMS_i + u_i$

Modelo 3 $P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$

Modelo 4 $P_i = \beta_1 + \beta_3\, BEDRMS_i + \beta_4\, BATHS_i + u_i$

Estos cuatro modelos difieren en las variables explicativas incluidas. El **Modelo 3** es el más general e incluye al resto de modelos. Esto quiere decir que cada uno de los restantes se obtiene imponiendo una o más restricciones sobre los coeficientes de este modelo. En este caso son restricciones de exclusión, es decir que algún coeficiente o coeficientes son iguales a cero. A este tipo de modelos se les llama modelos anidados. Los resultados de la estimación del **Modelo 3** con Gretl son los siguientes:

Modelo 3: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATH	−12,192	43,2500	−0,2819	0,7838

[10]Estos no son los únicos indicadores. Por ejemplo, analizar el gráfico de residuos o utilizar diversos contrastes de algunas de las hipótesis básicas son elementos importantes a la hora de evaluar los resultados de la especificación y estimación de un modelo.

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16700,1
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657
R^2	0,835976
$\bar{R}^2$ corregido	0,786769
$F(3, 10)$	16,9889
valor p para $F()$	0,000298587
Log-verosimilitud	−69,453
Criterio de información de Akaike	146,908
Criterio de información Bayesiano de Schwarz	149,464
Criterio de Hannan–Quinn	146,671

El **Modelo 1** es el más reducido y también está incluido en los modelos 2 y 3, no así en el 4. Estos son los resultados de su estimación:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	52,3509	37,2855	1,4041	0,1857
F2	0,138750	0,0187329	7,4068	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	18273,6
Desviación típica de los residuos ($\hat{\sigma}$)	39,0230
R^2	0,820522
$\bar{R}^2$ corregido	0,805565
Grados de libertad	12
Log-verosimilitud	−70,084
Criterio de información de Akaike	144,168
Criterio de información Bayesiano de Schwarz	145,447
Criterio de Hannan–Quinn	144,050

El **Modelo 2** está anidado en el 3. Los resultados de la estimación de este modelo se muestran a continuación:

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	121,179	80,1778	1,5114	0,1589
F2	0,148314	0,0212080	6,9933	0,0000
BEDRMS	−23,910	24,6419	−0,9703	0,3527

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16832,8
Desviación típica de los residuos ($\hat{\sigma}$)	39,1185
R^2	0,834673
$\bar{R}^2$ corregido	0,804613
$F(2, 11)$	27,7674
valor p para $F()$	5,02220e-05
Log-verosimilitud	−69,509
Criterio de información de Akaike	145,019
Criterio de información Bayesiano de Schwarz	146,936
Criterio de Hannan–Quinn	144,841

Finalmente el **Modelo 4** solamente está anidado en el modelo 3. Los resultados de la estimación por MCO son:

Modelo 4: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	27,2633	149,652	0,1822	0,8588
BEDRMS	−10,137	46,9811	−0,2158	0,8331
BATHS	138,795	52,3450	2,6515	0,0225

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	55926,4
Desviación típica de los residuos ($\hat{\sigma}$)	71,3037
R^2	0,450706
$\bar{R}^2$ corregido	0,350834
$F(2, 11)$	4,51285
valor p para $F()$	0,0370619
Log-verosimilitud	−77,914
Criterio de información de Akaike	161,829
Criterio de información Bayesiano de Schwarz	163,746
Criterio de Hannan–Quinn	161,651

Comparación de los resultados para los modelos 1,2 y 3.

- Se observa que a medida que se introducen más variables explicativas, la suma de cuadrados residual va disminuyendo y el coeficiente de determinación R^2 aumenta.

- En términos del coeficiente de determinación R^2, en el **Modelo 1** el tamaño de la vivienda (F2) explica el $82,1\,\%$ de la variación en los precios de la vivienda, pasando a ser de un $83,6\,\%$ al incluir el número de habitaciones (BEDRMS) y número de baños (BATHS).

- A medida que se incluyen más variables explicativas, primero BEDRMS y luego BATHS, el coeficiente de determinación corregido $\bar{R}^2$ disminuye y la desviación típica de los residuos aumenta[11]. Esto indica que la ganancia en un mayor valor del R^2 o menor suma de cuadrados residual no se compensa en ningún caso por la pérdida de grados de libertad.

- En cuanto a la significatividad individual, en los tres modelos la única variable significativa a los niveles de significación habituales es F2[12]. Así, una vez hemos controlado por el tamaño de la vivienda, las variables BEDRMS y BATHS no afectan significativamente el precio de la vivienda.

- El estadístico F de significación conjunta señala en los tres casos no aceptar la hipótesis nula de que todos los coeficientes excepto el asociado al término constante son igual a cero. Al menos hay un coeficiente que es significativamente distinto de cero. Por lo obtenido en los contrastes de significatividad individual, sabemos que éste es el coeficiente que acompaña a F2.

  Si nos fijamos, a medida que vamos del **Modelo 1** al **3**, el valor muestral del estadístico F disminuye. Esto es lógico, ya que este estadístico es función del R^2 pero también de los grados de libertad. Otra vez estaría recogiendo que, a medida que aumenta el número de parámetros a estimar K, las diferencias en R^2 son demasiado pequeñas para compensar la disminución en el ratio $(N-K)/(K-1)$. Ahora bien, en general, las diferencias en el estadístico F no son relevantes. Lo que es de interés es el resultado del contraste.

- Si consideramos los criterios de información AIC, BIC y HQC, de los tres modelos el elegido es el **Modelo 1**, reafirmando lo que indica el $\bar{R}^2$. La ganancia en un mejor ajuste, o una menor suma de cuadrados residual, no es suficiente para compensar el factor que penaliza en función de grados de libertad.

  Dado que el tamaño de la vivienda depende del número de habitaciones y de baños, este resultado parece indicar que una vez se controla por F2 indirectamente esta variable incluye casi todo lo que pueden aportar BEDRMS y BATHS.

¿Qué ocurre con el Modelo 4?

En este modelo no hemos incluido la variable F2, que en el análisis anterior era la variable que más explica el precio de la vivienda y hemos dejado las variables que no eran significativas una vez que incluíamos esta variable. Podríamos argumentar que de esta forma se podría analizar el efecto de BEDRMS y BATHS, ya que F2 parecía recoger la información relevante de estas dos variables.

Si lo comparamos con el **Modelo 3**, que es en el que está anidado el **Modelo 4**, se obtiene menor valor de R^2 y $\bar{R}^2$, mayor valor de AIC, BIC y HQC, mayor suma de cuadrados residual y mayor desviación típica de los residuos. Todos ellos señalan en la misma dirección siendo, en términos de estos criterios, peor modelo el 4. Vemos que el omitir F2 empeora mucho

[11] Notar que los estadísticos t asociados a cada coeficiente son menores que uno en valor absoluto.

[12] Por ejemplo, con nivel de significación del 5 por ciento los valores críticos serían para el modelo **1** $t_{(12)0,025} = 2,179$, para el **Modelo 2** $t_{(11)0,025} = 2,201$ y para el **Modelo 3** $t_{(10)0,025} = 2,228$.

el ajuste sin compensar por la ganancia en grados de libertad. Además cambia sustancialmente la estimación y la significatividad del coeficiente que acompaña a BATHS, pasando la estimación de signo positivo a negativo y ser significativamente distinto de cero al 5 % de significación. ¿Qué puede estar ocurriendo? ¿Serán esta estimación y este contraste fiables si hemos omitido una variable que parece ser relevante? ¿Se verán afectadas las propiedades del estimador MCO por esta omisión? Todo esto lo veremos en el tema de error de especificación.

Bibliografía

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.

Tema 4

Contrastes de restricciones lineales y predicción

Contenido

4.1. Contrastes de restricciones lineales

En el Tema 3 hemos estudiado la forma más común de realizar los contrastes de significatividad individual y el contraste de significatividad conjunta sobre los coeficientes que acompañan a las variables explicativas en un modelo de regresión lineal general. Estos contrastes son los más habituales y en general cualquier programa econométrico, como también es el caso de Gretl, muestra por defecto los valores de los estadísticos correspondientes para contrastar estas restricciones en el mismo output de estimación.

En ocasiones, además de éstas, también podemos estar interesados en contrastar hipótesis que implican otro tipo de restricciones lineales en los coeficientes poblacionales del modelo. En general, podemos denotar la hipótesis nula y la alternativa como:

$$H_0: \underset{(q\times K)}{R} \cdot \underset{(K\times 1)}{\beta} = \underset{(q\times 1)}{r}$$

$$H_a: R\beta \neq r$$

siendo q el número de restricciones bajo la hipótesis nula y K el número de parámetros en el modelo no restringido. La hipótesis alternativa implicaría que **al menos una** de las igualdades no se satisface[1].

Por ejemplo en el modelo sobre el precio de la vivienda que hemos visto ya en temas anteriores,

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \tag{4.1}$$

podemos expresar de esta forma los siguientes contrastes:

1. Contraste de significación individual de la variable $BEDRMS$: $H_0: \beta_3 = 0$

$$H_0: R\beta = r \;\Rightarrow\; \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 0$$

2. Contraste de significación conjunta: $H_0: \beta_2 = \beta_3 = \beta_4 = 0$

$$H_0: R\beta = r \;\Rightarrow\; \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

3. Contraste de un subconjunto de coeficientes igual a cero, por ejemplo los que acompañan a las variables BEDRMS y BATHS: $H_0: \beta_3 = \beta_4 = 0$

$$H_0: R\beta = r \;\Rightarrow\; \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

[1] Cuidado que esto no es lo mismo que **todas** las igualdades **no** se satisfagan.

Podemos ilustrar el interés de contrastar otro tipo de restricciones lineales en el siguiente modelo para la inversión agregada de un país,

$$INVERR_t = \beta_1 + \beta_2\, t + \beta_3 PNBR_t + \beta_4 INTERES_t + \beta_5 INFLACION_t + u_t \qquad (4.2)$$

donde las variables implicadas son:

INVERR:	Inversión agregada,, en términos reales.
t :	Tiempo $t = 1, 2, \ldots, T$
PNBR:	Producto Nacional Bruto, en términos reales.
INTERES:	Tipo de Interés nominal.
INFLACION:	Tasa de Inflación.

Además de realizar los contrastes de significatividad individual y conjunta, podríamos estar interesados en contrastar las siguientes restricciones lineales:

1. $H_0 : \beta_3 = 1$, la propensión marginal a invertir es igual a 1, esto es, si aumenta el PNB real en una unidad, la inversión aumentará en la misma proporción, manteniendo el valor del resto de variables constante.

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

2. $H_0 : \beta_4 + \beta_5 = 0$, los inversores tienen en cuenta el tipo de interés real. Esto es, la inversión no variará si un aumento del tipo de interés nominal viene acompañado por un aumento de la misma magnitud de la tasa de inflación, manteniendo el resto de factores constantes.

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

3. $H_0 : \beta_2 = 0,\ \beta_3 = 1,\ \beta_4 + \beta_5 = 0$. Contraste conjunto de las dos restricciones anteriores además de la restricción de que la inversión en media no presenta una tendencia lineal.

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

El siguiente estadístico, conocido como estadístico F de Wald, se puede utilizar para contrastar una o más restricciones lineales en el contexto de un MRLG. Esta forma de realizar el contraste solamente requiere estimar el modelo sin restringir.

Como ya hemos visto en el Tema 3, bajo las hipótesis básicas la distribución del estimador MCO del modelo sin restringir es: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$. Por lo tanto, dado que R es una matriz de constantes de rango q, se tiene que **bajo la hipótesis nula**:

$$\underset{(q\times 1)}{R\hat{\beta}} \sim \mathcal{N}(\underset{(q\times 1)}{r}, \underbrace{\sigma^2 R(X'X)^{-1}R'}_{(q\times q)}) \tag{4.3}$$

Utilizando este resultado y el estimador $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T-K}$ del parámetro σ^2, tenemos que el estadístico de contraste y su distribución bajo la hipótesis nula es el siguiente:

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{u}'\hat{u}/(T-K)} \overset{H_0}{\sim} \mathcal{F}(q, T-K) \tag{4.4}$$

Si no es cierta la H_0, la diferencia $(R\hat{\beta} - r)$ será grande por lo que el estadístico F tomará valores grandes en ese caso. Rechazaremos la H_0 con un nivel de significatividad α si el valor muestral del estadístico es mayor que el valor crítico, $F > \mathcal{F}(q, T-K)_\alpha$, no rechazando H_0 en caso contrario.

4.2. Contrastes utilizando Gretl

En esta sección vamos a utilizar Gretl para contrastar las restricciones vistas en los ejemplos anteriores utilizando ese estadístico. En general, una vez que hemos leído los datos de las variables de interés la forma de proceder es la siguiente:

- Especificar y estimar por MCO el modelo sin imponer las restricciones o **el modelo no restringido** en *Modelo* ⇒ *Mínimos cuadrados ordinarios*
- En la ventana donde se muestran los resultados de la estimación del modelo no restringido, **gretl: modelo1** elegir *Contrastes* ⇒ *Restricciones lineales*
- Dentro de la ventana que aparece *gretl: restricciones lineales* podemos escribir las restricciones a contrastar.

  Cada restricción del conjunto de restricciones tiene que ir en una línea como una ecuación, donde a la izquierda del signo igual tiene que ir la combinación lineal de los parámetros y a la derecha el valor numérico correspondiente. Los parámetros en la restricción se denotan de la forma `bJ` donde `J` representa la posición en la lista de regresores comenzando por `J`= 1. Lo que nosotros hemos denotado en el MRLG como β_1, coeficiente que normalmente, aunque no necesariamente, acompaña a la constante, en Gretl se denomina `b1`, nuestro β_2 es `b2`, β_3 es `b3` y así sucesivamente con todos los coeficientes del modelo.

En el **ejemplo del modelo para el precio de la vivienda**, que hemos utilizado en el Tema 3, vamos a contrastar la hipótesis de que conjuntamente variaciones en el número de habitaciones y el número de baños, manteniendo el tamaño de la vivienda constante, no influyen en el precio de la vivienda. Vamos a denotar los coeficientes como Gretl lo haría,

suponiendo que al especificar el modelo mantenemos el mismo orden en el listado de variables explicativas

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \tag{4.5}$$

Una vez estimado el modelo con *Modelo* ⇒ *Mínimos cuadrados ordinarios*, en la ventana de resultados de la estimación *gretl:modelo1* seleccionamos con el cursor

Contrastes ⇒ *Restricciones lineales*

Aparecerá la ventana *gretl: restricciones lineales*. Dentro de la ventana escribimos

```
b3=0
b4=0
```

Al seleccionar *Aceptar* en esta ventana obtenemos los siguientes resultados:

```
Conjunto de restricciones
 1: b[BEDRMS] = 0
 2: b[BATHS] = 0

Estadístico de contraste:

F(2, 10) = 0,471106, con valor p = 0,637492

Estimaciones restringidas:

 VARIABLE        COEFICIENTE       DESV.TÍP.          ESTAD T   VALOR P
  const               52,3509          37,2855          1,404   0,18565
  F2                   0,138750         0,0187329       7,407  <0,00001 ***
  BEDRMS               0,000000         0,000000          indefinido
  BATHS                0,000000         0,000000          indefinido

  Desviación típica de los residuos = 39,023
```

No rechazamos la hipótesis nula al nivel de significación por ejemplo del 5 % ya que el valor $p = 0,637492 > 0,05$. Si miramos a las tablas de la distribución F con 2 y 10 grados de libertad, eligiendo en la ventana principal de Gretl

Herramientas → *Tablas estadísticas* → *F con gln 2 y gld 10*

obtenemos la siguiente información,

```
Valores críticos aproximados de F(2, 10)
 10% en la cola derecha 2,92
  5%                4,10
  1%                7,56
```

De igual forma vemos que, para los tres niveles de significación del 1, 5 y 10 % no se rechaza la hipótesis nula, ya que el valor muestral del estadístico es menor que el valor crítico correspondiente. Además también se muestran las estimaciones del modelo restringido bajo esas dos restricciones. Notar que los coeficientes que acompañan a BEDRMS y BATHS son igual a cero y sus desviaciones típicas también. La razón es que esos coeficientes no son estimaciones ya que toman un valor dado conocido.

Cuando las restricciones a contrastar son simplemente de exclusión de uno o más regresores del modelo de partida, otra forma de llevar a cabo este contraste en Gretl es elegir en el menú de la ventana de estimación del modelo de partida,

Contrastes $\Rightarrow$ *Omitir variables*

Seguidamente en la ventana que surge, **gretl: contrastes del modelo**, se seleccionan las variables que acompañan a los coeficientes que bajo la hipótesis nula son cero. En el ejemplo en concreto que estamos viendo, sería elegir las variables BEDRMS y BATHS. Al pulsar *Aceptar* se muestra una nueva ventana con la estimación del modelo restringido bajo esas dos restricciones

$$P_i = \beta_1 + \beta_2 F2_i + u_i \tag{4.6}$$

que implican excluir de la regresión a BEDRMS y BATHS,

```
Modelo Restringido: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: P

  VARIABLE       COEFICIENTE       DESV.TÍP.      ESTAD T     VALOR P
  const           52,3509          37,2855        1,404      0,18565
  F2               0,138750         0,0187329     7,407     <0,00001 ***

  Media de la var. dependiente = 317,493
  Desviación típica de la var. dependiente. = 88,4982
  Suma de cuadrados de los residuos = 18273,6
  Desviación típica de los residuos = 39,023
  R-cuadrado = 0,820522
  R-cuadrado corregido = 0,805565
  Grados de libertad = 12
  Log-verosimilitud = -70,0842
  Criterio de información de Akaike (AIC) = 144,168
  Criterio de información Bayesiano de Schwarz (BIC) = 145,447
  Criterio de Hannan-Quinn (HQC) = 144,05

Comparación entre el modelo restringido y no restringido:

  Hipótesis nula: los parámetros de regresión son cero para las variables
    BEDRMS
    BATHS

Estadístico de contraste:  F(2, 10) = 0,471106, con valor p = 0,637492
```

La ventaja de realizar de esta forma el contraste es que, además de tener la estimación del modelo restringido (4.6), en esta nueva ventana tenemos otra vez todos los menús que Gretl ofrece para el análisis de esta nueva especificación[2].

En esta ventana también se muestra el resultado del contraste, esto es, el valor muestral del estadístico F que contrasta esas dos restricciones de exclusión, y el *valor-p*. Como se puede observar, el resultado que se obtiene es exactamente el mismo que el que se ofrece en la ventana **gretl: restricciones lineales**.

Seguidamente vamos a utilizar el ejemplo del modelo de la Función de Inversión, para ilustrar otro tipo de restricciones lineales que no sean simplemente de exclusión.

Escribimos el modelo no restringido

$$INVERR_t = \beta_1 + \beta_2\, t + \beta_3\ PNBR_t + \beta_4 INTERES_t + \beta_5 INFLACION_t + u_t \qquad (4.7)$$

y para el análisis usamos los datos de la siguiente Tabla[3]:

Año	PNB nominal	Inversión nominal	IPC	Tipo de Interés
1968	73,4	133,3	82,54	5,16
1969	944,0	149,3	86,79	5,87
1970	992,7	144,2	91,45	5,95
1971	1077,6	166,4	96,01	4,88
1972	1185,9	195,0	100,00	4,50
1973	1326,4	229,8	105,75	6,44
1974	1434,2	228,7	115,08	7,83
1975	1549,2	206,1	125,79	6,25
1976	1718,0	257,9	132,34	5,50
1977	1918,3	324,1	140,05	5,46
1978	2163,9	386,6	150,42	7,46
1979	2417,8	423,0	163,42	10,28
1980	2633,1	402,3	178,64	11,77
1981	2937,7	471,5	195,51	13,42
1982	3057,5	421,9	207,23	11,02

Tabla 4.1: Datos para el estudio de la Función de Inversión

Las series de Inversión y Producto Nacional Bruto en términos reales, INVERR y PNBR, se han obtenido de dividir las series nominales por el IPC con año base en 1972 y multiplicar por 10^{-1}, tal que están medidas en trillones de dólares. La tasa de inflación se ha calculado como el porcentaje de variación del IPC. Por lo tanto, los datos utilizados para estimar el modelo, son los de la siguiente tabla:

[2] El estimador restringido será $\hat{\beta}_R = \left[\ \hat{\beta}_{R,1}\ \ \hat{\beta}_{R,2}\ \ 0\ \ 0\ \right]'$ donde $\hat{\beta}_{R,1}$ y $\hat{\beta}_{R,2}$ son los obtenidos de la regresión excluyendo BEDRMS y BATHS.

[3] Corresponden a la Tabla F3.1 publicada en Greene (2008), p.1082 y disponible en: http://pages.stern.nyu.edu/∼wgreene/Text/econometricanalysis.htm. Fuente: Economic Report of the President, U.S. Government Printing Office, Washington, D.C., 1983. El IPC de 1967 es 79,06. El tipo de interés es el promedio anual de la tasa de descuento del Banco de la Reserva Federal de Nueva York.

Año	INVERR	PNBR	INFLACION	INTERES
1968	0,161	1,058	4,40	5,16
1969	0,172	1,088	5,15	5,87
1970	0,158	1,086	5,37	5,95
1971	0,173	1,122	4,99	4,88
1972	0,195	1,186	4,16	4,50
1973	0,217	1,254	5,75	6,44
1974	0,199	1,246	8,82	7,83
1975	0,163	1,232	9,31	6,25
1976	0,195	1,298	5,21	5,50
1977	0,231	1,370	5,83	5,46
1978	0,257	1,439	7,40	7,46
1979	0,259	1,479	8,64	10,28
1980	0,225	1,474	9,31	11,77
1981	0,241	1,503	9,44	13,42
1982	0,204	1,475	5,99	11,02

Tabla 4.2: Datos en términos reales

Primeramente creamos el fichero de datos a partir de la tabla anterior incluyendo la variable $t = 1, \ldots, 15$, con la opción de Gretl

Archivo → Nuevo conjunto de datos

Seguidamente estimamos por MCO el modelo no restringido arriba especificado, eligiendo en el menú *Modelo → Mínimos Cuadrados ordinarios* y obtenemos los siguientes resultados

Modelo 1: estimaciones MCO utilizando las 15 observaciones 1968–1982
Variable dependiente: INVERR

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−0,509071	0,0551277	−9,2344	0,0000
t	−0,0165804	0,00197176	−8,4089	0,0000
PNBR	0,670383	0,0549972	12,1894	0,0000
INTERES	−0,00232593	0,00121887	−1,9083	0,0854
INFLACION	−9,40107e-05	0,00134748	−0,0698	0,9458

Media de la var. dependiente	0,203333
D.T. de la variable dependiente	0,0341774
Suma de cuadrados de los residuos	0,000450812
Desviación típica de los residuos ($\hat{\sigma}$)	0,00671425
R^2	0,972433
$\bar{R}^2$ corregido	0,961406
$F(4, 10)$	88,1883
Estadístico de Durbin–Watson	1,96364
Coef. de autocorr. de primer orden	−0,0981367
Criterio de información de Akaike	−103,62
Criterio de información Bayesiano de Schwarz	−100,07

Contrastes de restricciones lineales:

a) Contraste de que la propensión marginal a invertir es la unidad, $H_0 : \beta_3 = 1$, frente a la hipótesis alternativa de que es distinto de la unidad. En la ventana **gretl: modelo1** seleccionamos *Contrastes* → *Restricciones lineales* y en la ventana que surge escribimos b3 = 1. Al aceptar se obtiene el siguiente resultado,

```
Restricción:
 b[PNBR] = 1

Estadístico de contraste:
F(1, 10) = 35,92, con valor p = 0,000133289

Estimaciones restringidas:
  VARIABLE      COEFICIENTE       DESV.TÍP.      ESTAD T      VALOR P
  const          -0,837112         0,0134320   -62,322    <0,00001 ***
  t              -0,0276707        0,00139136  -19,888    <0,00001 ***
  PNBR            1,00000          0,000000     indefinido
  INTERES        -0,00311914       0,00247563   -1,260     0,23377
  INFLACION      -0,000342359      0,00275183   -0,124     0,90323

Desviación típica de los residuos = 0,0137184
```

Se muestran también las estimaciones de los coeficientes del modelo restringido, donde se ha impuesto que el coeficiente que acompaña a PNBR es igual a la unidad. Como damos ese valor a β_3, no estamos estimando ese coeficiente, por lo tanto su desviación típica es cero y el estadístico t no está definido.

Dado que el *valor-p*, asociado al valor muestral del estadístico de contraste, es más pequeño que $0,01$ se rechaza la hipótesis nula al $1\,\%$ de significación.

b) Contraste de que la inversión real responde al tipo de interés real, $H_0 : \beta_4 + \beta_5 = 0$, frente a $H_a : \beta_4 + \beta_5 \neq 0$. De la misma forma que antes, en la ventana **gretl: modelo1** seleccionamos *Contrastes* → *Restricciones lineales*. En la nueva ventana que aparece escribimos b4+b5= 0. Al aceptar se obtiene el siguiente resultado

```
Restricción:
 b[INTERES] + b[INFLACION] = 0

Estadístico de contraste:
  F(1, 10) = 3,25354, con valor p = 0,10143

Estimaciones restringidas:
  VARIABLE     COEFICIENTE       DESV.TÍP.      ESTAD T      VALOR P
  const         -0,505855        0,0604801      -8,364    <0,00001 ***
  t             -0,0170255       0,00214732     -7,929    <0,00001 ***
  PNBR           0,657533        0,0598599      10,985    <0,00001 ***
  INTERES       -0,00133784      0,00119517     -1,119     0,28683
  INFLACION      0,00133784      0,00119517      1,119     0,28683
```

De nuevo se muestran las estimaciones del modelo restringido. En este caso se estiman todos los coeficientes bajo la restricción de que $\beta_4 = -\beta_5$. El coeficiente estimado que acompaña a INTERES es el mismo valor pero con signo contrario que el obtenido para el coeficiente de INFLACION. Este resultado surge de la restricción impuesta ($\beta_4 = -\beta_5$). De igual forma coinciden las varianzas estimadas y las desviaciones típicas.

Dado que el *valor-p*, asociado al valor muestral del estadístico de contraste, es mayor que $0,1$ no se rechaza la hipótesis nula al $10\,\%$ (ni al $5\,\%$ o $1\,\%$) de significación.

c) Por último, realizamos el contraste conjunto de estas dos restricciones lineales, la propensión marginal a invertir es la unidad y la inversión real responde al tipo de interés real. Esto es $H_0 : \beta_3 = 1$, $\beta_4 + \beta_5 = 0$ frente a la alternativa de que al menos una de ellas no se satisface, $H_a : \beta_3 \neq 1, \text{y} \setminus \text{o}\, \beta_4 + \beta_5 \neq 0$.

De nuevo, en la ventana **gretl: modelo1** seleccionamos

Contrastes → *Restricciones lineales*

y escribimos

```
b3=1
b4+b5=0
```

Al aceptar se obtiene el siguiente resultado:

```
Conjunto de restricciones
 1: b[PNBR] = 1
 2: b[INTERES] + b[INFLACION] = 0

Estadístico de contraste:
   F(2, 10) = 21,3453, con valor p = 0,000246226

Estimaciones restringidas:
 VARIABLE       COEFICIENTE      DESV.TÍP.      ESTAD T     VALOR P
  const          -0,851039       0,00799803    -106,406   <0,00001 ***
  t           -0,0289471      0,000989688   -29,249    <0,00001 ***
  PNBR            1,00000        0,000000        indefinido
  INTERES        -0,00172664     0,00227790     -0,758    0,46308
  INFLACION       0,00172664     0,00227790      0,758    0,46308

  Desviación típica de los residuos = 0,0140693
```

Se rechaza la hipótesis nula al $1\,\%$ de significación, ya que el *valor-p* es menor que 0,01. Por lo tanto, al menos una de las restricciones parece no satisfacerse. Viendo los resultados de los contrastes individuales, parece que la evidencia es contra la primera restricción.

4.3. Estimación bajo restricciones lineales

El estimador resultante de minimizar la suma de los residuos al cuadrado sujeto a restricciones lineales del tipo $R\beta = r$, esto es

$$\min_{\hat{\beta}_R} \sum_{i=1}^{N} (Y_i - \hat{\beta}_{R,1} - \hat{\beta}_{R,2} X_{2i} - \hat{\beta}_{R,3} X_{3i} - \cdots - \hat{\beta}_{R,K} X_{Ki})^2$$

$$\texttt{sujeto a} \qquad R\hat{\beta}_R = r$$

se puede expresar como:

$$\hat{\beta}_R = \hat{\beta} - (X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - r) \tag{4.8}$$

donde $\hat{\beta} = (X'X)^{-1} X'Y$ es el estimador de los parámetros β sin imponer las restricciones. Dado que el estimador no restringido $\hat{\beta}$ se ha obtenido sin imponer que éste satisfaga tales restricciones, en general $(R\hat{\beta} - r) \neq 0$. La solución restringida, $\hat{\beta}_R$, es igual a la solución no restringida, $\hat{\beta}$, menos un término de ajuste que tiene en cuenta en qué medida la solución no restringida no satisface las restricciones. Si hemos obtenido ya $\hat{\beta}$ podemos utilizar directemente la expresión (4.8) para obtener el estimador de β restringido, es decir $\hat{\beta}_R$.

Hemos visto en la sección anterior que el programa Gretl muestra las estimaciones del modelo restringido cuando se selecciona la opción de contrastar restricciones lineales, a la vez que el valor muestral del estadístico de contraste.

Otra posibilidad es la de estimar el modelo imponiendo la o las restricciones. Cuando las restricciones implican solamente la exclusión de variables explicativas del modelo de partida, no hay mayor problema en llevar a cabo la estimación del modelo restringido. Bien se realiza la regresión eliminando del listado de regresores esas variables o, como hemos visto antes en Gretl, se puede utilizar la opción *Contrastes* ⇒ *Omitir variables* a la vez que se contrasta.

Si las restricciones no son simplemente de exclusión, entonces se pueden sustituir en el modelo de partida y reorganizarlo en función del conjunto de $(K - q)$ parámetros que quedan sin determinar. Una ventaja de proceder así es que se dispone de las mismas opciones que en la ventana de estimación de un modelo por mínimos cuadrados ordinarios. Por ejemplo, se pueden hacer otro tipo de contrastes en el modelo restringido, guardar sus residuos, etc.

Por ejemplo, si queremos obtener **el estimador de los parámetros bajo la restricción de que la propensión marginal a invertir sea la unidad**, podemos hacerlo sustituyendo en el modelo

$$INVERR_t = \beta_1 + \beta_2\, t + \beta_3\, PNBR_t + \beta_4\, INTERES_t + \beta_5\, INFLACION_t + u_t \tag{4.9}$$

la restricción $\beta_3 = 1$ y reorganizar tal que nos quedaría la siguiente regresión:

$$INVERR_t - PNBR_t = \beta_1 + \beta_2\, t + \beta_4 INTERES_t + \beta_5 INFLACION_t + u_t \tag{4.10}$$

en función de $K - q = 5 - 1 = 4$ parámetros a estimar. El quinto ya está determinado por la restricción. Definimos una nueva variable llamémosla R, calculada como $R_t = INVERR_t - PNBR_t$, utilizando la opción en Gretl de

Variable → Definir nueva variable

y en la ventana que aparece escribimos R = INVERR-PNBR. De esta forma se añade la variable R al conjunto de variables disponibles que aparecen en la ventana principal o de inicio. Seguidamente, se realiza la regresión de esta variable sobre la constante, t, INTERES e INFLACION con *Modelo → Mínimos cuadrados ordinarios* y se obtienen los siguientes resultados:

Modelo Restringido (4.10): estimaciones MCO utilizando las 15 observaciones 1968–1982
Variable dependiente: R

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−0,837112	0,0134320	−62,3223	0,0000
t	−0,0276707	0,00139136	−19,8875	0,0000
INTERES	−0,00311914	0,00247563	−1,2599	0,2338
INFLACION	−0,000342359	0,00275183	−0,1244	0,9032

Media de la var. dependiente	−1,0840
D.T. de la variable dependiente	0,131901
Suma de cuadrados de los residuos	0,00207013
Desviación típica de los residuos ($\hat{\sigma}$)	0,0137184
R^2	0,991501
$\bar{R}^2$ corregido	0,989183
$F(3, 11)$	427,751
Estadístico de Durbin–Watson	0,995558
Coef. de autocorr. de primer orden.	0,441936
Log-verosimilitud	45,3774
Criterio de información de Akaike	−82,754
Criterio de información Bayesiano de Schwarz	−79,922
Criterio de Hannan–Quinn	−82,784

Recordamos lo que se obtenía al realizar el contraste de esa restricción en la ventana de estimación por MCO del modelo no restringido mediante *Contrastes → Restricciones Lineales*:

```
Restricción: b[PNBR] = 1

Estadístico de contraste: F(1, 10) = 35,92, con valor p = 0,000133289

Estimaciones restringidas:
   VARIABLE       COEFICIENTE       DESV.TÍP.    ESTAD T    VALOR P
    const          -0,837112         0,0134320   -62,322    <0,00001 ***
    t          -0,0276707        0,00139136  -19,888    <0,00001 ***
    PNBR            1,00000          0,000000        indefinido
    INTERES        -0,00311914       0,00247563   -1,260     0,23377
    INFLACION      -0,000342359      0,00275183   -0,124     0,90323
  Desviación típica de los residuos = 0,0137184
```

Los coeficientes estimados corresponden a las realizaciones del estimador de Mínimos Cuadra-

dos Restringidos para los cuatro coeficientes que quedaban sin determinar por la restricción[4]. El valor para el coeficiente de PNBR viene dado por la restricción y es igual a la unidad. Su varianza por lo tanto es igual a cero ya que su valor está dado.

Hay que notar que el R^2, y por lo tanto el corregido, obtenidos en este ajuste no son comparables con los resultantes de estimar el modelo no restringido, ya que en este caso la Suma de Cuadrados Total corresponde a la variable $R = INVERR - PNBR$ que es el regresando de esta regresión y no a $INVERR$ que es realmente la variable endógena de interés a explicar. Para que los R^2 sean comparables entre el modelo no restringido y el restringido la Suma de Cuadrados Total tiene que ser la misma. Veremos en la sección siguiente los que sí son comparables y un estadístico de contraste basado en ellos.

4.4. Estadísticos equivalentes

Partimos del modelo $Y = X\beta + u$ donde se quiere contrastar las restricciones lineales $H_0 : R\beta = r$. Podemos obtener la suma de los residuos al cuadrado y el coeficiente de determinación correspondientes a la estimación del modelo sin restringir y al modelo restringido, de la siguiente forma:

$$SCR_{NR} = \hat{u}'\hat{u} = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \qquad R^2_{NR} = 1 - \frac{\hat{u}'\hat{u}}{\sum(Y_t - \overline{Y})^2}$$

$$SCR_R = \hat{u}'_R\hat{u}_R = (Y - X\hat{\beta}_R)'(Y - X\hat{\beta}_R) \qquad R^2_R = 1 - \frac{\hat{u}'_R\hat{u}_R}{\sum(Y_t - \overline{Y})^2}$$

Por otra parte, utilizando las sumas de cuadrados de los residuos correspondientes a la estimación del modelo restringido y no restringido, SCR_R y SCR_{NR} respectivamente y sus grados de libertad, gl_R y gl_{NR}, es posible realizar el contraste de las restricciones lineales con el siguiente estadístico:

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-K)} \overset{H_0}{\sim} \mathcal{F}(q, T-K) \tag{4.11}$$

Nótese que los grados de libertad de la distribución del estadístico bajo la hipótesis nula son en el numerador $gl_R - gl_{NR} = (T - (K - q)) - (T - K) = q$, el número de restricciones, y en el denominador $gl_{NR} = T - K$. Se puede demostrar que este estadístico es el mismo que el estadístico anterior (4.4). La diferencia radica en que calcularlo de esta forma requiere estimar tanto el modelo sin restringir como el restringido.

Su interpretación puede ser más intuitiva. Imponer restricciones en la estimación siempre empeora el ajuste tal que la diferencia de las sumas de cuadrados residuales del modelo restringido y no restringido, $(SCR_R - SCR_{NR})$, es mayor o igual a cero. Ahora bien, cuanto más grande sea esta diferencia más evidencia habrá de que las restricciones no sean ciertas, es decir contra la hipótesis nula. Se rechazará esta hipótesis nula si el valor muestral del estadístico es suficientemente grande como para caer en una región crítica establecida.

[4]El estimador restringido será $\hat{\beta}_R = \begin{bmatrix} \hat{\beta}_{R,1} & \hat{\beta}_{R,2} & 1 & \hat{\beta}_{R,4} & \hat{\beta}_{R,5} \end{bmatrix}'$ donde $\hat{\beta}_{R,1}$, $\hat{\beta}_{R,2}$, $\hat{\beta}_{R,4}$ y $\hat{\beta}_{R,5}$, son los obtenidos de la regresión bajo la restricción de que el coeficiente que acompaña al PNBR en el modelo para la Inversión real es igual a 1.

Si dividimos numerador y denominador por la suma de cuadrados total $SCT = \sum_t (Y_t - \overline{Y})^2$ podemos expresar el estadístico en términos de los coeficientes de determinación[5] :

$$F = \frac{(R^2_{NR} - R^2_R)/q}{(1 - R^2_{NR})/(T-K)} \overset{H_0}{\sim} \mathcal{F}_{(q,T-K)} \tag{4.12}$$

El contraste se realizará del mismo modo que con los otros estadísticos equivalentes.

Vamos a ilustrar esta forma de realizar el contraste en el ejemplo del modelo para la inversión agregada. Para realizar el contraste de la restricción de que la propensión marginal a invertir es igual a la unidad, utilizamos las sumas de cuadrados residuales de la estimación del modelo restringido (4.10) y el modelo no restringido (4.9). Esto ya lo obtuvimos en la secciones anteriores. En la ventana donde hemos realizado la regresión en cada caso podemos guardar las sumas de cuadrados residuales y añadirlo a las variables ya definidas con *Guardar* → *Suma de cuadrados de lo residuos*. En concreto se obtienen las siguientes sumas de cuadrados residuales:

$$SCR_R = 0,00207013 \qquad SCR_{NR} = 0,000450812$$

Sustituyendo en el estadístico (4.11) obtenemos el siguiente valor muestral[6]:

$$F = \frac{(0,00207013 - 0,000450812)/(15-4) - (15-5)}{0,000450812/(15-5)} = 35,92$$

siendo este el mismo valor que obtuvimos anteriormente con el estadístico utilizando *Contrastes* → *Restricciones lineales*, y por lo tanto obtenemos la misma conclusión del contraste, se rechaza la hipótesis nula de que la propensión marginal a invertir sea la unidad.

A su vez, utilizando el dato que nos da Gretl de la Desviación típica para la variable dependiente $INVERR$, podemos obtener la Suma de Cuadrados Total como,

$$SCT = \sum (INVERR_t - \overline{INVERR})^2 = (15-1)(D.T.\,INVERR)^2 = 14(0,0341774)^2$$

obteniendo el valor $SCT = 0,016353325$. Por lo tanto la realización de R^2_R es en este caso,

$$R^2_R = 1 - \frac{\hat{u}'_R \hat{u}_R}{SCT} = 1 - (0,00207013/0,016353325) = 0,87341$$

que no coincide con el que muestra la regresión del modelo (4.10). Esta vez este valor sí es comparable con el valor obtenido para el coeficiente de determinación de estimar el modelo no restringido, $R^2_{NR} = 0,972433$. Se puede apreciar, como era de esperar, que el valor obtenido del R^2_R es menor que el del R^2_{NR}, el ajuste empeora al imponer la restricción. La cuestión es si esto es aceptable, con un nivel de confianza elegido, para aceptar la hipótesis nula como cierta o no.

[5] Este es el estadístico que se introdujo en el Tema 3. En ese tema se vió como caso particular el estadístico de significación conjunta

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

En ese caso $R^2_R = 0$

[6] Se puede hacer el cálculo con Gretl utilizando *Datos* → *Definir nueva variable* y escribiendo la fórmula del estadístico en términos de los nombres asignados a las variables sumas de cuadrados residuales.

El valor del estadístico (4.12) para este caso es,

$$F = \frac{(R^2_{NR} - R^2_R)/q}{(1 - R^2_{NR})/(T - K)} = F = \frac{(0,972433 - 0,87341)/1}{(1 - 0,972433)/(15 - 5)} = 35,92$$

obteniendo de nuevo el mismo valor para el estadístico y la misma conclusión del contraste.

4.5. Predicción

Uno de los objetivos de la econometría consiste en predecir. Una vez estimado un modelo que se considera que recoge bien el comportamiento de una variable en función de otros factores o variables explicativas, se quiere determinar con cierta confianza el valor o intervalo de valores que puede tomar la variable dependiente, supuestos unos valores para esos factores.

Supongamos que se ha estimado el siguiente modelo[7]:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_K X_{Kt} + u_t$$

con una muestra de tamaño T, obteniendo la siguiente función de regresión muestral (FRM):

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \ldots + \hat{\beta}_K X_{Kt}$$

Entonces, disponiendo de nuevas observaciones de las variables explicativas,

$$X'_p = \left[\begin{array}{cccc} 1 & X_{2p} & \ldots & X_{Kp} \end{array} \right] \qquad p \notin \{1, 2, \ldots, T\}$$

podemos utilizar el modelo estimado por MCO para predecir el valor que tomará la variable endógena en el periodo de predicción p. A este proceso se le llama predicción por punto, donde el valor estimado para la variable endógena Y en el periodo de predicción se obtiene sustituyendo estos valores de las variables exógenas en la FRM.

$$\hat{Y}_p = X'_p \, \hat{\beta}_{MCO}$$

Equivalentemente:

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

El error de predicción se define como $e_p = Y_p - \hat{Y}_p = -X'_p(\hat{\beta} - \beta) + u_p$. Para obtener la predicción por intervalo, nos basaremos en la distribución del error de predicción, ya que si u_p y $\hat{\beta}$ son variables aleatorias normales, el error de predicción también lo será:

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

Sin embargo, en general, σ^2 es desconocido por lo que utilizaremos su estimador insesgado propuesto en temas anteriores obteniendo el siguiente resultado:

$$\frac{e_p}{\hat{\sigma}\sqrt{1 + X'_p \, (\, X'X \,)^{-1} \, X_p}} \sim t_{(T-K)}$$

[7] En lo que sigue, como siempre, se satisfacen las hipótesis básicas tanto en el periodo de estimación como de predicción

A partir de este estadístico podemos obtener un intervalo con un nivel de confianza del $1-\alpha$ alrededor de la predicción por punto para la variable endógena en el momento p.

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

donde $\hat{\sigma}^2_{e_p} = \hat{\sigma}^2(\,1 + X_p'\,(\,X'X\,)^{-1}\,X_p\,)$.

¿Cómo utilizar Gretl para predecir por punto y por intervalo?

Utilizaremos el ejemplo de los precios de las viviendas para analizar los pasos a seguir en el programa Gretl.

Uno de los modelos propuestos era

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$$

Supongamos que tenemos información de una nueva vivienda, por ejemplo, $F2 = 3200$, $BEDRMS = 5$ y $BATHS = 3$ y nos piden $P = 500$, en miles de euros, por ella. Mediante este modelo, podemos obtener una predicción del precio que tendría una vivienda con estas características y analizar si el precio solicitado es razonable o no.

Para ello, incorporamos los nuevos datos (X_p) a la base de datos mediante

Datos → Seleccionar todos

A continuación, pincharemos la opción

Datos → Añadir Observaciones

indicando el número de observaciones que queremos añadir, en este caso 1. En la fila correspondiente incluimos los valores de las variables explicativas en el periodo de predicción, en este caso la observación 15, incorporando cada observación en la casilla correspondiente. Si no incorporamos el valor para la variable P 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 (recordar que inicialmente teníamos 14 observaciones en la muestra). Para ello, tenemos que especificar el rango muestral, es decir, en la opción

Muestra → Establecer rango

especificaremos del rango de observaciones de la muestra para estimar el modelo, en nuestro caso de la 1 a la 14 y elegimos *Aceptar*.

Tal y como explicamos en los temas anteriores, estimaremos el modelo por MCO y en la ventana de los resultados elegimos

Análisis → Predicciones

En la nueva ventana podemos determinar el dominio de predicción, es decir el *Inicio* y *Fin* que en este caso es en ambos la observación número 15, y también cuantas observaciones se quieren representar antes de la prediccion[8].

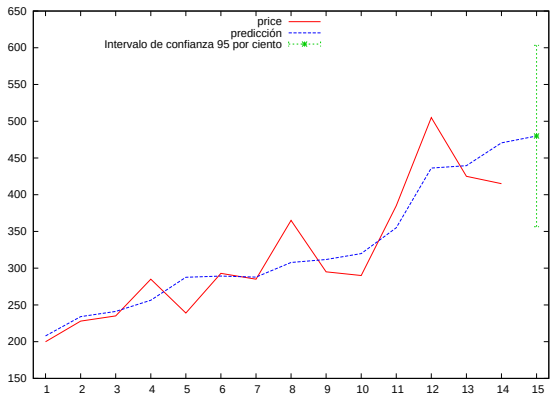
[8] En este caso hemos elegido todas pero esto es opcional.

Los resultados que muestra Gretl son los siguientes:

```
Para intervalos de confianza 95%, t(10, .025) = 2,228

     Obs        price    predicción  desv. típica   Intervalo de confianza 95%

       1        199,9        207,8
       2        228,0        234,0
       3        235,0        241,2
       4        285,0        256,3
       5        239,0        287,6
       6        293,0        289,2
       7        285,0        287,8
       8        365,0        307,8
       9        295,0        311,8
      10        290,0        319,9
      11        385,0        355,1
      12        505,0        436,3
      13        425,0        439,6
      14        415,0        470,5
      15                     479,9        55,39        356,5 -    603,3
```



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 $95\,\%$.

Bibliografía

Greene, W. (2008), *Econometric Analysis*, 6ª edn., Prentice-Hall.

Tema 5

Errores de especificación en la elección de los regresores

Contenido

5.1. Introducción

La primera especificación de un modelo de regresión implica tomar varias decisiones, a menudo previas a la confrontación de éste con los datos. Algunas de estas decisiones son:

- Elección de la variable dependiente.
- Elección de las variables explicativas.
- Medición de las variables.
- Forma funcional de la relación. Estabilidad.
- Especificación de las propiedades del término de error.

En los temas anteriores hemos especificado un modelo de regresión donde se satisfacen una serie de hipótesis básicas. Algunas de estas hipótesis pueden no mantenerse si las decisiones adoptadas son erróneas o porque simplemente, dadas las características de las variables del modelo y de los datos a utilizar, estas hipótesis pudieran no ser adecuadas. Esto puede influir negativamente en las propiedades del estimador utilizado y en la inferencia, siendo las decisiones posteriores sobre el modelo erróneas. En muchos casos la evaluación de un modelo puede estar influenciada por esta primera especificación. Por ello, es importante disponer de instrumentos o contrastes que nos permitan hacer un diagnóstico sobre si son aceptables ciertas decisiones o hipótesis adoptadas. Estos instrumentos pueden ser un análisis gráfico de los residuos o contrastes estadísticos donde se traten de detectar problemas de mala especificación.

En este tema nos vamos a centrar en ilustrar las implicaciones que pueden tener decisiones erróneas en términos de la elección de las variables explicativas o regresores. Para ello vamos a proponer que conocemos el modelo correcto y consideramos separadamente dos situaciones:

a) Omisión de variables explicativas relevantes. Analizaremos las implicaciones en el estimador MCO y en la validez de los contrastes de significatividad. Veremos la utilización del gráfico de residuos y algún contraste de mala especificación con algunos ejemplos empíricos.

b) Inclusión de variables irrelevantes. En este caso nos interesaremos por los efectos de haber incluido variables que sabemos no tendrían que estar en el modelo. La cuestión es cómo detectar y decidir en la práctica qué variables son o no relevantes. También discutiremos estas cuestiones utilizando un caso práctico.

Aunque teóricamente analizaremos cada uno de estos efectos por separado y asumiremos que conocemos la especificación correcta, en la práctica podemos tener combinados estos efectos.

5.2. Efectos de omisión de variables relevantes

Podemos seguir con nuestro ejemplo sobre el precio de la vivienda en el que queríamos explicar esta variable, medida en miles de dólares, en función de una serie de variables explicativas

como podían ser el tamaño de la vivienda $F2$, el número de habitaciones $BEDRMS$ y el número de baños $BATHS$. En principio, vamos a considerar que el modelo correcto para explicar el precio de la vivienda es

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, \ldots, N \tag{5.1}$$

donde se satisfacen las hipótesis básicas pero se estima por MCO el siguiente,

$$P_i = \beta_1 + \beta_3\, BEDRMS_i + \beta_4\, BATHS_i + v_i \qquad i = 1, \ldots, N \tag{5.2}$$

En el modelo considerado a la hora de estimar se ha omitido la variable F2 o tamaño de la vivienda. Si esta variable es relevante entonces $\beta_2 \neq 0$ por lo que el error v_i recogerá la variable omitida, esto es $v_i = \beta_2 F2_i + u_i$, siendo $E(v_i) = \beta_2 F2_i \neq 0$. Luego en el modelo mal especificado no se satisface una de las hipótesis básicas. Esto a su vez implica que la covarianza entre las variables incluidas y el error del modelo (5.2) dependerá de la covarianza entre la variable omitida $F2_i$ y cada una de las incluidas $BEDRMS_i$ y $BATHS_i$. Si estas no son cero, esto introducirá un sesgo en los coeficientes estimados que será función de estas covarianzas. El signo del sesgo dependerá del signo del coeficiente β_2 y de los signos de estas covarianzas. Se puede demostrar que los sesgos de estimar por MCO β_3 y β_4 en el modelo (5.2) son

$$E(\hat{\beta}_3) - \beta_3 = \beta_2 \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \qquad E(\hat{\beta}_4) - \beta_4 = \beta_2 \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \tag{5.3}$$

donde $S_{js} = \sum_i (X_{ji} - \bar{X}_j)(X_{is} - \bar{X}_s)$, siendo la covarianza muestral entre dos variables j, s si $j \neq s$, y la varianza muestral de la variable j si $j = s$. Como se puede apreciar, el sesgo en la estimación de ambos coeficientes depende de las covarianzas entre las variables relevante excluida $F2$ y cada una de las variables incluidas $BEDRMS$ y $BATHS$[1]. Además depende del coeficiente β_2 que en el modelo correcto (5.1) se esperaba fuera positivo, pero la dirección del signo de cada sesgo no es clara ya que depende del signo del cociente que acompaña a β_2. Para que no hubiera sesgo en la estimación de cualquiera de estos dos coeficientes **ambas variables incluidas**, $BEDRMS$ y $BATHS$ tendrían que estar **incorreladas con** el tamaño de la vivienda o **variable excluida**, cosa poco probable en este ejemplo.

[1] Si el modelo de partida correcto hubiera sido

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + u_i \quad i = 1, \ldots, N \tag{5.4}$$

pero hubiéramos considerado para estimar

$$P_i = \beta_1 + \beta_3 BEDRMS_i + v_i \quad i = 1, \ldots, N \tag{5.5}$$

entonces el sesgo en estimar β_3 en (5.5) sería simplemente

$$E(\hat{\beta}_3) - \beta_3 = \beta_2 \frac{S_{23}}{S_{33}} \tag{5.6}$$

El sesgo sigue dependiendo de la covarianza entre la variable omitida $F2$ y la incluida $BEDRMS$ dada por S_{23}. En este caso se puede esperar que el sesgo fuera positivo ya que tanto S_{23} como β_2 se esperan sean positivos. El efecto de omitir $F2$ o no controlar por el tamaño de la vivienda en el modelo (5.5) será sobreestimar el efecto marginal de tener una habitación más en la vivienda sobre el precio de ésta. Por tanto, el número de habitaciones estaría también de alguna forma representando el papel del tamaño de la vivienda, que no se ha incluido en el modelo. No se estimaría con sesgo si $S_{23} = 0$, cosa que no parece factible ya que el número de habitaciones estará correlacionado con el tamaño de la vivienda.

En cuanto al sesgo en la estimación del coeficiente que acompaña al término constante se puede demostrar que es[2]

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \left(\bar{X}_2 - \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_3 - \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_4 \right) \tag{5.7}$$

Vemos que en este caso aún siendo $S_{23} = S_{24} = 0$ el sesgo no se anularía, ya que todavía depende de la media de la variable omitida $\bar{X}_2$, que generalmente no va a ser cero. De este resultado se puede argumentar que el coeficiente que acompaña al término constante, generalmente va a recoger efectos de variables omitidas aún cuando esto no influya en la estimación del resto de parámetros o pendientes por estar estas variables incorreladas con las incluidas. Por ello, normalmente es conveniente no excluir el término constante, a no ser que se tengan fuertes razones teóricas para hacerlo.

Si se estiman con sesgo los coeficientes β_j, también serán incorrectos los contrastes de significatividad individual, conjunta y otro tipo de contrastes sobre los coeficientes del modelo utilizando estas estimaciones sesgadas. Ahora bien, ¿serán fiables los contrastes sobre las pendientes si se dan las condiciones para que los estimadores de estos parámetros no sean sesgados? La respuesta es que no, ya que aún dandose las condiciones de incorrelación entre regresores incluidos y variables relevantes excluidas, el estimador de la matriz de varianzas y covarianzas de esos coeficientes estimados seguirá siendo sesgada. Esto se debe a que el estimador del parámetro σ^2 utilizando la suma de cuadrados residual de la estimación del modelo mal especificado estará sesgado en cualquiera de los casos.

Luego vemos que en general las consecuencias de omitir variables relevantes en la especificación de un modelo son serias, especialmente en la inferencia.

¿Cómo detectar que esto pueda estar ocurriendo? Una primera cuestión es tener en cuenta el modelo teórico de interés y pensar qué variables pueden faltar en el modelo empírico. Por otro lado, podemos ayudarnos de contrastes que puedan señalar la existencia de algún problema de mala-especificación[3].

Además, el análisis de los residuos nos puede ayudar a ver si hemos dejado fuera factores relevantes. Por ejemplo, podemos ver el gráfico de los residuos por observación y ver si estos presentan algún comportamiento sistemático que pueda apuntar en esa dirección.

Por ejemplo, consideremos los resultados de la estimación de los modelos (5.1) y (5.2) para explicar el precio de la vivienda[4]

[2] Ocurre lo mismo si consideramos que el modelo estimado es (5.5) y el verdadero modelo es (5.4).

[3] En este tema ilustraremos alguno de estos contrastes, aunque no todos. Incluso algunos contrastes diseñados para analizar si el término de error no está autocorrelacionado, puede capturar también otro tipo de cuestiones de mala especificación.

[4] Los valores entre paréntesis son los correspondientes estadísticos t de significatividad individual.

Variable	**Modelo (5.1)**	**Modelo (5.2)**
	Supuestamente Correcto	
CONSTANT	129,062	27,2633
	(1,462)	(0,182)
F2	0,1548	
	(4,847)	
BEDRMS	-21,588	-10,1374
	(-0,799)	(-0,216)
BATHS	-12,193	138,795
	(-0,282)	(2,652)
Suma de cuadrados de los residuos	16700,1	55926,4
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657	71,3037
R^2	0,836	0,450706
$\bar{R}^2$	0,787	0,350834
F de significación conjunta	16,989	4,51285
Grados de libertad	10	11
Criterio de Akaike (AIC)	146,908	161,829
Criterio de Schwarz (BIC)	149,464	163,746

Tabla 5.1: Modelos (5.1) y (5.2) estimados para el precio de la vivienda

Como ya comentamos en el capítulo anterior, la omisión de la variable $F2$ empeora bastante el ajuste tanto en términos del R^2 como del $\bar{R}^2$, AIC y BIC. El coeficiente estimado que más ha cambiado es el que acompaña a la variable $BATHS$ pasando a tener signo positivo y ser significativamente distinto de cero. Parece que, dado que ambas variables representan también tamaño de la vivienda, el efecto indirecto de la omisión de esta variable puede estar siendo capturando más por el coeficiente de $BATHS$ que por el de $BEDRMS$.

Podemos mirar a las correlaciones entre la variable excluida $F2$ y las incluidas $BEDRMS$ y $BATHS$. En la ventana principal de Gretl donde tenemos estas variables, las seleccionamos con el botón izquierdo del ratón, mientras mantenemos la tecla de mayúsculas ⇑, y en *Ver* → *matriz de correlación* obtenemos

Coeficientes de correlación, usando las observaciones 1 - 14
valor crítico al 5 % (a dos colas) = 0,5324 para n = 14

F2	BEDRMS	BATHS	
1,0000	0,4647	0,7873	F2
	1,0000	0,5323	BEDRMS
		1,0000	BATHS

Vemos que, aunque tanto el número de habitaciones $BEDRMS$ como el número de baños $BATHS$ presenta una correlación positiva con la variable excluida, tamaño de la vivienda $F2$, es la variable $BATHS$ la que presenta una mayor correlación con esta última.

Seguidamente vamos a analizar diversos gráficos de los residuos del ajuste del modelo (5.2) donde hemos omitido $F2$ que parece ser relevante. De la estimación de este modelo en la ventana de estimación **gretl:modelo2** elegimos

Gráficos → *Grafico de residuos* → *Por número de observación*

que nos muestra el gráfico de residuos por observación según están las 14 observaciones ordenadas en la muestra. Lo podemos guardar posicionando el cursor dentro de la ventanta del gráfico y pinchando con el botón derecho del ratón, aparece un menú con distintas opciones y formatos para guardarlo.

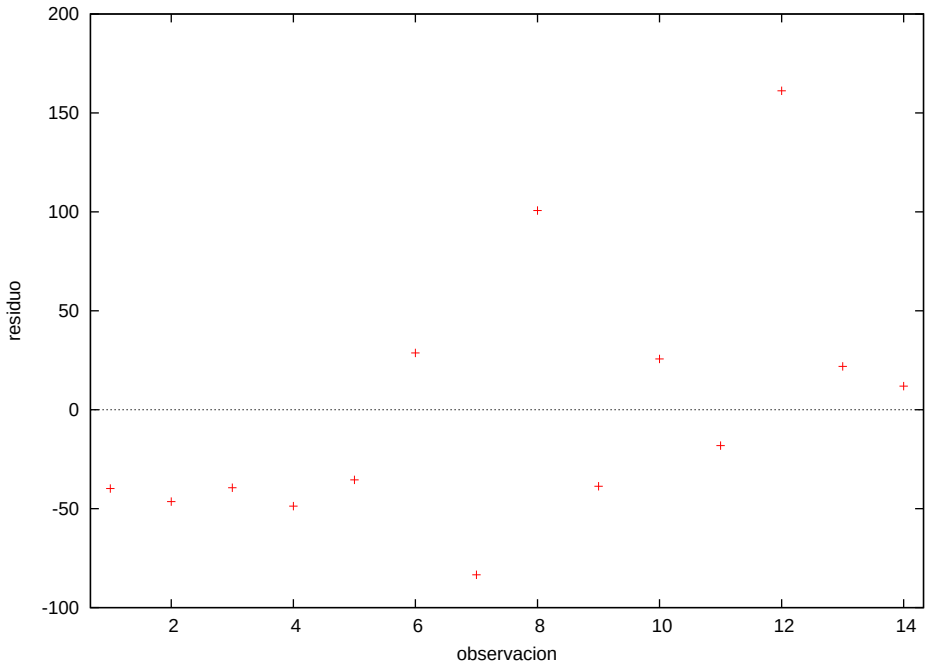


Gráfico 5.1: Gráfico de los residuos del Modelo (5.2) por observación

En el gráfico se puede apreciar que hay demasiados residuos negativos juntos al comienzo de la muestra y a medida que vamos hacia las últimas observaciones o viviendas, estos se concentran más en la parte positiva. Si observamos la disposición de las viviendas en la muestra, veremos que están ordenadas en función creciente del tamaño de la vivienda. Luego los residuos negativos estarían asociados en general con viviendas de menor tamaño y los positivos con viviendas de mayor tamaño. Esto sugiere un comportamiento sistemático en la disposición de los residuos alrededor de su media muestral que es cero.

El gráfico de los residuos sobre la variable $F2$ puede ayudar a ver si hay alguna relación. De hecho el gráfico nos mostrará la recta de regresión de los residuos sobre esta variable si es que existe una relación significativa. Para obtener el gráfico primero tenemos que guardar los residuos de la estimación del modelo (5.2). Para ello, en la ventana de estimación **gretl:modelo2** elegimos

Guardar → *Residuos*

y le damos un nombre a la serie de residuos. Esta serie aparecerá en la ventana principal **gretl** y la podremos utilizar posteriormente. En esta misma ventana elegimos

Ver → Gráficos → Grafico X-Y (scatter)

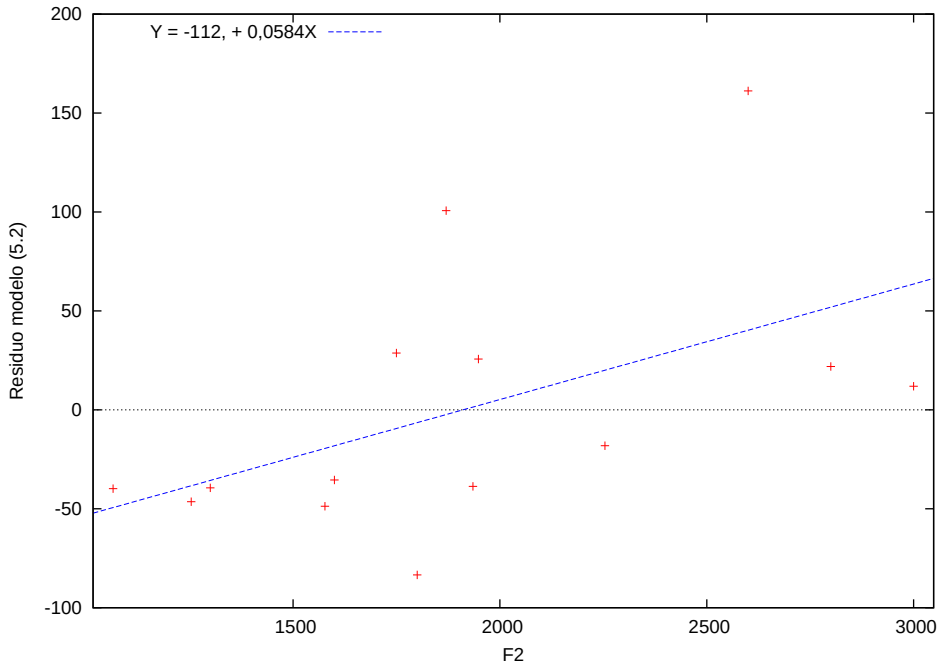


Gráfico 5.2: Gráfico de los residuos del Modelo (5.2) sobre F2

En la ventana que aparecerá posteriormente, especificamos que variable se representa en el eje de ordenadas *eje X*, en este caso $F2$, y en el eje de abcisas o *eje Y*, en este caso los residuos de la estimación del Modelo (5.2). En este gráfico podemos apreciar que hay una relación positiva significativa entre los residuos de la estimación del modelo (5.2) y la variable $F2$ omitida en ese modelo. De hecho, la línea que aparece en el gráfico representa la recta de regresión de los residuos sobre esa variable. Esto indica que cierto componente residual puede ser explicado por la variable que no hemos incluido.

Lo detectado en estos gráficos puede ser contrastado utilizando el siguiente contraste que se debe a Engle (1982). Este contraste utiliza el R^2 de la regresión auxiliar de los residuos del modelo que se está analizando sobre la variable o variables que sospechamos puedan ser candidatas a ser incluidas en él por ser relevantes. En nuestro caso sería realizar la regresión

$$\hat{u}_i = \delta_1 + \delta_2 F2_i + \xi_i \qquad i = 1, \ldots, N \tag{5.8}$$

El estadístico de contraste es NR^2 donde el R^2 es el coeficiente de determinación de esta regresión auxiliar. La distribución exacta del estadístico, bajo la hipótesis nula de que la variable $F2$ no es una variable relevante a incluir en el modelo, no es conocida pero se puede aproximar por la distribución χ^2 con un grado de libertad[5]. Esta aproximación será mejor cuanto mayor sea el tamaño muestral.

[5]En general, los grados de libertad serán el número de regresores de la regresión auxiliar sin contar el término constante.

En el ejemplo que nos ocupa esta regresión auxiliar la podemos obtener con Gretl eligiendo

Modelo → Minimos Cuadrados Ordinarios

y en la ventana que emerge elegir como variable dependiente la serie de residuos de la estimación del modelo (5.2) que teníamos guardada y como regresores a $F2$ además de la constante. Los resultados de esta regresión auxiliar (5.8) para el ejemplo que nos ocupa son

$$\widehat{\mathrm{u}}_i = \underset{(-1,995)}{-111,588} + \underset{(2,078)}{0,0583946}\,\mathrm{F2}_i$$
$$N = 14 \quad R^2 = 0,264584$$

Si queremos guardar el valor muestral NR^2 podemos hacerlo en esa misma ventana eligiendo

*Guardar → T*R-cuadrado*

El valor muestral del estadístico $NR^2 = 3,70417$ se muestra en la ventana principal con el resto de variables. Este valor habrá que compararlo en este caso con el valor crítico $\chi^2_{(1)\alpha}$ utilizando en el contraste un nivel de significación α concreto.

Para buscar el valor crítico en las tablas de la Chi-cuadrado con 1 grado de libertad podemos elegir en la ventana principal de Gretl, *Herramientas → Tablas Estadísticas* y en la ventana que aparece seleccionar la chi-cuadrado especificando 1 grado de libertad. Aparece una ventana con los valores críticos de la distribución Chi-cuadrado para distintos niveles de significación.

También podemos obtener el *valor-p* dado el valor muestral del estadístico. En la ventana principal de nuevo en *Herramientas → Buscador de valores-p*, y en la ventana que aparece seleccionar la chi-cuadrado especificando en la primera casilla 1 grado de libertad y el valor muestral del estadístico en la segunda casilla. Aparece una ventana con la siguiente información: Chi-cuadrado(1): área a la derecha de 3,70417 = 0,0542767 (a la izquierda: 0,945723).

Por lo tanto, como el *valor-p* obtenido es $0,0542767$ que, aunque poco, es algo mayor que $0,05$, no se rechazaría la hipótesis nula de que $F2$ sea una variable importante a añadir al modelo al 5 %, pero sí al 10 % al ser el *valor-p* en ese caso menor que ese nivel de significación. Vemos que la hipótesis nula se rechazaría al 10 % de significación ya que el valor muestral en ese caso $NR^2 = 3,70417$ sería mayor que el valor crítico $\chi^2_{(1)0,1} = 2,706$, aunque no se rechazaría al 5 %. Luego existe cierta evidencia de que $F2$ sea una variable relevante a añadir en el modelo.

¿Cómo cambiarían los gráficos (5.1) y (5.2) si consideramos los residuos del modelo (5.1) que incluye a la variable F2? Estos corresponden a los gráficos de la Figura (5.3). En este caso la disposición de los residuos positivos y negativos es más aleatoria alrededor de su media muestral. Por otro lado, el gráfico de los residuos del modelo (5.1) sobre la variable $F2$ ya no muestra esa relación positiva entre ambas variables.

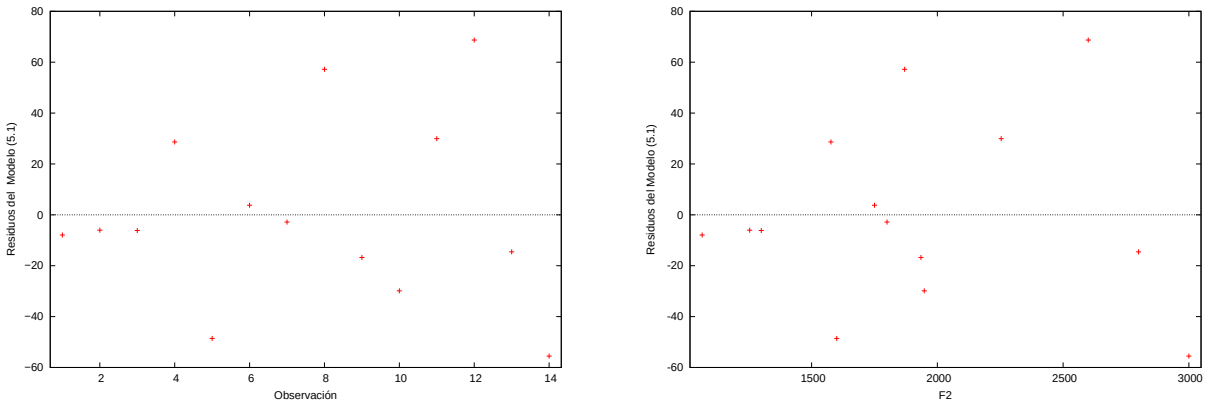


Gráfico 5.3: Gráficos de los residuos del Modelo (5.1) sobre observación y sobre F2

5.3. Efectos de inclusión de variables irrelevantes

Supongamos ahora que el modelo correcto para el precio de la vivienda es

$$P_i = \beta_1 + \beta_2 F2_i + u_i \quad i = 1, \ldots, N \tag{5.9}$$

donde se satisfacen las hipótesis básicas, pero incluimos en la regresión una variable más que no es relevante, $BEDRMS$. El modelo que ajustamos es

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + u_i \quad i = 1, \ldots, N \tag{5.10}$$

En este modelo se siguen satisfaciendo las hipótesis básicas, ya que el valor poblacional del coeficiente que acompaña a la variable $BEDRMS$ es cero al no ser una variable relevante, por lo que el término de error no cambia. Pero en la regresión se estimarán todos los coeficientes, también los de las variables irrelevantes y la estimación puntual de β_3 no será en general cero. ¿Qué consecuencias tendrá este error de especificación?

- En este caso, los estimadores de todos los coeficientes son insesgados, por lo que $E(\hat{\beta}_j) = \beta_j \ \forall j$. En particular, $E(\hat{\beta}_3) = 0$.

- La matriz de varianzas y covarianzas se estimará correctamente con el estimador habitual. Por lo que tanto los intervalos de confianza como los procedimientos habituales de contraste sobre los coeficientes β_j siguen siendo válidos.

- El coste de este error de especificación es la pérdida de eficiencia en la estimación. Si se comparan las varianzas de los coeficientes estimados en el modelo incorrecto relativamente al correctamente especificado, estas serán mayores en el primero. Por ejemplo, se puede demostrar que esta pérdida de eficiencia depende de la correlación entre $F2$ y $BEDRMS$ siendo mayor cuanto mayor sea esta correlación.

  En particular, para β_2 el ratio de la varianza del estimador de este coeficiente en el modelo incorrecto (5.10) sobre la varianza del estimador en el modelo correcto (5.9) es

$$\frac{var(\hat{\beta}_2)_{(10)}}{var(\hat{\beta}_2)_{(9)}} = \frac{1}{\left(1 - \rho_{23}^2\right)} \geq 1 \tag{5.11}$$

siendo $0 \leq \rho_{23}^2 \leq 1$ el coeficiente de correlación al cuadrado entre $F2$ y $BEDRMS$. En el caso de los datos que estamos utilizando *data4-1* sobre 14 viviendas este ratio es $\left(1/\left(1-(0,5323)^2\right)\right) = 1,4$, luego hay cierta pérdida de eficiencia en la estimación de β_2 en el modelo (5.10) relativamente a (5.9). La inclusión de la variable supuestamente irrelevante $BEDRMS$ hace que estimemos con menor precisión el coeficiente β_2. Lo mismo ocurre con el coeficiente β_1.

¿Cómo podemos detectar la presencia de variables innecesarias?

Una posibilidad es comenzar por un modelo relativamente general y utilizar los contrastes de significatividad individual, así como las medidas de bondad de ajuste $\bar{R}^2$ o los criterios de información AIC o BIC por ejemplo. Estos indicadores nos pueden ayudar en la toma de esta decisión. Los resultados obtenidos de la estimación de los modelos (5.9) y (5.10) se muestran en la tabla (5.2)[6]. Considerando que nuestro modelo de partida es el modelo más general, **Modelo (5.10)**, y utilizando el contraste de significatividad individual para el coeficiente que acompaña a $BEDRMS$, podríamos considerar que esta variable no es relevante en explicar la variación en el precio de la vivienda una vez hemos incluido el tamaño de ésta. Eliminar esta variable del modelo también mejora el resto de indicadores de ajuste, mayor $\bar{R}^2$, menores AIC y BIC. Se puede observar también que las desviaciones típicas estimadas se reducen bastante. Por otro lado, tanto en el modelo (5.10) como en el (5.9), la variable $F2$ es significativa indicando su relevancia en explicar la variación en el precio de la vivienda.

Variable	**Modelo (5.9)**	**Modelo (5.10)**
	supuestamente correcto	
CONSTANT	52,351	121,179
	(1,404)	(1,511)
	[37,28]	[80,1778]
F2	0,13875	0,14831
	(7,407)	(6,993)
	[0,0187]	[0,0212]
BEDRMS		-23,911
		(-0,970)
		[24,642]
Suma de cuadrados de los residuos	18273,6	16832,8
Desviación típica de los residuos ($\hat{\sigma}$)	39,023	39,1185
R^2	0,821	0,835
$\bar{R}^2$	0,806	0,805
F de significación conjunta	54,861	27,767
Grados de libertad	12	11
Criterio de Akaike (AIC)	144,168	145,019
Criterio de Schwarz (BIC)	145,447	146,936

Tabla 5.2: Modelos estimados para el precio de la vivienda.

[6]Entre paréntesis estadísticos t y entre corchetes las desviaciones típicas estimadas.

La aproximación de ir de un modelo más general a uno más restringido suele ser más conveniente que la aproximación contraria. En el caso de comenzar por un modelo más reducido e ir añadiendo variables secuencialmente, decidiendo mantenerlas o no en función de si son o no significativas, se corre el peligro de lo que se conoce con el nombre inglés de *data mining* o *torturar a los datos*.

El problema en la aproximación contraria es que, si el modelo de partida es demasiado general y los regresores están muy correlacionados, la precisión con la que estimemos los parámetros puede ser poca. Por esa falta de precisión en la estimación podemos tener coeficientes no significativamente distintos de cero, no siendo capaces de identificar el efecto de esas variables ya que la potencia de los contrastes de significación puede ser muy poca[7]. No rechazar en ese caso la hipótesis nula no es evidencia de que esas variables no sean relevantes sino de que el contraste tiene poca potencia.

[7]Este problema será tratado más en detalle en el tema de Multicolinealidad.

Bibliografía

Engle, R. F. (1982), "A general approach to Lagrangian Multiplier Modelo Diagnostics", *Journal of Econometrics*, vol. 20, pp. 83-104.

Tema 6

Multicolinealidad

Contenido

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. Por ejemplo, la población y el PIB en general suelen estar altamente correlacionados. A este fenómeno se le conoce como multicolinealidad. En algún caso puede que los datos de una variable se obtengan como resultado de una identidad contable o de una combinación lineal exacta entre otros regresores. Este último caso se denomina de multicolinealidad exacta o perfecta.

Cuando dos o más variables explicativas en 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 el caso extremo de multicolinealidad exacta no es posible estimar separadamente estos efectos sino una combinación lineal de ellos. En este tema analizaremos las implicaciones que tiene en la estimación por el método de Mínimos Cuadrados Ordinarios este fenómeno muestral.

6.1. Multicolinealidad perfecta

Dada la especificación del modelo y los datos de las variables, si al menos una de las variables **explicativas** se puede obtener como combinación lineal exacta de alguna o algunas de las restantes, diremos que existe multicolinealidad exacta o perfecta.

Consideremos el siguiente ejemplo. ¿Qué ocurrirá si definimos una nueva variable $F25$ que es una combinación lineal exacta de otra variable explicativa en el modelo, $F25 = 5 \times F2$ y pretendemos estimar los parámetros del siguiente modelo?

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 F25_i + u_i \qquad i = 1, 2, \ldots, N \tag{6.1}$$

Las variables $F25$ y $F2$ son combinación lineal exacta por lo que el rango de la matriz X es $3 = K - 1$, menor que el número de parámetros a estimar, ya que la cuarta columna se obtiene de multiplicar por 5 la segunda columna. El sistema de ecuaciones normales que se obtiene del criterio de estimación del método de Mínimos Cuadrados Ordinarios sería un sistema de cuatro ecuaciones pero solamente tres serán linealmente independientes[1].

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \hat{\beta}_4 \sum X_{4i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} + \hat{\beta}_4 \sum X_{4i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \hat{\beta}_4 \sum X_{4i} X_{3i} \\
\sum Y_i X_{4i} &= \hat{\beta}_1 \sum X_{4i} + \hat{\beta}_2 \sum X_{2i} X_{4i} + \hat{\beta}_3 \sum X_{3i} X_{4i} + \hat{\beta}_4 \sum X_{4i}^2
\end{aligned}$$

Si sustituimos en estas ecuaciones la relación lineal exacta $X_{4i} = 5X_{2i}$ y reorganizamos,

[1] La notación utilizada es $Y_i \equiv P_i$, $X_{2i} \equiv F2_i$, $X_{3i} \equiv BEDRMS_i$, $X_{4i} \equiv F25_i$.

obtenemos:

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i} + \hat{\beta}_3 \sum X_{3i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \\
5\,[\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + (\hat{\beta}_2 + 5\hat{\beta}_4)\sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i}]
\end{aligned}$$

Se puede observar que la cuarta ecuación es la misma que la segunda excepto por un factor de escala igual a 5. Por lo tanto, hay cuatro incógnitas $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\beta}_4$ pero solamente tres ecuaciones linealmente independientes. Consecuentemente, no es posible estimar de forma única todos los coeficientes del modelo. Ahora bien, las tres primeras ecuaciones si podemos resolverlas para $\hat{\beta}_1$, $\hat{\beta}_3$ y la combinación lineal $(\hat{\beta}_2 + 5\hat{\beta}_4)$. Esto mismo se puede comprobar sustituyendo $F25_i = 5 \times F2_i$ en el modelo (6.1).

$$P_i = \beta_1 + (\beta_2 + 5\beta_4)\, F2_i + \beta_3\, BEDRMS_i + u_i \qquad i = 1, 2, \ldots, N \tag{6.2}$$

Vemos que en esta regresión son estimables de forma separada y única los coeficientes β_1 y β_3 pero no β_2 y β_4. El coeficiente que acompaña a $F2_i$ recogería la combinación lineal $\beta_2 + 5\beta_4$.

¿Qué hace el programa GRETL si hay multicolinealidad perfecta? Elimina una variable cualquiera de las que forman parte de esa relación exacta, mostrando el siguiente resultado.

```
Modelo 8: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: P

Omitidas debido a colinealidad exacta: F25

     VARIABLE       COEFICIENTE       DESV.TÍP.      ESTAD T      VALOR P
  const               121,179           80,1778        1,511   0,15888
  F2                  0,148314         0,0212080     6,993   0,00002 ***
  BEDRMS              -23,9106          24,6419       -0,970   0,35274

  Media de la var. dependiente = 317,493
  Desviación típica de la var. dependiente. = 88,4982
  Suma de cuadrados de los residuos = 16832,8
  Desviación típica de los residuos = 39,1185
  R-cuadrado = 0,834673
  R-cuadrado corregido = 0,804613
  Estadístico F (2, 11) = 27,7674 (valor p = 5,02e-005)
  Log-verosimilitud = -69,5093
  Criterio de información de Akaike (AIC) = 145,019
  Criterio de información Bayesiano de Schwarz (BIC) = 146,936
  Criterio de Hannan-Quinn (HQC) = 144,841
```

Por lo tanto, **avisa** de que ha eliminado una variable explicativa de la regresión, en este caso $F25$, y muestra los resultados de la regresión excluyendo esa variable. De hecho, el coeficiente que acompaña a F2 podría considerarse como $\widehat{(\beta_2 + 5\beta_4)}$. Este ha sido un ejemplo ilustrativo de las implicaciones que tiene el problema de multicolinealidad perfecta.

6.2. Multicolinealidad de grado alto

En general es difícil tener en un modelo de regresión variables explicativas o regresores que no presenten cierta correlación muestral. La multicolinealidad, de no ser perfecta, se puede considerar un problema cuando la correlación entre los regresores es tan alto que se hace casi imposible estimar con precisión los efectos individuales de cada uno de ellos.

Si la correlación entre la variables explicativas es alta, es común tener los siguientes síntomas:

- Pequeños cambios en los datos o en la especificación provocan grandes cambios en las estimaciones de los coeficientes.
- La estimaciones de los coeficientes suelen presentar signos distintos a los esperados y magnitudes poco razonables.
- El efecto más pernicioso de la existencia de un alto grado de multicolinealidad es el de incrementar las varianzas de los coeficientes estimados por MCO. Es decir, es difícil estimar separadamente los efectos marginales o individuales de cada variable explicativa por lo que estos se estiman con poca precisión.[2] Como consecuencia, el valor del estadístico para realizar contrastes de significatividad individual tiende a ser pequeño y aumenta la probabilidad de no rechazar la hipótesis nula, por lo que se tiende a concluir que las variables no son significativas individualmente. El problema no reside en que los contrastes no sean correctos estadísticamente, sino en que no estimamos con suficiente precisión estos efectos individuales.
- Se obtienen valores altos del R^2 aún cuando los valores de los estadísticos t de significatividad individual son bajos. El problema reside en la identificación del efecto individual de cada variable explicativa, no tanto en su conjunto. Por eso, si se realiza un contraste de significatividad conjunta de las variables explicativas, el resultado normalmente será rechazar la hipótesis nula por lo que conjuntamente son significativas aunque individualmente cada una de ellas no lo sea.

Si se presentan estos síntomas se puede sospechar que el problema de multicolinealidad esté afectando a nuestros resultados, especialmente a la inferencia sobre los efectos individuales de cada variable explicativa. De todas formas es importante analizar e interpretar adecuadamente los resultados obtenidos sin tomar conclusiones precipitadamente.

¿Cómo podemos analizar si existe un problema de multicolinealidad?

1) 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.

   Utilizando el ejemplo de los precios de los pisos (Fichero de muestra del Ramanathan *data4-1*) con las variables que ya analizamos en temas anteriores,

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$$

[2] Los estimadores MCO siguen siendo los de menor varianza dentro de la clase de lineales e insesgados si las hipótesis básicas se satisfacen. Luego no es un problema de pérdida de eficiencia relativamente a otro estimador lineal e insesgado.

obtenemos los siguientes valores de los coeficientes de correlación:

```
Coeficientes de correlación, usando las observaciones 1 - 14
          valor crítico al 5% (a dos colas) = 0,5324 para n = 14

           P            F2          BEDRMS         BATHS
      1,0000        0,9058          0,3156        0,6696  P
                    1,0000          0,4647        0,7873  F2
                                    1,0000        0,5323  BEDRMS
                                                  1,0000  BATHS
```

Como podemos observar, todas las variables explicativas presentan cierto grado de correlación dos a dos, siendo la correlación mayor entre F2 y BATH con un coeficiente igual a 0,7873. Excepto por este valor, no parece que los coeficientes de correlación simple sean demasiado grandes para sospechar que haya un problema de multicolinealidad. De todas formas, aunque es condición suficiente para que exista este problema que todos estos coeficientes fueran altos, lo contrario no necesariamente es cierto. Se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y sin embargo las correlaciones simples entre pares de variables no ser mayores que $0,5$.

2) Otra forma de **detectar la multicolinealidad** consiste en realizar la regresión de cada una de las variables explicativas sobre el resto[3] 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.

   Siguiendo con el ejemplo sobre el modelo del precio de la vivienda, esto consistiría en realizar las siguientes regresiones:

```
Modelo 1: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: F2

  VARIABLE     COEFICIENTE       DESV.TÍP.      ESTAD T  VALOR P
  const        -657,612          809,640        -0,812   0,43389
  BEDRMS         73,9671         254,175         0,291   0,77646
  BATHS         975,371          283,195         3,444   0,00548 ***

  R-cuadrado = 0,622773

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: BEDRMS

  VARIABLE     COEFICIENTE       DESV.TÍP.     ESTAD T  VALOR P
  const         2,29560          0,700852       3,275   0,00739 ***
  F2            0,000103288      0,000354931    0,291   0,77646
  BATHS         0,487828         0,459485       1,062   0,31113
```

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

```
  R-cuadrado = 0,288847

Modelo 3: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: BATHS

  VARIABLE    COEFICIENTE       DESV.TÍP.    ESTAD T  VALOR P
  const        0,646527         0,583914      1,107   0,29182
  F2           0,000531961      0,000154452   3,444   0,00548 ***
  BEDRMS       0,190531         0,179461      1,062   0,31113

  R-cuadrado = 0,655201
```

Los resultados parecen mostrar que las variaciones muestrales de las variables $F2$ y $BATHS$ son las más explicadas por el resto de variables explicativas, aunque los coeficientes de determinación de esas dos regresiones no son excesivamente altos; alrededor de un 60 % de la variación de $F2$ y de $BATHS$ vienen explicadas por variaciones en el resto de variables explicativas. Si recordamos los resultados obtenidos en el Tema 3, donde al estimar el modelo 3 una vez que incluíamos $F2$ en la regresión, obteníamos que las variables $BATH$ y $BEDRMS$ no eran significativas. ¿Puede ser este hecho consecuencia de un problema de multicolinealidad? ¿Podríamos tener problemas de multicolinealidad entre las variables $F2, BATHS$ y $BEDRMS$? Vamos a utilizar algún procedimiento más formal para detectar si existe este problema.

3) Neter, Wasserman & Kutner (1990) 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:

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

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 depende 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.

Los indicadores TOL y VIF se pueden obtener con el programa GRETL de forma muy sencilla. Siguiendo con el ejemplo de los precios de las viviendas, calcularemos la Inflación de la Varianza para analizar la posible presencia de multicolinealidad. Para ello, en la ventana de la estimación por MCO del modelo de interés, elegimos la opción

Contrastes → *Colinealidad*

obteniendo la siguiente información:

```
Factores de inflación de varianza (VIF)

 Mínimo valor posible = 1.0
 Valores mayores que 10.0 pueden indicar un problema de colinealidad

   2)            F2      2,651
   3)          BEDRMS    1,406
   4)           BATHS    2,900

VIF(j) = 1/(1 - R(j)^2), donde R(j) es el coeficiente de
correlación múltiple entre la variable j y las demás variables
independientes
```

Como podemos observar, según los valores del VIF_j, podríamos concluir que no existen problemas de multicolinealidad.

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.

- Quitar del modelo alguna de las variables colineales. Es una medida que puede provocar otro tipo de problemas, ya que si la variable que eliminamos del modelo realmente sí es significativa, estaremos omitiendo una variable relevante. Por consiguiente, los estimadores de los coeficientes del modelo y de su varianza serían sesgados por lo que la inferencia realizada no sería válida.

- Existen otros métodos de estimación sugeridos en la literatura econométrica que mejorarían la estimación en términos de eficiencia o precisión, pero los estimadores así obtenidos serían sesgados. Explicar estos métodos no entran dentro de los objetivos de este curso.

Bibliografía

Neter, J., Wasserman, W. y M. H. Kutner (1990), *Applied Linear Statistical Models*, 3ª edn., M.A: Irwin.

Tema 7

Variables Cualitativas

Contenido

7.1. Introducción. Un ejemplo

A lo largo del curso únicamente se han especificado 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 este tema, aunque seguimos manteniendo que la variable dependiente es cuantitativa, vamos a considerar que ésta puede venir explicada por variables cualitativas y/o cuantitativas.

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].

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

En este tema estudiamos la estimación, interpretación de los coeficientes y contrastes de hipótesis en modelos con presencia de variables cualitativas como regresores.

7.2. Modelo con una variable cualitativa

Consideremos el caso más sencillo, una variable cualitativa como único regresor del modelo. Vamos a suponer que queremos explicar el precio de la vivienda basándonos únicamente en si la vivienda tiene piscina o no[2]. Para ello, definimos la siguiente variable ficticia:

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

Abrimos el fichero de datos *data7-3* de Ramanathan (2002), 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), utilizados en capítulos anteriores y añade 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

[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.

[2] Por simplicidad vamos a ignorar el efecto del resto de variables que afectan al precio de la vivienda.

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. Para ello, 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[3]. A continuación, se obtienen los estadísticos principales:

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

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

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

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.1}$$

Interpretación y estimación de los coeficientes

En nuestro ejemplo, la función de regresión poblacional varía en función de si la vivienda tiene piscina o no:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2$, puesto que la variable POOL toma el valor 1 y $E(u_i) = 0$.
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1$, puesto que la variable POOL toma el valor 0 y $E(u_i) = 0$.

Por tanto, los coeficientes se interpretan como sigue:

- 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.

Utilizando las ecuaciones normales que derivamos en el Tema 2 para estimar el modelo de regresión simple y teniendo en cuenta que al ser POOL una variable ficticia que toma valores 0 y 1 coincide con su cuadrado, obtenemos que los estimadores de los coeficientes del modelo (7.1) se pueden calcular a partir de simples medias muestrales[4]:

- $\hat{\alpha}_1 = \overline{PRICE}_{nopool} = 306{,}111 \quad \Rightarrow$ precio estimado medio de las viviendas sin piscina.
- $\hat{\alpha}_2 = \overline{PRICE}_{pool} - \overline{PRICE}_{nopool} = 337{,}980 - 306{,}111 = 31{,}869 \quad \Rightarrow$ diferencia estimada en el precio medio de las viviendas con piscina con respecto a las que no la tienen.

En efecto, si estimamos el modelo por Mínimos Cuadrados Ordinarios utilizando Gretl obtenemos que las estimaciones de los coeficientes son las siguientes:

Modelo (7.1): 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

Que coinciden con las calculadas utilizando los valores obtenidos en ambas submuestras mediante los Estadísticos Principales:

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

[4] $\overline{PRICE}_{pool}$ es la media muestral del precio de las viviendas con piscina, de igual forma $\overline{PRICE}_{nopool}$ es la media muestral del precio de las viviendas sin piscina.

El modelo (7.1) no es la única especificación correcta posible para explicar las variaciones del precio de la vivienda en función de si tiene piscina o no. Al igual que hemos definido la variable ficticia POOL, podemos crear la variable NOPOOL, tomando el valor 1 si la vivienda no tiene piscina y 0 en caso contrario. Con esta nueva variable podemos especificar los dos modelos siguientes:

$$PRICE_i = \gamma_1 + \gamma_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.2}$$

$$PRICE_i = \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.3}$$

La interpretación de los coeficientes se haría de forma análoga a como hemos visto para el modelo (7.1). Notar que la equivalencia entre los coeficientes de los distintos modelos (7.1), (7.2) y (7.3) es la siguiente:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 = \gamma_1 = \beta_1$
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1 = \gamma_1 + \gamma_2 = \beta_2$

Una especificación que no sería adecuada es la siguiente:

$$PRICE_i = \alpha + \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14$$

ya que si analizamos la matriz de datos X para este modelo observamos que la suma de la segunda y tercera columnas es igual a la primera y tendríamos un problema de multicolinealidad exacta, por lo que la matriz $X'X$ no sería invertible. En estas circunstancias no se podría obtener una única solución para $\hat{\alpha}, \hat{\beta}_1$ y $\hat{\beta}_2$ del sistema de ecuaciones normales.

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

Contraste de hipótesis

Los contrastes de hipótesis se realizan con la metodología estudiada en los capítulos previos. Por ejemplo, si quisiéramos contrastar en el modelo (7.1) 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 $H_0 : \alpha_2 = 0$.[5] Este contraste se puede realizar utilizando el estadístico t habitual cuyo *valor-p* es 0,5402, 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 (7.1) el modelo no restringido mientras que el modelo restringido es $PRICE_i = \alpha_1 + u_i \quad i = 1, \ldots, 14$.

7.2.1. Incorporación de variables cuantitativas

En el modelo (7.1) el único regresor para explicar el precio de la vivienda es una característica cualitativa, el hecho de tener o no piscina sin embargo, en un modelo pueden convivir variables cualitativas y cuantitativas. Vamos a comenzar añadiendo un regresor cuantitativo, la variable SQFT (número de pies cuadrados habitables de la vivienda) y manteniendo la variable ficticia POOL afectando a la ordenada.

Cambio en la ordenada

Suponer que el precio de la vivienda únicamente depende de si tiene piscina o no es poco realista, por lo que añadimos como regresor a la variable cuantitativa SQFT (número de pies cuadrados habitables de la vivienda) de la siguiente manera:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta\, SQFT_i + u_i \quad i = 1, \ldots, 14 \tag{7.4}$$

Estimación e interpretación de los coeficientes:

La función de regresión poblacional se puede expresar como:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 + \beta\, SQFT_i$
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1 + \beta\, SQFT_i$

Por tanto podemos interpretar α_1 como el precio esperado de una vivienda sin piscina y cero pies cuadrados, α_2 como el diferencial en el precio esperado en una vivienda por el hecho de tener piscina, manteniendo el número de pies cuadrados habitables constante. A igual número de pies cuadrados habitables el hecho de tener piscina se puede considerar una mejora en la vivienda por lo que sería preferida, así tener piscina es una característica que sube el precio de la vivienda y esperaríamos que α_2 tuviese signo positivo. Finalmente interpretamos β como la variación en el precio esperado de una vivienda por incrementar su superficie en un pie cuadrado. Esperaríamos signo positivo, a mayor superficie mayor precio esperado para la vivienda. Gráficamente, obtenemos dos rectas con igual pendiente, β, y distinta ordenada como podemos observar en el Gráfico 7.1:

[5] Equivalentemente, $H_0 : \gamma_2 = 0$ ó $H_0 : \beta_1 = \beta_2$ para los modelos (7.2) y (7.3), respectivamente.

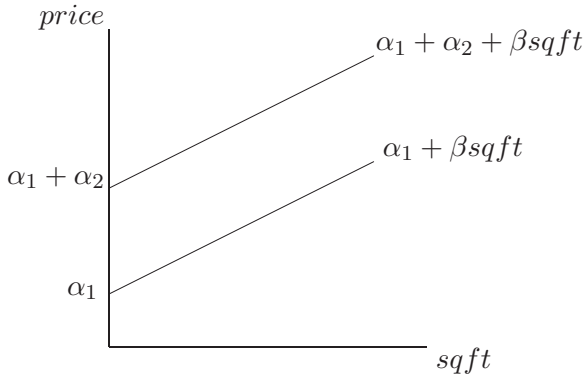


Gráfico 7.1: Cambio en ordenada

El resultado de la estimación del modelo (7.4) por Mínimos Cuadrados Ordinarios es:

Modelo (7.4): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	22,6728	29,5058	0,7684	0,4584
pool	52,7898	16,4817	3,2029	0,0084
sqft	0,144415	0,0141849	10,1809	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	9455,36
Desviación típica de los residuos ($\hat{\sigma}$)	29,3186
R^2	0,907132
$\bar{R}^2$ corregido	0,890247
$F(2, 11)$	53,7238
Log-verosimilitud	−65,472
Criterio de información de Akaike	136,944
Criterio de información Bayesiano de Schwarz	138,861

El modelo estimado es:

$$\underset{(estad.\,t)}{\widehat{PRICE}_i} = \underset{(0,768)}{22,673} + \underset{(3,203)}{52,790} POOL_i + \underset{(10,181)}{0,144}\, SQFT_i$$

donde se puede observar que ambos regresores son significativos para explicar el precio medio de la vivienda y tienen los signos adecuados[6]. Por tanto, existen diferencias significativas en el precio medio de la vivienda que tiene piscina con respecto a la que no la tiene.

Los coeficientes estimados se interpretan como sigue:

[6]El valor de los estadísticos t para los coeficientes de ambos regresores es superior al valor crítico de una distribución t de Student de $N - K = 14 - 3 = 11$ grados de libertad para un nivel de significación del 5 %, que es 2,201.

- $\hat{\alpha}_1 = 22,673 \Rightarrow$ el precio medio estimado de las viviendas sin piscina y con cero pies cuadrados habitables es 22.673 dólares.
- $\hat{\alpha}_2 = 52,790 \Rightarrow$ se estima que entre dos viviendas con el mismo número de pies cuadrados habitables el precio medio de una con piscina es 52.790 dólares más caro que el de una sin piscina.
- $\hat{\beta} = 0,144 \Rightarrow$ el precio medio estimado de una vivienda se incrementa en 144 dólares al aumentar en un pie cuadrado habitable la vivienda.

Cambio en la ordenada y en la pendiente

También es posible pensar que la variación en el precio de las viviendas ante el incremento en un pie cuadrado habitable sea diferente para aquéllas que tienen piscina. En este caso se especifica el siguiente modelo, donde la variable ficticia POOL afecta tanto a la ordenada como a la pendiente de la recta:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta_1 SQFT_i + \beta_2 POOL \cdot SQFT_i + u_i \qquad i = 1, \ldots, 14 \qquad (7.5)$$

La interacción $POOL \cdot SQFT$ mide el número de pies cuadrados habitables para las viviendas que tienen piscina, mientras que toma el valor 0 para las que no la tienen.

Estimación e interpretación de los coeficientes:

Una vez definida la interacción $POOL \cdot SQFT$ en Gretl, estimamos el modelo (7.5):

Modelo (7.5): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	77,1332	25,6379	3,0086	0,0131
pool	−82,648	39,7759	−2,0779	0,0644
sqft	0,116667	0,0125934	9,2641	0,0000
pool· sqft	0,0722955	0,0203274	3,5566	0,0052

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	4174,72
Desviación típica de los residuos ($\hat{\sigma}$)	20,4321
R^2	0,958997
$\bar{R}^2$ corregido	0,946696
$F(3, 10)$	77,9615
Log-verosimilitud	−59,749
Criterio de información de Akaike	127,499
Criterio de información Bayesiano de Schwarz	130,055

La función de regresión poblacional se puede expresar como:

- $E(PRICE_i | i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 + (\beta_1 + \beta_2) SQFT_i$

- $E(PRICE_i|i \text{ es una vivienda sin piscina}) = \alpha_1 + \beta_1 SQFT_i$

El parámetro poblacional α_1 se interpreta como el precio esperado de una vivienda sin piscina y con cero pies cuadrados habitables. α_2 mide el diferencial en el precio esperado de una vivienda con cero pies cuadrados habitables por el hecho de tener piscina. Esperaríamos que ambos coeficientes tuviesen signo positivo por las razones argumentadas anteriormente.

β_1 se interpreta como la variación en el precio esperado de una vivienda sin piscina por incrementar su superficie en un pie cuadrado habitable mientras que β_2 mide el diferencial en la variación en el precio esperado de una vivienda ante un incremento de su superficie en un pie cuadrado por el hecho de tener piscina. Esperaríamos que ambos coeficientes tuviesen signo positivo, a mayor superficie de la vivienda mayor precio esperado. Si además la vivienda tiene piscina el cambio en el precio esperado por pie cuadrado más de superficie será mayor ya que la posesión de piscina es una mejora.

La representación gráfica corresponde a dos rectas que varían tanto en el punto de corte con el eje de ordenadas como en la pendiente:

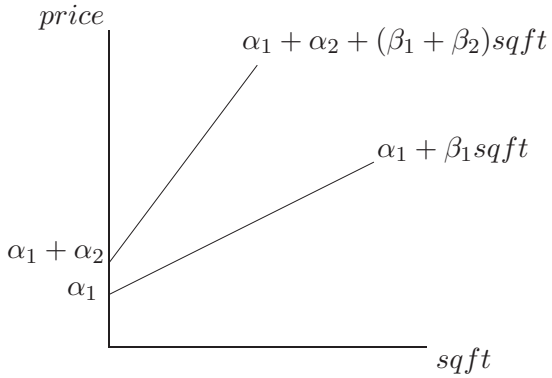


Gráfico 7.2: Cambio en ordenada y en pendiente

Interpretación de los coeficientes estimados:

- $\hat{\alpha}_1 = 77,133 \Rightarrow$ el precio medio estimado de las viviendas que no tienen piscina y con cero pies cuadrados habitables es 77.133 dólares.
- $\hat{\alpha}_2 = -82,648 \Rightarrow$ entre dos viviendas con 0 pies cuadrados habitables el precio medio estimado de una con piscina es 82.648 dólares más barato que el de una sin piscina.
- $\hat{\beta}_1 = 0,117 \Rightarrow$ al incrementar en un pie cuadrado la superficie habitable, el precio medio estimado de una vivienda sin piscina aumenta en 117 dólares.
- $\hat{\beta}_2 = 0,072 \Rightarrow$ al incrementar en un pie cuadrado la superficie habitable, el precio medio estimado de una vivienda con piscina aumenta en 72 dólares.

Contraste de hipótesis

La hipótesis nula para contrastar si tener piscina influye significativamente en el precio medio de las viviendas es $H_0 : \alpha_2 = \beta_2 = 0$. El resultado del contraste es:

Contraste de omisión de variables –
  Hipótesis nula: los parámetros son cero para las variables
    pool
    poolsqft
  Estadístico de contraste: $F(2, 10) = 16{,}886$
  con valor p = $P(F(2, 10) > 16{,}886) = 0{,}000622329$

por lo que se rechaza la hipótesis nula para un nivel de significación del 5 % y por lo tanto tener piscina es una variable significativa para explicar el precio de las viviendas.

También se puede contrastar mediante un contraste de significatividad individual si el incremento en un pie cuadrado de superficie afecta al precio de manera diferente según la vivienda tenga o no piscina, para ello podemos contrastar $H_0 : \beta_2 = 0$. Como vemos en los resultados de la estimación del modelo este coeficiente es significativo, como esperábamos la influencia de la superficie habitable de una vivienda en su precio varía si la vivienda tiene piscina o no. Por otro lado, $\hat{\alpha}_2$ no tiene el signo esperado y a su vez no es significativo a nivel individual, aparentemente el hecho de incluir la variable ficticia en la pendiente ha restado significatividad a la discriminación en la ordenada.

7.3. Modelo con dos o más variables cualitativas

Al igual que ocurría con los regresores cuantitativos sobre una variable endógena pueden influir más de una variable cualitativa. Por ejemplo en el precio de una vivienda podría influir no sólo el hecho de tener o no piscina, su superficie habitable, el número de habitaciones, el número de baños, si no también si tiene o no chimenea, si tiene o no ascensor o la zona de la ciudad donde esté situada.

7.3.1. Varias categorías

Supongamos que creemos que la zona de la ciudad donde esté situada la vivienda es un determinante de su precio. Pensemos por ejemplo en precios de viviendas situadas en una gran ciudad en la que podemos distinguir como zonas a la zona centro, zona norte, zona sur, zona este y zona oeste. En general el centro de las ciudades es una zona valorada por ser el centro neurálgico económico-comercial y el resto de zonas se valorará en función del tipo de viviendas que recoja y sus comunicaciones, por ejemplo en una ciudad como Madrid esperaríamos mayor precio en el centro, norte y oeste que en el sur o en el este que agrupan a barrios, en general, con menor nivel económico y peor comunicados. Para el ejemplo necesitamos definir cinco variables ficticias una para cada zona ya que la situación geográfica de la vivienda la hemos

dividido en cinco categorías[7].

Definimos las siguiente variables:

$$D_{1i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona centro} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{2i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona norte} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{3i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona sur} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{4i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona este} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{5i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona oeste} \\ 0 & \text{en caso contrario} \end{cases}$$

Si además de la situación geográfica de la vivienda creemos que la superficie habitable influye en su precio podemos definir, por ejemplo, el siguiente modelo:

$$PRICE_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N \quad (7.6)$$

Donde β se interpreta de la forma habitual y α_1 se interpreta como el precio esperado de una vivienda con cero pies cuadrados situada en la zona centro, así $\alpha_i \quad i = 1, \ldots, 5$ se interpretan como el precio esperado de una vivienda con cero pies cuadrados situadas en la zona correspondiente, centro, norte, sur, este u oeste.

En la especificación (7.6) se ha optado por no incluir término independiente en el modelo e incluir las cinco variables ficticias para no incurrir en un problema de multicolinealidad exacta como se expuso en el punto anterior pero, podríamos especificar un modelo con término independiente siempre y cuando dejemos fuera una de las variables ficticias o categorías para no tener dicho problema. Por ejemplo una especificación alternativa sería:

$$PRICE_i = \alpha + \alpha_2^\star D_{2i} + \alpha_3^\star D_{3i} + \alpha_4^\star D_{4i} + \alpha_5^\star D_{5i} + \beta\, SQFT_i + u_i \qquad i = 1, \ldots, N \quad (7.7)$$

En el modelo anterior la interpretación del parámetro poblacional β no varía, α se interpreta como el precio esperado de una vivienda con cero pies cuadrados situada en la zona centro, $\alpha_i^\star \quad i = 2, \ldots, 5$ se interpretan como el diferencial en el precio esperado de una vivienda, a igual superficie habitable, por estar situada en la zona norte, (sur, este y oeste respectivamente) con respecto a una vivienda situada en la zona centro. Qué variable ficticia (o categoría) dejemos fuera no es relevante siempre y cuando interpretemos adecuadamente los parámetros. Naturalmente podemos afectar las variables ficticias a la variable cuantitativa como en el caso anterior siempre y cuando no incurramos en multicolinealidad exacta.

[7] En el ejemplo anterior la vivienda tenía o no piscina, solo había dos casos posibles y por tanto sólo había dos categorías.

Contraste de hipótesis

Para contrastar en el modelo (7.6) que por ejemplo no existen diferencias significativas en el precio medio de la vivienda por su situación la hipótesis de contraste es $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$. Hipótesis que podemos contrastar utilizando el estadístico F basado en las sumas de cuadrados de los residuos siendo en este caso el modelo (7.6) el modelo no restringido mientras que el modelo restringido sería $PRICE_i = \alpha_1 + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N$. El mismo contraste puede llevarse a cabo en el modelo (7.7) con la hipótesis $H_0 : \alpha_2^\star = \alpha_3^\star = \alpha_4^\star = \alpha_5^\star = 0$ siendo el modelo no restringido el modelo (7.7) y el restringido $PRICE_i = \alpha + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N$.

7.3.2. Varios conjuntos de variables ficticias

Supongamos que ampliamos el modelo (7.4) incorporando regresores que podrían explicar el precio de la vivienda como por ejemplo el número de habitaciones, el número de baños, que la vivienda tenga sala de estar o no y que tenga chimenea o no. 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 el número de baños y el número de habitaciones se definen como en los temas anteriores:

$BEDRMS$ número de habitaciones de la vivienda i-ésima
$BATHS$ número de cuartos de baño de la vivienda i-ésima

Con todas ellas podemos definir el siguiente modelo para explicar el precio de la vivienda:

$$\begin{aligned} PRICE_i = {} & \gamma_1 + \gamma_2 POOL_i + \gamma_3 FAMROOM_i + \gamma_4 FIREPL_i \\ & + \beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \ldots, 14 \end{aligned} \tag{7.8}$$

Donde lo primero a notar es que en el modelo (7.8), 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, γ_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 (7.8): 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, número de habitaciones, número de baños, existencia de sala de estar y/o 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 y el resto de características de la vivienda.

- $\hat{\beta}_2 = -7,046$: el precio medio estimado de una vivienda disminuye en 7.046 dólares al aumentar en 1 el número de habitaciones, permaneciendo constantes el número de baños y los pies cuadrados habitables y el resto de características de la vivienda. 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 el resto de características de la vivienda. 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 *valor-p* 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 (7.4) y (7.8), ninguna de las variables añadidas en el último modelo es significativa individualmente[8]. 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 (7.4)) y no restringido (modelo (7.8)). 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 (7.4) 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.

[8] 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.

7.4. Contraste de cambio estructural

En ocasiones puede ocurrir que la relación entre la variable dependiente y los regresores cambie a lo largo del periodo muestral, es decir, puede que exista un cambio estructural. Por ejemplo, si estamos analizando el consumo de tabaco y durante el período muestral se ha producido una campaña de salud pública informando sobre los peligros que conlleva el consumo de tabaco, podemos pensar que tras dicha campaña el comportamiento de la demanda de tabaco haya cambiado, reduciéndose significativamente. Si esto ocurre no podemos especificar una única función de demanda para todo el período muestral si no que deberíamos especificar dos funciones, una hasta la campaña antitabaco y otra para el período siguiente. Por tanto, ante sospechas de que exista un cambio estructural, debemos de contrastar la estabilidad de los parámetros de nuestra relación.

El contraste de cambio estructural, llamado habitualmente contraste de Chow, puede realizarse de manera sencilla mediante el estadístico de sumas de cuadrados de los residuos sin más que especificar adecuadamente el modelo restringido y el no restringido. También podemos llevarlo a cabo utilizando variables ficticias. Veamos un ejemplo.

El fichero *data7-19* contiene datos para 1960-1988 sobre la demanda de tabaco y sus determinantes en Turquía. Las variables de interés para el ejemplo son las siguientes:

Q: consumo de tabaco por adulto (en kg).
Y: PNB real per cápita en liras turcas de 1968.
P: precio real del kilogramo de tabaco, en liras turcas.
$D82$: variable ficticia que toma valor 1 a partir de 1982.

A mediados de 1981 el gobierno turco lanza una campaña de salud pública advirtiendo de los peligros de salud que conlleva el consumo de tabaco. Nuestro objetivo es determinar si existen cambios en la demanda de tabaco tras la campaña institucional en cuyo caso la especificación:

$$LnQ_t = \alpha + \beta LnY_t + \gamma LnP_t + u_t \qquad t = 1960, \ldots, 1988 \tag{7.9}$$

no es correcta para todo el período muestral y deberíamos especificar dos ecuaciones:

$$LnQ_t = \alpha_1 + \beta_1 LnY_t + \gamma_1 LnP_t + u_{1t} \qquad t = 1960, \ldots, 1981 \tag{7.10}$$
$$LnQ_t = \alpha_2 + \beta_2 LnY_t + \gamma_2 LnP_t + u_{2t} \qquad t = 1982, \ldots, 1988 \tag{7.11}$$

Si existe cambio estructural rechazaríamos $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ y $\gamma_1 = \gamma_2$

Este contraste podemos llevarlo a cabo utilizando el estadístico F basado en las sumas de cuadrados de los residuos siendo en este caso el modelo restringido el recogido en la ecuación (7.9) mientras que el modelo no restringido está constituido por las ecuaciones (7.10) y (7.11). Utilizando Gretl una vez abierto el fichero de datos y tomado las correspondientes transformaciones estimaríamos el modelo (7.9) por MCO y en la ventana de resultados de la estimación elegimos:

Contrastes $\longrightarrow$ *Contraste de Chow*

A la pregunta *Observación en la cual dividir la muestra* contestaríamos *1982* y la correspondiente devolución es:

Modelo (7.9): estimaciones MCO utilizando las 29 observaciones 1960-1988
Variable dependiente: lnQ

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−4,58987	0,724913	−6,332	0,00001***
lnY	0,688498	0,0947276	7,268	0,00001***
lnP	0,485683	0,101394	−4,790	0,00006***

Media de la var. dependiente = 0,784827
Desviación típica de la var. dependiente. = 0,108499
Suma de cuadrados de los residuos = 0,0949108
Desviación típica de los residuos = 0,0604187
R-cuadrado = 0,712058
R-cuadrado corregido = 0,689908
Estadístico F (2, 26) = 32,148 (valor p < 0,00001)
Estadístico de Durbin-Watson = 1,00057
Coef. de autocorr. de primer orden. = 0,489867
Log-verosimilitud = 41,8214
Criterio de información de Akaike (AIC) = -77,6429
Criterio de información Bayesiano de Schwarz (BIC) = -73,541
Criterio de Hannan-Quinn (HQC) = -76,3582

Contraste de Chow de cambio estructural en la observación 1982 -
Hipótesis nula: no hay cambio estructural
Estadístico de contraste: F(3, 23) = 20,1355
con valor p = P(F(3, 23) > 20,1355) = 1,25619e-006

El estadístico calculado es $F_c = 20,135 > F_{0,05(3,23)}$ por lo que rechazamos H_0 para un nivel de significatividad del 5 %, es decir existe cambio estructural, la campaña institucional ha tenido efecto y la demanda de tabaco en Turquía de 1960 a 1988 queda especificada por las ecuaciones (7.10) y (7.11). Los resultados de la estimación mínimo cuadrática de estas ecuaciones son los siguientes:

$$\underset{(estad.\,t)}{\widehat{LnQ}_t} = \underset{(-10,614)}{-5,024} + \underset{(11,587)}{0,735}\,LnY_t - \underset{(-4,227)}{0,381}\,LnP_t \qquad t = 1960,\ldots,1981 \quad SCR_1 = 0,01654$$

$$\underset{(estad.\,t)}{\widehat{LnQ}_t} = \underset{(2,170)}{8,837} - \underset{(-1,941)}{0,953}\,LnY_t + \underset{(0,654)}{0,108}LnP_t \qquad t = 1982,\ldots,1988 \quad SCR_2 = 0,00965$$

7.4.1. Cambio estructural utilizando variables ficticias

Alternativamente, el contraste anterior podríamos haberlo realizado mediante la variable ficticia $D82$ especificando el siguiente modelo donde $t = 60,\ldots,88$:

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + \beta_1^\star D82_t + \beta_2^\star D82_t \cdot LnY_t + \beta_3^\star D82_t \cdot LnP_t + u_t \quad (7.12)$$

En el cual, si existe cambio estructural rechazaríamos $H_0 : \beta_1^\star = \beta_2^\star = \beta_3^\star = 0$. De nuevo el contraste puede realizarse con el estadístico F habitual de sumas residuales donde el modelo no restringido es el (7.12) y el modelo restringido es

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + u_t \tag{7.13}$$

Utilizando Gretl, el proceso después de abierto el fichero de datos, tomado logaritmos y construido las interacciones $D82 \cdot LnY$ y $D82 \cdot LnP$, sería: estimaríamos el modelo (7.12) por MCO y en la ventana de resultados de la estimación haríamos

Contrastes $\longrightarrow$ *Omitir variables*

elegiríamos $D82$, $D82 \cdot LnY$ y $D82 \cdot LnP$ y obtendríamos el siguiente resultado:

Modelo 1: estimaciones MCO utilizando las 29 observaciones 1960-1988
Variable dependiente: lnQ

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−4,58987	0,724913	−6,332	0,00001***
lnY	0,688498	0,0947276	7,268	0,00001***
lnP	0,485683	0,101394	−4,790	0,00006***

Media de la var. dependiente = 0,784827
Desviación típica de la var. dependiente. = 0,108499
Suma de cuadrados de los residuos = 0,0949108
Desviación típica de los residuos = 0,0604187
R-cuadrado = 0,712058
R-cuadrado corregido = 0,689908
Estadístico F (2, 26) = 32,148 (valor p < 0,00001)
Estadístico de Durbin-Watson = 1,00057
Coef. de autocorr. de primer orden. = 0,489867
Log-verosimilitud = 41,8214
Criterio de información de Akaike (AIC) = -77,6429
Criterio de información Bayesiano de Schwarz (BIC) = -73,541
Criterio de Hannan-Quinn (HQC) = -76,3582

Comparación entre el modelo (7.12) y el modelo (7.13):
Hipótesis nula: los parámetros de regresión son cero para las variables
D82
D82Y
D82P
Estadístico de contraste: F(3, 23) = 20,1355, con valor p = 1,25619e-006
De los 3 estadísticos de selección de modelos, 0 han mejorado.

Dado el *valor-p* rechazamos la hipótesis nula para un nivel de significatividad del 5 % y existe cambio estructural. La demanda de tabaco en Turquía de 1960 a 1988 queda mejor especificada por el modelo (7.12). O lo que es lo mismo las ecuaciones (7.10) y (7.11) si no utilizamos

la variable ficticia $D82$ en la especificación del modelo. Notar que ambas especificaciones son idénticas, son dos formas alternativas y por lo tanto equivalentes de especificar la demanda de tabaco en Turquía para ese periodo temporal.

Bibliografía

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.

*

A.1. Repaso de probabilidad

Las variables económicas tienen un componente sistemático y otro aleatorio, ya que con anterioridad a su observación no podemos predecir con certeza los valores que van a tomar. Este apartado revisa los conceptos de probabilidad que aplicaremos este curso: qué es una variable aleatoria o *estocástica*, cuáles son sus propiedades y, finalmente, se presentan las distribuciones de probabilidad más usuales.

A.1.1. Una variable aleatoria

Una *variable aleatoria*, que denotamos por X, es aquella cuyo valor no es conocido con anterioridad a su observación. La probabilidad es un medio para expresar la incertidumbre sobre el resultado. Se distinguen dos tipos de variables aleatorias: *discretas*, cuando el conjunto de todos sus posibles valores es finito o infinito numerable, y *continuas*, cuando el conjunto de realizaciones es infinitamente divisible y, por tanto, no numerable. Por ejemplo, la superficie de una vivienda es una variable continua mientras que el número de baños es una variable discreta. En general, en este curso nos ocuparemos de variables continuas.

Si X es una variable discreta, podemos asignar una probabilidad $p(x_i) = Prob(X = x_i)$ a cada posible resultado x_i. El conjunto de probabilidades, que se denomina *función de probabilidad*, debe cumplir que $p(x_i) \geq 0$ y $\sum_i p(x_i) = 1$.

Si X es continua, la probabilidad asociada a cualquier punto en particular es cero, por lo que nos referimos a la probabilidad de que X tome valores en un intervalo $[a, b]$. La *función de densidad* $f(x)$ de una variable aleatoria continua X es una función tal que

$$Probabilidad(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Es decir, el área por debajo de la función entre dos puntos a y b es la probabilidad de que la variable tome valores en el intervalo $[a, b]$ (ver panel izquierdo del Gráfico A.3). La función de densidad toma valores no negativos, $f(x) \geq 0$, y el área total por debajo de la función es la unidad, $\int_{-\infty}^{\infty} f(x)dx = 1$.

Un ejemplo de variable aleatoria continua es la **distribución normal**. Su función de densidad tiene forma de campana (ver panel izquierdo del Gráfico A.3). Es muy utilizada en la práctica para modelar variables que se distribuyen simétricamente alrededor de un valor central, con

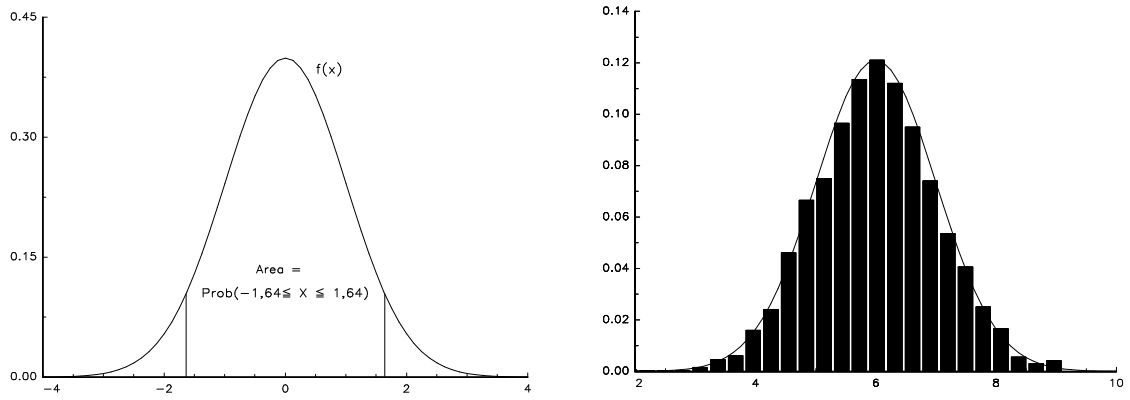


Gráfico A.3: La función de densidad *normal* y el histograma

mucha probabilidad acumulada en valores cercanos a dicho punto central y poca en valores alejados.

El panel derecho del Gráfico A.3 ilustra la relación entre la función de densidad y el histograma de los datos. Tal y como mencionan Peña & Romo (1997): "*La función de densidad constituye una idealización de los histogramas de frecuencia o un* ***modelo*** *del cual suponemos que proceden las observaciones. El histograma representa frecuencias mediante áreas; análogamente, la función de densidad expresa probabilidades por áreas. Además, conserva las propiedades básicas del histograma: es no negativa y el área total que contiene es uno.* "

La distribución de una variable aleatoria puede resumirse utilizando medidas de posición (media, mediana y moda), dispersión (varianza, desviación típica y coeficiente de variación) o forma (coeficiente de asimetría y coeficiente de curtosis). Estos conceptos se definen de forma similar a los utilizados para resumir las características de un conjunto de datos. Definiremos los elementos que utilizaremos a lo largo del curso.

La media o valor esperado, μ, de una variable aleatoria X se define como el promedio ponderado de todos los posibles valores que puede tomar X, donde la ponderación es la probabilidad de cada valor. Si la variable es continua se define:

$$\mu \;=\; E(X) \;=\; \int_{-\infty}^{\infty} x\, f(x) dx$$

donde E se conoce como el operador de esperanzas matemáticas o, simplemente, esperanzas. La media recoge el centro de gravedad sobre el que se distribuye la variable. Así, cuanto mayor sea la media, mayor es el valor que se espera que tomen las realizaciones del experimento (ver panel izquierdo del Gráfico A.4).

La varianza de una variable aleatoria X es su momento central, o respecto a la media, de orden 2. Es decir,

$$var(X) \;=\; \sigma_X^2 \;=\; E[(X-\mu)^2] \geq 0$$

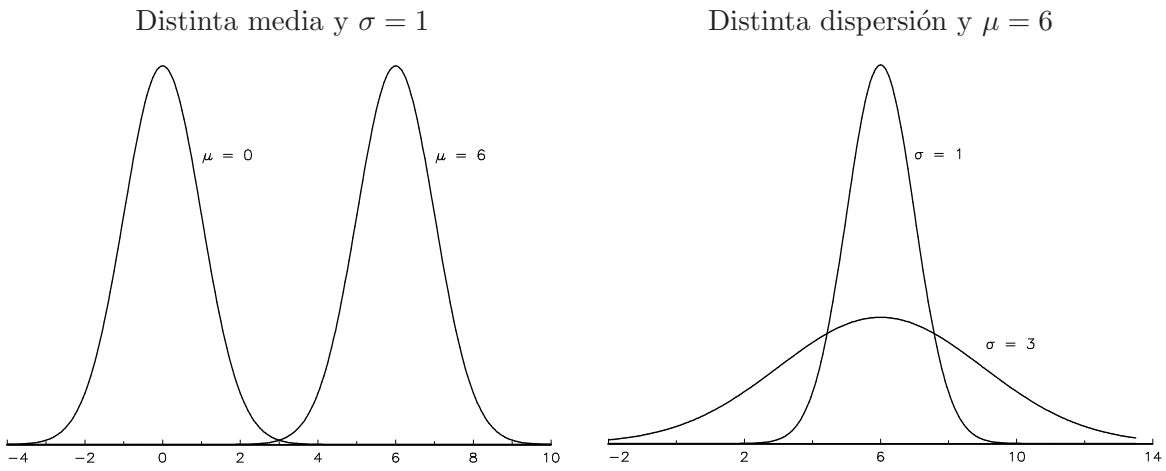


Gráfico A.4: Ejemplos de distribución normal

La varianza es una medida de dispersión de la distribución. Su raíz cuadrada positiva se conoce como **desviación típica** o **desviación estándar** de la variable aleatoria X, es decir:

$$des(X) = \sigma_X = \sqrt{var(X)}$$

El panel derecho del Gráfico A.4 muestra que cuanto menor es la varianza de la variable, mayor es la probabilidad concentrada alrededor de la media.

Distribución normal estándar. La distribución normal se caracteriza por el valor de su media y su varianza. Si Z es una variable aleatoria normal de media igual a 0 y varianza igual a la unidad, se dice que Z es una variable normal estándar y se denota $Z \sim N(0,1)$. Existen tablas de esta distribución que a cada posible resultado z le asigna la probabilidad acumulada hasta ese punto, $Prob(Z \leq z)$.

En general, si X es una variable normal con media μ y varianza σ^2 se denota $X \sim N(\mu, \sigma^2)$. Dado que la transformación $Z = (X - \mu)/\sigma$ es una normal estándar, con la tabla de esta distribución normal se obtiene la probabilidad acumulada $Prob(X \leq x)$.

Ejercicio 1: simulación normal estándar. Crea un conjunto de datos artificiales (N=250 observaciones), generados a partir de variables aleatorias normales estándar independientes. El proceso es el siguiente:

1. En Gretl, crea el conjunto de datos siguiendo los pasos: *Archivo →Nuevo conjunto de datos*, en *Número de observaciones:* escribe 250, elige la estructura de datos *de sección cruzada* y pincha en *No* desea empezar a introducir los valores. Se crea un conjunto de datos con dos variables que genera Gretl automáticamente: la constante *const* y la variable índice *index*, que toma valores 1,2,3,...,250.
2. Crea una serie de 250 realizaciones independientes de una variable normal con:

   Añadir →Variable aleatoria →Normal ...

Aparece un cuadro titulado *gretl: variable normal* donde debes indicar el nombre de la variable, su media y su desviación típica σ. Por ejemplo, para generar observaciones de una variable que llamamos z1 y que se distribuye como una N(0,1), escribimos:

z1 0 1

Tras pinchar en *Aceptar*, en la ventana principal de Gretl aparece la variable creada, $z1$, con la nota explicativa $z1 = normal()$.

3. Repitiendo el paso 2, crea una nueva realización de la normal estándar y llámala $z2$.
4. Haz dos gráficos, uno con $z1$ y otro con $z2$, sobre la variable índice con la opción: *Ver →Gráficos →Gráfico X-Y (scatter)*. Observa sus características comunes: los datos oscilan en torno al valor cero, y la mayor parte de ellos se encuentra en el intervalo (-2, 2).
5. Compara el histograma de las frecuencias relativas con la función de densidad normal. Para ello debes situar el cursor sobre una de las variables y seguir la ruta:

   Variable →Gráfico de frecuencias →contra la normal

   El resultado es un gráfico similar (no idéntico) al Gráfico A.5.

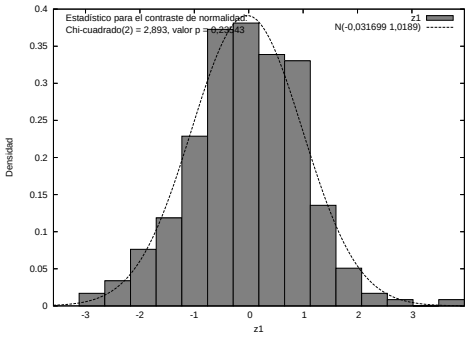


Gráfico A.5: Simulación 1: histograma

En este gráfico aparece el histograma junto con la función de densidad de la distribución normal de media $\mu = 0,1087$ y desviación típica $\sigma = 1,0055$. Estos valores aparecen en la parte superior derecha del gráfico y se eligen en función de la media y varianza de los datos.

Ejercicio 2: simulación normal general. En el mismo fichero crea dos series de datos:

- $x3$= 250 datos generados con una variable normal de media 25 y desviación típica 6 (es decir, $\sigma^2 = 36$). En *Añadir →Variable aleatoria →Normal ...* escribir *x3 25 6*.
- $x4$, generados a partir de una distribución normal de media 50 y desviación típica 0.

Haz el gráfico de los datos sobre la variable *index* y su distribución de frecuencias frente a la normal. ¿Hay algún problema al crear o representar la distribución de

$x4$? ¿Por qué?

Ejercicio 3: transformación lineal. Se trata de construir una nueva serie de datos, que llamaremos $z3$ y que se define a partir de la variable $x3$ del ejercicio anterior:

$$z3 = \frac{x3 - 25}{6}$$

1. Pincha en la opción *Añadir →Definir nueva variable*.
2. En la siguiente ventana escribe el nombre de la nueva serie y su fórmula de cálculo, es decir *z3=(x3-25)/6*.

Si has realizado el proceso correctamente, en la ventana principal de Gretl aparece la variable creada, $z3$. Haz el histograma de $z3$, comparándola con la de la variable inicial $x3$. Compara sus estadísticos descriptivos, en particular, las medias y las varianza. ¿Cambian mucho?

A.1.2. Dos o más variables aleatorias

Para responder a preguntas relativas a dos o más variables aleatorias debemos conocer su función de densidad conjunta. Si las variables aleatorias X e Y son discretas, a cada posible par de resultados (x_i, y_j) podemos asignar una probabilidad $p(x_i, y_j)$. El conjunto de probabilidades es la *función de probabilidad conjunta*, cumpliéndose que $0 \leq p(x_i, y_j) \leq 1$ y $\sum_i \sum_j p(x_i, y_j) = 1$.

Si las variables aleatorias son continuas, su distribución conjunta se recoge mediante la *función de densidad conjunta* $f(x, y)$. Si las dos variables siguen una distribución normal, la forma típica de su función de densidad conjunta se encuentra en el Gráfico A.6.

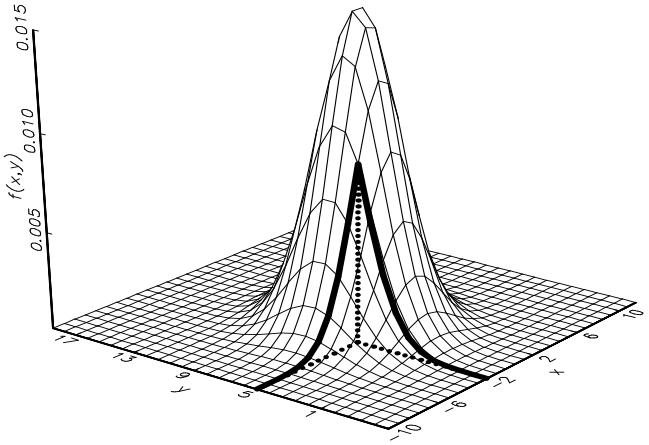


Gráfico A.6: Distribución normal bivariante

El volumen total recogido bajo esta superficie es la masa de probabilidad total que es igual a la unidad, es decir, $\int_x \int_y f(x, y)\, dx\, dy = 1$. Además, la función no toma valores negativos, $f(x, y) \geq 0$. Así, el volumen debajo del rectángulo definido por dos puntos (a, b) mide la probabilidad de que X tome valores por debajo de a e Y por debajo de b. Es decir,

$$Probabilidad(X \leq a,\, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) dx\, dy$$

Por ejemplo, el volumen recogido bajo la superficie marcada en el Gráfico A.6 es la probabilidad de que $X \leq -2$ e $Y \leq 4{,}5$. La **función de densidad marginal** de cada variable puede obtenerse mediante integración. Así:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{A.14}$$

La distribución conjunta de dos variables aleatorias se puede resumir mediante:

- El centro de gravedad de cada variable, es decir, las medias (μ_X, μ_Y), que se obtienen de las distribuciones marginales (A.14).
- Medidas de dispersión de cada variable alrededor de su media, por ejemplo, las varianzas de X e Y, σ_X^2 y σ_Y^2, que se derivan de las distribuciones marginales (A.14).
- Medida de la relación lineal entre las dos variables aleatorias, para lo que se utiliza la covarianza σ_{XY}:

  $$cov(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

  o bien el coeficiente de correlación entre las variables,

  $$corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1,1]$$

  Covarianza y correlación de las variables aleatorias tienen una interpretación similar a sus homólogas en los datos. Así, si $\sigma_{XY} = \rho_{XY} = 0$ se dice que las variables X e Y están incorrelacionadas.

La distribución conjunta se resume en el vector de medias μ y la matriz de varianzas y covarianzas Σ ó V:

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

Distribución condicionada. Al estudiar un conjunto de variables, interesa evaluar la posibilidad de que un suceso ocurra dado que otro suceso ha tenido lugar. Por ejemplo, ¿cuál es la probabilidad de que una mujer casada y con hijos en edad escolar participe en el mercado de trabajo? La **probabilidad condicionada** permite responder este tipo de preguntas. Si las variables son discretas, se define la distribución condicional de Y dado que la variable aleatoria X toma el valor x_i como:

$$Prob(Y = y_j | X = x_i) = \frac{Prob(Y = y_j, X = x_i)}{Prob(X = x_i)} = \frac{p(x_i, y_j)}{\sum_j p(x_i, y_j)} \quad \text{para } Prob(X = x_i) > 0$$

Si las variables son continuas, se define la función de densidad de Y condicionada a que la variable aleatoria X tome el valor x (para $f(x) > 0$):

$$f(y|X = x) = \frac{f(x,y)}{f(x)}$$

De esta forma se obtiene una nueva distribución, con las propiedades ya vistas. Los momentos de interés de esta distribución se denominan media y varianza condicionada de Y para el valor dado de $X = x$, y se denotan $E(Y|X = x)$ y $var(Y|X = x)$.

Independencia. Dos variables aleatorias X y Y son estadísticamente independientes o están independientemente distribuidas si conocido el valor que toma una de ellas, no aporta ninguna información sobre el valor que puede tomar la segunda. Si las variables X e Y son independientes, entonces su función de densidad conjunta puede descomponerse según:

$$f(x,y) = f(x) \times f(y) \qquad -\infty < x, y < \infty$$

Además, se tiene que $f(y|X = x) = f(y)$. Se demuestra que si X e Y son independientes, entonces $Cov(X,Y) = 0$. También se demuestra que, si las variables X e Y se distribuyen conjuntamente según una normal y $Cov(X,Y) = 0$, entonces X e Y son independientes.

Más de dos variables. Los resultados anteriores se pueden generalizar a un conjunto de n variables, $X_1, X_2, \ldots, X_n$, que se recogen en un vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

La distribución conjunta de estas variables se resume en el vector de **medias** $E(\mathbf{X})$ ó $\vec{\mu}$ y la matriz de **varianzas y covarianzas** $V(\mathbf{X})$ ó Σ_X. Así:

$$E(\mathbf{X}) = \vec{\mu} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \text{y}$$

$$\Sigma_X = \begin{pmatrix} var(X_1) & cov(X_1,X_2) & \ldots & cov(X_1,X_n) \\ cov(X_1,X_2) & var(X_2) & \ldots & cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_1,X_n) & cov(X_2,X_n) & \ldots & var(X_n) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \ldots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \ldots & \sigma_n^2 \end{pmatrix}$$

donde Σ_X es una matriz cuadrada de orden n, simétrica y definida no negativa. Esto implica que los elementos de la diagonal principal son no negativos, $\sigma_i^2 \geq 0$, $\forall i$.

Si las variables son mutuamente independientes, entonces están incorrelacionadas, es decir, $\sigma_{i,j} = 0, \forall i \neq j$, por lo que la matriz Σ_X es diagonal:

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}$$

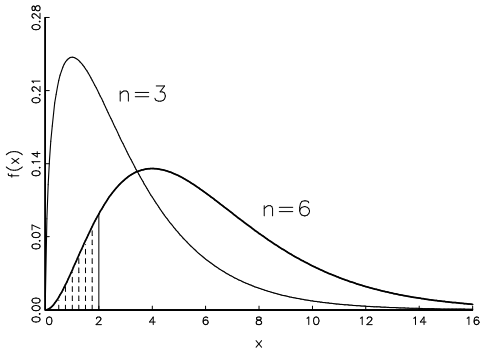


Gráfico A.7: Función de densidad de la distribución Chi-cuadrado

Si, además, $X_1, \ldots, X_n$ siguen la misma distribución, con la misma media y la misma varianza:

$$E(\mathbf{X}) = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \qquad \Sigma_X = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

entonces se dice que son variables aleatorias idéntica e independientemente distribuidas con media μ y varianza σ^2 y se denota $X_i \sim iid(\mu, \sigma^2), \forall i = 1, \ldots, n$.

Si $X_1, \ldots, X_n$ son variables aleatorias normales, se dice que el vector $\mathbf{X}$ sigue una **distribución normal multivariante**, y queda caracterizada por su vector de medias $\vec{\mu}$ y su matriz de varianzas y covarianzas Σ_X. Se denota $\mathbf{X} \sim N(\vec{\mu}, \Sigma_X)$. Si además las variables son independientes, con media y varianza común, se denota $X_i \sim NID(\mu, \sigma^2)$, $i = 1, \ldots, n$.

Además de la distribución normal, a lo largo del curso utilizaremos otras distribuciones, todas ellas relacionadas con la distribución normal. Veamos sus propiedades.

A.1.3. Algunas distribuciones de probabilidad

La distribución Chi-cuadrado. Si $(Z_1, \ldots, Z_n)$ son variables aleatorias independientes con distribución normal estándar, es decir, $Z_i \sim NID(0, 1)$, se dice que $X = \sum_{i=1}^{n} Z_i^2$ es una variable aleatoria chi-cuadrado de n grados de libertad y se denota $X \sim \chi^2(n)$. Para valores negativos de X, $f(x) = 0$ y la forma general de su función de densidad se recoge en el Gráfico A.7.

Es una distribución asimétrica, con media igual a n y varianza $2n$. Existen tablas que proporcionan la probabilidad acumulada hasta un punto $Prob(X \leq x)$, es decir, el área rayada del gráfico, en función de los grados de libertad, n.

> **Ejercicio 4: transformación no lineal.** Siguiendo el procedimiento del ejercicio 3, crea una nueva serie de datos, $y = z1^2 + z2^2 + z3^2$. En este caso debes escribir:
>
> $$y = z1\hat{}2 + z2\hat{}2 + z3\hat{}2$$

Haz la representación gráfica de la distribución de frecuencias de esta variable frente a la normal. El histograma que obtengas tendrá un patrón bastante diferente a la distribución normal. ¿Puedes justificar el resultado? ¿Con qué distribución la compararías?

La distribución F de Snedecor. Si $Z_1 \sim \chi^2(n_1)$ y $Z_2 \sim \chi^2(n_2)$ y además se distribuyen independientemente, entonces la distribución $X = (n_2/n_1)(Z_1/Z_2)$ se conoce como distribución F de n_1, n_2 grados de libertad y se escribe:

$$X = \frac{Z_1/n_1}{Z_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

El Gráfico A.8 muestra su función de densidad para distintos grados de libertad.

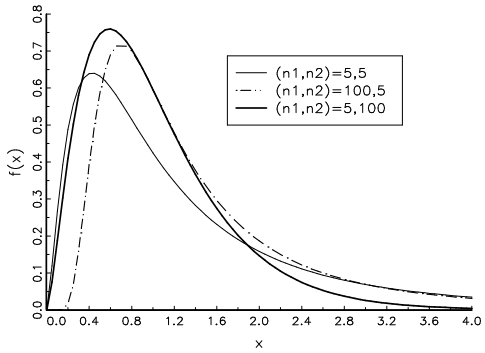


Gráfico A.8: Función de densidad de la distribución F-Snedecor

La probabilidad se acumula en la parte positiva de la recta real, $x > 0$. A medida que aumentan los grados de libertad del denominador, $n_2 \to \infty$, la distribución de $n_1\mathcal{F}(n_1, n_2)$ converge a la distribución $\chi^2(n_1)$.

La distribución t de Student. Si $Z \sim N(0, 1)$ e $Y \sim \chi^2(n)$ y además, Z e Y se distribuyen independientemente, entonces la distribución de $X = Z/\sqrt{Y/n}$ se denomina distribución t de Student de n grados de libertad y se denota:

$$X = \frac{Z}{\sqrt{Y/n}} \sim t(n)$$

El Gráfico A.9 incluye ejemplos de la función de densidad de la t-Student comparándolas con la distribución normal estándar:

Se trata de una distribución simétrica alrededor de 0. Para $n > 1$, la media de la distribución es cero y para $n > 2$ su varianza es igual a $n/(n-2)$. Esta distribución tiene las colas más gruesas que la normal, es decir, su exceso de curtosis es positivo, pero, a medida que aumentan sus grados de libertad, la distribución t converge a la normal estándar.

A.2. Repaso de inferencia estadística

Supongamos que interesa conocer cuál es el salario medio de los recién licenciados. Se trata de una población o conjunto de individuos muy amplio, por lo que se recoge la información

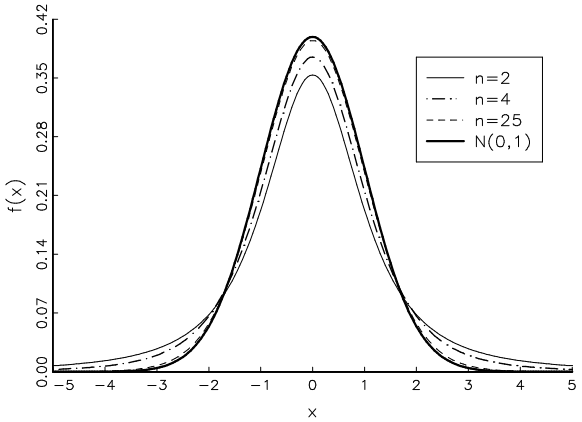


Gráfico A.9: Función de densidad de la distribución t-Student

únicamente de una muestra o un subconjunto de recién licenciados seleccionados al azar. Con esta información, ¿qué es posible inferir del salario esperado de un recién licenciado? Para responder a esta pregunta y, en general, saber usar los datos para examinar conjeturas y relaciones sobre la población repasaremos algunos conceptos de inferencia estadística.

El objetivo de la inferencia estadística es aprender determinadas características de una población a partir del análisis de una muestra. La **población** es un conjunto bien definido de elementos que son el objeto del estudio, por ejemplo, el conjunto de familias de un país, el conjunto de viviendas de una ciudad o los clientes de una empresa de telecomunicaciones. La **muestra** está formada por un subconjunto representativo de elementos de la población.

Una vez definida la población, hay que especificar un modelo para los datos que recoja las características poblacionales que interesan. En Econometría suponemos que los datos $y_1, y_2, \ldots, y_N$ son realizaciones de N variables aleatorias cuya distribución conjunta depende de varios parámetros desconocidos Θ. Un **modelo** para los datos especifica las características generales de la distribución junto con el vector de parámetros desconocidos Θ. Por ejemplo, supongamos que nos interesa conocer el precio *medio* del metro cuadrado de un piso en una ciudad y la muestra está formada por 50 pisos. Suponemos que los valores recogidos del precio por m^2 de los 50 pisos, $y_1, \ldots, y_{50}$, son realizaciones de variables normales idéntica e independientemente distribuidas. Por tanto, el modelo especificado para los datos es:

$$Y_i \sim NID(\mu, \sigma^2)$$

Los parámetros que determinan la distribución son la media y la varianza del precio del m^2, que son desconocidos, es decir, $\Theta = (\mu,\, \sigma^2)$. Además, la media es el parámetro de interés en el estudio y queremos *aprender* sobre ella a partir de los datos.

En grandes líneas, aplicaremos dos herramientas de la estadística, la estimación y el contraste de hipótesis. En la estimación se trata de calcular posibles valores para parámetros de interés, por ejemplo, una elasticidad o el precio medio por metro cuadrado de la vivienda. En el contraste de hipótesis hay que establecer una hipótesis o conjetura específica sobre la población, por ejemplo, que no hay discriminación salarial por sexo o que el estado de un piso es un factor determinante de su precio, y analizar los datos para decidir si la hipótesis es correcta.

A.2.1. Estimación

El objetivo de la estimación es aproximar el valor de un conjunto de parámetros desconocidos de una distribución a partir de las observaciones muestrales de la misma. Denotaremos como θ a un parámetro desconocido y $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)'$ a un vector de K parámetros desconocidos. Un **estadístico** es una función de los datos, $g(y_1, \ldots, y_N)$. Un **estimador puntual** de θ es un estadístico que pretende ser un aproximación al parámetro desconocido y se denota por $\hat{\theta}$. Por ejemplo, la media de los datos puede ser un estimador de la media de una variable aleatoria y la varianza de los datos un estimador de su varianza. Es decir,

$$\hat{\mu} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad\qquad \hat{\sigma}^2 = S_y^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Un estimador es una regla que está definida antes de que los datos se observen. El valor numérico que se obtiene al aplicarlo a los datos se denomina *estimación*. Por ejemplo, la estimación de la media del precio por metro cuadrado de un piso con la muestra de la Tabla 1.1 es:

$$\hat{\mu} = \frac{3,82 + 5,246 + \ldots + 3,434 + 4,20}{50} = 3,91 \text{ miles de euros}$$

Es decir, se estima que el precio de un piso oscila alrededor de 3910 euros/m^2. Sin embargo, ¿qué confianza podemos tener en este resultado? Por ejemplo, ¿valoraríamos igual esta cantidad si se hubiera calculado con una muestra de 5 observaciones? La respuesta obvia es NO, sino que consideramos más fiables los resultados con 50 datos que con 5. Por tanto, un estimador (y sus estimaciones) deben complementarse con una medida de su fiabilidad o precisión.

Un estimador es una variable aleatoria que depende de las variables Y_i, $i = 1, \ldots, N$. Su distribución de probabilidad se denomina distribución muestral o distribución empírica del estimador. En el ejemplo anterior, si $Y_i \sim NID(\mu, \sigma^2)$, entonces el estimador $\hat{\mu} = \bar{y}$ es una combinación lineal de N variables normales independientes, por lo que su distribución muestral es:

$$\hat{\mu} = \bar{y} \sim N(\mu, \sigma^2/N) \tag{A.15}$$

La media muestral se distribuye alrededor de la media poblacional y se concentra más probabilidad alrededor de μ cuanto mayor es N (es decir, menor es la varianza). Por tanto, hay mayor probabilidad de obtener una estimación cercana a μ con 50 datos que con $N = 5$. En este caso, es sensato utilizar como indicador de la *precisión* la desviación típica $\sigma/\sqrt{N}$: menor desviación típica indica mayor precisión. Normalmente, σ es desconocido, por lo que sustituimos su valor poblacional por el correspondiente muestral, S_y^*. La estimación de la desviación típica de la distribución muestral de $\bar{y}$,

$$\hat{\sigma}_{\bar{y}} = S_{\bar{y}} = S_y^*/\sqrt{N}$$

se conoce como *error típico* de $\bar{y}$. En el ejemplo del precio del m^2, obtenemos que el error típico de estimación es $0,993341/\sqrt{50} = 0,14$. Es fácil comprobar que si obtuviéramos los mismos valores de $\bar{y}$ y S_y con una muestra de 5 observaciones, el error típico se triplicaría, $S_{\bar{y}} = 0,993341/\sqrt{5} = 0,44$ miles de euros.

Ejercicio 5. Estimación de la media y la varianza del precio por m^2 de un piso.

1. Abre el fichero de datos de Gretl pisos.gdt.
2. Crea la variable precio por metro cuadrado, que denotaremos *pr_ m2*:
   a) Usa las opción *definir nueva variable* que está en el menú *Añadir* o en *Variable*.
   b) En la nueva ventana escribe *nombre de la nueva variable = fórmula*, es decir,
   $$pr\_\ m2 = precio/m2$$
3. Una vez creados los nuevos datos, las estimaciones de la media, m, y la desviación típica, S, se obtienen de la tabla de estadísticos descriptivos. La estimación de la varianza es el cuadrado de S. El error típico de estimación es $S/\sqrt{50}$.

Ejercicio 6: Estimación de media y varianza. Utilizando la opción de estadísticos descriptivos o estadísticos principales, obtén las medias y las desviaciones típicas de $z1$, $z2$, $x3$ y $x4$ generados en el ejercicio 1. Completa la siguiente tabla, incluyendo junto con los momentos poblacionales las estimaciones que has obtenido, es decir, correspondientes los momentos muestrales.

Modelo 1	$\mu =$	$\sigma =$
Muestra: $z1$	Estimación =	Estimación =
Modelo 2	$\mu =$	$\sigma =$
Muestra: $z2$	Estimación =	Estimación =
Modelo 3	$\mu =$	$\sigma =$
Muestra: $x3$	Estimación =	Estimación =
Modelo 4	$\mu =$	$\sigma =$
Muestra: $x4$	Estimación =	Estimación =

Criterios para comparar estimadores

Para un problema determinado existen distintos métodos de estimación y, obviamente, unos son mejores que otros. En algunos casos, distintos métodos pueden dar lugar a un mismo estimador de un parámetro. Es posible elegir entre distintos métodos de estimación basándonos en ciertas propiedades de la distribución muestral del estimador. En general, buscamos los estimadores que más se aproximen a los verdaderos valores. Así, exigimos que los estimadores cumplan una serie de propiedades basadas en una medida de la distancia entre θ y $\hat{\theta}$. En este curso nos fijamos en tres propiedades: insesgadez, eficiencia y el error cuadrático medio mínimo.

Insesgadez. Un estimador es insesgado si la media de su distribución empírica es el verdadero valor del parámetro, es decir,

$$E(\hat{\theta}) = \theta$$

Si se pudieran obtener todas las posibles realizaciones muestrales de $\hat{\theta}$, el promedio de todas estas estimaciones sería el valor del parámetro. Es una propiedad deseable porque indica que si un estimador es insesgado, el error de estimación, $\hat{\theta} - \theta$, se anula en promedio. Un ejemplo de estimador insesgado de la media poblacional de una distribución normal es $\bar{y}$, ya que de (A.15) tenemos que $E(\bar{y}) = \mu$. Un estimador insesgado de la varianza de una distribución es la varianza muestral, S^2.

En caso contrario, se dice que el estimador es sesgado. Se define el sesgo de un estimador como $Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$. La parte izquierda del Gráfico A.10 representa las distribuciones de 3 estimadores de un mismo parámetro, θ: el estimador $\hat{\theta}_1$ es insesgado; $\hat{\theta}_2$, tiene sesgo negativo, es decir, en promedio subestima el valor del parámetro; finalmente el sesgo de $\hat{\theta}_3$ es positivo, es decir, este estimador en promedio sobrevalora el valor del parámetro.

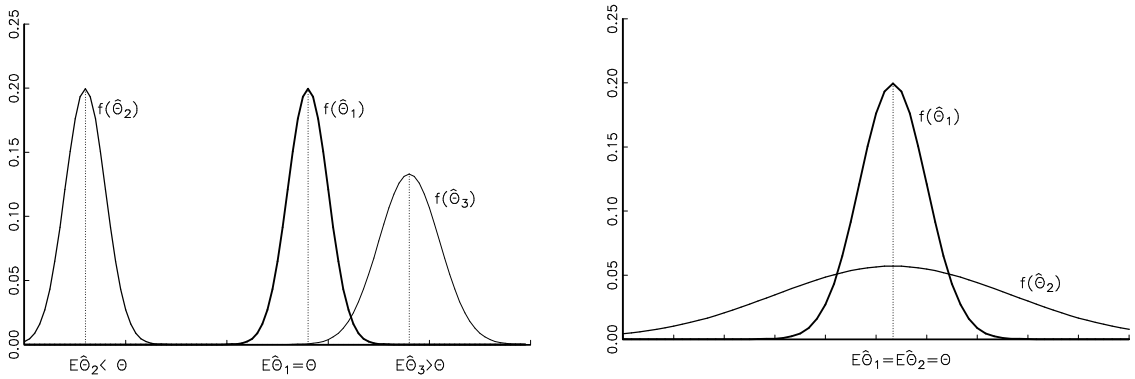


Gráfico A.10: Sesgo y varianza de estimadores

Eficiencia. Si nos fijamos únicamente en los estimadores insesgados, nos interesa establecer un criterio para elegir un estimador dentro de esta clase de estimadores. En la parte derecha del Gráfico A.10 se representa la distribución de dos estimadores, ambos insesgados. Claramente, el estimador con menor varianza, $\hat{\theta}_1$, tiene una probabilidad menor de obtener realizaciones *alejadas* del verdadero valor del parámetro. Por tanto, se considera que $\hat{\theta}_1$ supera al estimador $\hat{\theta}_2$ y se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$.

En general, si un estimador es el que tiene menor varianza dentro de una clase de estimadores se dice que es el estimador *eficiente dentro de esa clase*. Así, se dice que un estimador $\hat{\theta}$ es eficiente dentro de la clase de estimadores insesgados si no hay otro estimador insesgado $\tilde{\theta}$ con una varianza menor:

$$var(\tilde{\theta}) \geq var(\hat{\theta}) \qquad \forall \tilde{\theta} \quad insesgado$$

Por ejemplo, la media de los datos es un estimador eficiente dentro de la clase de estimadores insesgados de la media poblacional μ de una variable normal. Es decir, se demuestra que, si $Y_i \sim NID(\mu, \sigma^2)$, $i = 1, \ldots, N$, entonces para todo estimador insesgado de μ, $\tilde{\mu}$ con $E\tilde{\mu} = \mu$:

$$var(\bar{y}) = \frac{\sigma^2}{N} \leq var(\tilde{\mu})$$

Si se trata de estimar un conjunto de K parámetros Θ, se dice que un estimador insesgado $\widehat{\Theta}$ es más eficiente que otro estimador insesgado $\tilde{\Theta}$ si la diferencia $[V(\tilde{\Theta}) - V(\widehat{\Theta})]$ es una matriz semidefinida positiva. Esto implica que cada elemento de $\widehat{\Theta}$ tiene una varianza menor o igual que el correspondiente elemento de $\tilde{\Theta}$.

Error cuadrático medio Aunque la insesgadez es una propiedad deseable, esto no implica que un estimador insesgado siempre sea preferible a uno sesgado. El Gráfico A.11 ilustra una situación en la que un estimador insesgado $\hat{\theta}_1$ puede descartarse frente a otro sesgado, $\hat{\theta}_2$. El estimador $\hat{\theta}_1$ tiene mucha varianza, por lo que tiene una probabilidad mayor de obtener errores de estimación más grandes que el estimador con menor varianza, $\hat{\theta}_2$, aunque este sea sesgado.

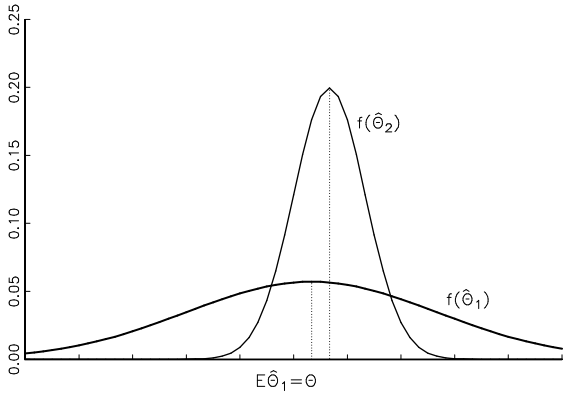


Gráfico A.11: Ejemplos de distribución de estimadores

Esto sugiere utilizar como criterio de elección de estimadores una medida del error del estimador. Se define el error cuadrático medio de un estimador:

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + [sesgo(\hat{\theta})]^2$$

que se descompone en un término de varianza y otro de sesgo. Así, entre un conjunto de estimadores se elige aquel que tiene menor error cuadrático medio.

A.2.2. Contraste de hipótesis

Como ya se mencionó, uno de los objetivos de la Econometría es el de *contrastar hipótesis*. Por ejemplo, nos planteamos si los datos del precio del m^2 de la vivienda son compatibles con una determinada distribución con media 3000 euros/m^2. En un contraste de hipótesis se trata de establecer si la diferencia entre la hipotética media poblacional (en el ejemplo, 3000€) y la media muestral (3910€) se debe únicamente a la naturaleza aleatoria de los datos.

Un contraste de hipótesis tiene tres etapas (Ramanathan, 2002): (1) Formulación de dos hipótesis opuestas; (2) derivación de un estadístico de contraste y su distribución muestral; y (3) 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. La hipótesis de interés se denomina **hipótesis nula**, H_0, mientras que la hipótesis frente a la que se contrasta se llama **hipótesis alternativa**, H_a. En el

ejemplo, consideramos que el precio del m^2 es una variable aleatoria normal y planteamos la hipótesis nula de que la media de Y sea igual a 3 (miles €) frente a la alternativa de que no lo sea, es decir,

$$H_0\colon \mu = 3 \qquad \text{frente a} \qquad H_a\colon \mu \neq 3$$

Normalmente, la hipótesis nula es una hipótesis simple, es decir, sólo se plantea un valor para μ. La hipótesis alternativa suele ser una hipótesis compuesta, que especifica un intervalo de valores. En el ejemplo, H_a es la negación de H_0 y se dice que es un contraste bilateral o *a dos colas*. Si la hipótesis alternativa se especifica H_a: $\mu < 3$, o bien H_a: $\mu > 3$, se dice que el contraste es unilateral o *a una cola*.

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. Por ejemplo, en el contraste bilateral sobre la media, se define la siguiente medida de la discrepancia:

$$\frac{\bar{y} - 3}{S_{\bar{y}}}$$

Esta discrepancia, que utilizaremos como estadístico de contraste, no depende de las unidades de medida y tiene en cuenta la diferencia entre los datos (resumidos en $\bar{y}$) y el valor establecido en H_0. Además, debe conocerse la distribución de esta variable aleatoria cuando la hipótesis nula es correcta. En el ejemplo, se demuestra que si los datos $y_1, y_2, \ldots, y_N$ son una muestra aleatoria de un conjunto de variables $Y_i \sim NID(\mu, \sigma^2)\, \forall i$, con μ y σ^2 desconocidas, entonces:

$$\frac{\bar{y} - \mu}{S_{\bar{y}}} \sim t(N-1)$$

y sustituyendo $\mu = 3$, tenemos la distribución muestral del estadístico bajo H_0:

$$t = \frac{\bar{y} - 3}{S_{\bar{y}}} \overset{H_0}{\sim} t(N-1) \tag{A.16}$$

Este estadístico se aplica mucho en la práctica y se denomina estadístico t de la media.

Finalmente, para determinar **el criterio de decisión** del contraste se divide el conjunto de posibles resultados del estadístico de contraste en dos zonas, la **región crítica** y su complementaria. Se rechaza H_0 cuando el valor del estadístico obtenido con la muestra t^m pertenece a la región crítica. El punto de partida para establecer la región crítica es que se rechaza H_0 si la discrepancia entre datos y H_0 es *grande*. En el contraste bilateral, se rechazaría H_0 si $\bar{y}$ se alejara *mucho* del valor establecido en H_0, lo que para el estadístico implica que:

$$|t^m| = \left| \frac{\bar{y} - 3}{S_{\bar{y}}} \right| > c \tag{A.17}$$

donde c es la discrepancia máxima que estamos dispuestos a asumir y se denomina *valor crítico*. En caso contrario, si $|t^m| \leq c$, no se rechaza la hipótesis nula. El valor de c depende de la distribución del estadístico de contraste cuando H_0 es cierta y del error que estemos dispuestos a aceptar. En un contraste siempre existe la posibilidad de cometer los siguientes errores:

- Rechazar la hipótesis nula cuando ésta es cierta, que se llama error tipo I. El *nivel de significación* o *tamaño* de un contraste es la probabilidad de incurrir en el error tipo I y se denota por α.

- No rechazar la hipótesis nula cuando ésta es falsa, llamado error tipo II. La *potencia* de un contraste es la probabilidad de no cometer un error tipo II.

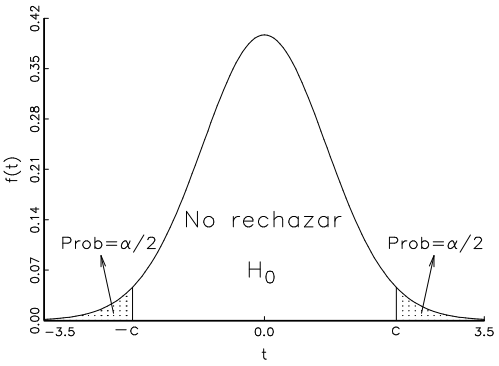
Deseamos cometer el menor error, pero no es posible eliminar los dos errores simultáneamente, es decir, que el tamaño sea 0 y la potencia igual a 1. En general, disminuir el error tipo I lleva consigo un aumento del error tipo II. Por ejemplo, no cometemos error tipo I si decidimos no rechazar nunca la hipótesis nula; pero la potencia del contraste sería 0 porque tampoco rechazaremos H_0 cuando sea falsa. Daremos más importancia al error tipo I, por lo que elegiremos el tamaño del contraste; los niveles más habituales son 10 %, 5 % y 1 %. Para el tamaño elegido, trataremos de utilizar el contraste con mayor potencia.

Ejemplo: zona crítica en un contraste bilateral sobre la media de una distribución normal. Veamos cómo se determina el valor crítico c en el ejemplo sobre la media del precio. El tamaño α es la probabilidad de rechazar H_0 cuando ésta es cierta. Como (A.17) es la condición para rechazar y (A.16) es la distribución del estadístico cuando H_0 es cierta, esto implica que:

$$\alpha = Prob(|t| > c) \qquad \text{cuando el estadístico } t \sim t(N-1)$$

En este caso, rechazaremos H_0 si el valor del estadístico t obtenido con los datos es un valor *poco probable* en la distribución del estadístico bajo H_0.

Este gráfico muestra la distribución del estadístico si H_0: $\mu = 3$ es cierta. La región crítica es la zona punteada en las **dos** colas de la distribución, de modo que en cada cola se acumula una probabilidad $\alpha/2$. Así, c es la ordenada de la distribución $t(N-1)$ que deja en la cola derecha una probabilidad $\alpha/2$. Por ejemplo, para $\alpha = 0,05$ y $N = 50$, entonces, $c = 2,01$ y se rechaza H_0 al nivel de significación del 5 % si $|t^m| > 2,01$.



Ejemplo 1: Contraste sobre la media del precio por m^2 en Gretl.
Suponiendo que la variable precio por metro cuadrado pr_m2 sigue una distribución normal, contrasta H_0: $\mu = 3$ frente a H_a: $\mu \neq 3$. Los pasos son los siguientes:

1. Cálculo del valor muestral del estadístico $t = (\bar{y} - 3)/S_{\bar{y}}$, siendo $\bar{y}$ la media muestral de pr_m2:

$$t^m = \sqrt{50}(3,9144 - 3)/0,99341 = 6,51$$

Se obtiene con la siguiente opción de Gretl:

Herramientas →Calculadora de estadísticos de contraste

En la siguiente ventana elige la pestaña *media* y en ella:

- Marca la opción *Utilice una variable del conjunto de datos.*
- Selecciona la variable pr_m2. Aparecerán los estadísticos descriptivos que intervienen en el cálculo de t^m. En este caso:

media muestral:	3,9144
desv. típica:	0,99341
tamaño muestral:	50

- Escribe la hipótesis nula a contrastar: *H0: media =* 3.
- Comprueba que la opción *Suponer que la desv. típica es un valor poblacional* no está activada y pincha en *Aplicar*.

El resultado es la tabla y el Gráfico A.12. En el gráfico se representa la distribución del estadístico bajo H_0, en este caso $t(49)$, junto con el valor muestral del estadístico (la línea verde).

```
Hipótesis nula: media poblacional = 3    Tamaño muestral: n = 50
Media muestral = 3,91439, desv. típica = 0,993407
Estadístico de contraste: t(49) = (3,91439 - 3)/0,140489 = 6,50864
valor p a dos colas = 3,83e-008 (a una cola = 1,915e-008)
```

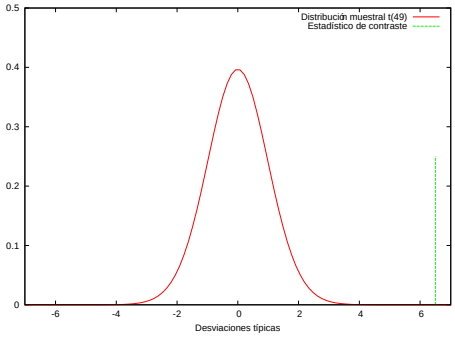


Gráfico A.12: Ejemplo 1: Resultado y distribución del estadístico bajo H_0

En este caso tenemos que el valor muestral del estadístico cae en la cola superior, en una intervalo de valores poco probable si H_0 es cierta. Por tanto, rechazaremos la hipótesis nula. Pero calcularemos exactamente la región crítica.

2. Región crítica o zona de rechazo. El valor crítico c se obtiene con la opción de Gretl *Herramientas* →*Tablas estadísticas*.

   En la nueva ventana hay que elegir la pestaña de la variable t y en el siguiente cuadro hay que rellenar:
   - *gl*= grados de libertad n, en este caso 49
   - probabilidad en la cola derecha = $\alpha/2$. Fijamos un nivel de significación del 5 %, por lo que escribimos 0,025.

   Tras pinchar en *Aceptar*, obtenemos el siguiente resultado:

```
t(49)     probabilidad en la cola derecha = 0,025
          probabilidad complementaria = 0,975
          probabilidad a dos colas = 0,05

          Valor crítico = 2,00958
```

   Interpretación: $Prob(t > 2,00958) = 0,025$ o bien $Prob(X < 2,00958) = 0,975$. Por tanto, el valor crítico con $alpha = 5\,\%$ es igual a $c = 2,00958$.

3. Aplicación de la regla de decisión. Como $|6,51| > c$, al nivel de significación del 5 %, se rechaza la hipótesis de que el precio medio sea igual a 3000€ frente a la alternativa. Cierra las ventanas de *calculadora de estadísticos* y *tablas estadísticas*.

Ejemplo: región crítica en el contraste unilateral sobre la media de una distribución normal. En los estudios econométricos a veces se plantean contrastes a una cola. Por ejemplo, en estudios sociales interesa analizar si hay discriminación salarial, de modo que las mujeres perciben salarios más bajos que los hombres. Habitualmente, se contrasta la hipótesis nula de que la media del salario que perciben las mujeres es igual al salario medio de los hombres frente a la hipótesis alternativa de que la media del salario es mayor en el grupo de hombres.

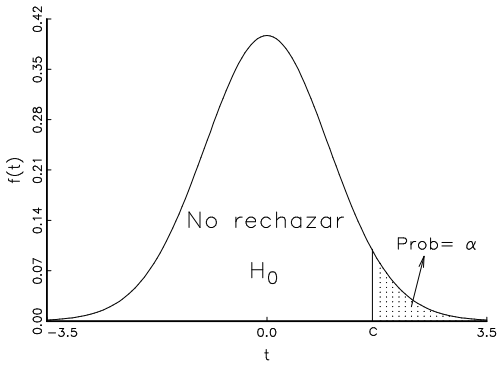
En el estudio del precio del m^2, supongamos que interesa contrastar si la media es tres o mayor, por lo que planteamos las hipótesis:

$$H_0 : \mu = 3 \qquad \text{frente a} \qquad H_a : \mu > 3$$

Al mantenerse la misma hipótesis nula, el estadístico de contraste es (A.16), $t = \sqrt{N}(\bar{y}-3)/S_y$, que bajo H_0 sigue una distribución $t(N-1)$. La hipótesis alternativa determina el criterio de decisión. Rechazaremos H_0 cuando la discrepancia tome valores alejados de H_0 y compatibles con H_a, es decir, cuando t tome valores positivos *grandes*. La región crítica está definida por la condición $t > c$. El valor crítico c se determina por:

$$\alpha = Prob(t > c) \qquad \text{cuando el estadístico } t \sim t(N-1)$$

La región crítica del contraste es la zona punteada en **una** cola de la distribución, la derecha. Así, c es la ordenada de la distribución $t(N-1)$ que acumula en la cola derecha una probabilidad α.
Por ejemplo, si $\alpha = 0,05$ y $N = 50$, entonces el nivel crítico es $c = 1,67655$ (usar herramienta de tabla estadística de Gretl) y no se rechaza H_0 al nivel de significación del 5 % si $t^m < 1,67655$.



En general, se usan las expresiones *rechazar* o *no rechazar* H_0. Esto es así porque en un contraste mantenemos la H_0 mientras no haya suficiente evidencia en contra. Los datos pueden rechazar la hipótesis, pero no pueden probar que H_0 sea correcta, por lo que no se dice que *se acepta* H_0. No rechazar H_0 significa que los datos no son capaces de mostrar su falsedad.

Ejemplo 2: Contraste de igualdad de varianzas. Los datos que estamos analizando sobre precio de la vivienda incluye dos tipos de viviendas:

- Viviendas a reformar, es decir, es necesario realizar un gasto adicional para acondicionar la vivienda.
- Viviendas acondicionadas para entrar a vivir.

Es posible que el precio medio de las viviendas a reformar y reformadas sigan

patrones diferentes. Esto implica que la distribución del precio de los dos tipos de vivienda es distinta. Por tanto, consideramos el siguiente modelo:

- El precio por metro cuadrado de la vivienda que no necesita reforma, Y_1 sigue una distribución normal de media μ_1 y varianza σ_1^2.
- El precio por metro cuadrado de la vivienda a reformar, Y_2 sigue una distribución normal de media μ_2 y varianza σ_2^2.
- Ambas variables Y_1 e Y_2 son independientes.

Vamos a contrastar si la varianza es la misma en ambas distribuciones frente a que sea menor en el grupo de pisos a reformar. Por tanto, planteamos el contraste de hipótesis:

$$H_0 \colon \sigma_1^2 = \sigma_2^2 \qquad \text{frente a} \qquad H_a \colon \sigma_1^2 > \sigma_2^2$$

El procedimiento de contraste consiste en comparar las dos varianzas muestrales, S_1^{*2} y S_2^{*2}, que son estimadores insesgados de las respectivas varianzas poblacionales. Valores cercanos de S_1^{*2} y S_2^{*2}, o ratios $S_1^{*2}/S_2^{*2} \simeq 1$, apoyan H_0. El estadístico de contraste y su distribución bajo H_0 son:

$$F = \frac{S_1^{*2}}{S_2^{*2}} \overset{H_0}{\sim} \mathcal{F}(N_1 - 1, N_2 - 1)$$

donde N_1 es el número de pisos que no necesita reforma y N_2 el número de pisos a reformar. Dada H_a, rechazamos H_0 si el ratio S_1^{*2}/S_2^{*2} está muy por encima de 1. La región crítica, por tanto, está definida por $S_1^{*2}/S_2^{*2} > c$, siendo c el valor crítico. Los pasos para realizar el contraste con Gretl son:

1. Seleccionar el subconjunto de pisos que no necesitan reforma. En el fichero de datos *pisos.gdt* son las observaciones para las que la variable *Reforma = 1*. En Gretl, seleccionamos la submuestra que cumple esta condición si:
   a) Vamos a *Muestra →Definir a partir de v. ficticia*.
   b) En la nueva ventana aparece como opción *Reforma* y pinchamos en *Aceptar*

   Si el proceso es correcto, en la parte inferior de la pantalla de *Gretl* aparece el mensaje *Sin fecha: rango completo n=50; muestra actual n=31*. Ahora sólo trabajamos con los datos de pisos que no necesitan reforma: si consultamos los datos en *Datos →Mostrar valores* ahora sólo aparece la información de los 31 pisos que pertenecen a esta clase.
2. Crear la serie de datos $y1$ que incluye únicamente los precios por m^2 de los pisos reformados: en *Añadir →Definir nueva variable...* escribimos *y1 = pr_m2*.
3. Seleccionar el subconjunto formado por los pisos que necesitan reforma, es decir, caracterizados por *Reforma = 0*:
   a) Vamos a *Muestra →Restringir, a partir de criterio*.
   b) En la nueva ventana escribimos el criterio de selección: *Reforma =0*
   c) Pinchamos en *Reemplazar restricción actual* y luego en *Aceptar*.

   Ahora debe aparecer *Sin fecha: rango completo n=50; muestra actual n=19*.
4. Crear la serie de datos $y2$ de precios por m^2 de pisos no reformados: en *Añadir →Definir nueva variable...* escribimos *y2 = pr_m2*.

5. Recuperar la muestra completa en *Muestra →Recuperar rango el completo*. Comprobamos que las series $y1$ e $y2$ no tienen errores editando los datos de estas series. Las celdas de $y1$ estarán vacías en pisos no reformados y lo recíproco para $y2$.

6. Calcular el valor muestral del estadístico F^m en *Herramientas →Calculadora de estadísticos de contraste →2 varianzas*. En la siguiente ventana rellenamos los datos:
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y1$. Aparecen los estadísticos necesarios de $y1$: $S_1^{*2} = 0,77702$ y $N_1 = 31$
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y2$. Aparecen los estadísticos necesarios de $y2$: $S_2^{*2} = 0,70340$ y $N_2 = 19$
   - Comprobar la marca en *Mostrar el gráfico de la distribución muestral* y *Aplicar*.

   El resultado es una tabla y un gráfico con la distribución del estadístico bajo H_0, $\mathcal{F}(30, 18)$ y el valor muestral del estadístico.

```
Hipótesis nula: Las varianzas poblacionales son iguales
Muestra 1: n = 31, varianza = 0,777054
Muestra 2:  n = 19, varianza = 0,703402
Estadístico de contraste: F(30, 18) = 1,10471
valor p a dos colas = 0,8436 (a una cola = 0,4218)
```

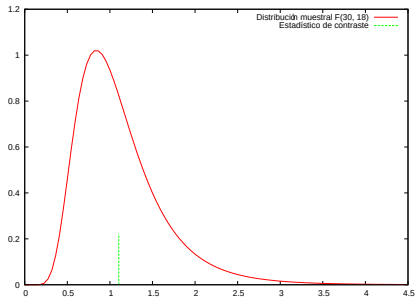


Gráfico A.13: Ejemplo 2: Resultado y distribución del estadístico bajo H_0

7. El gráfico anterior sugiere que no rechazaremos H_0. Calculamos la región crítica: se trata de un contraste a una cola, por tanto, buscamos c tal que $0,05 = Prob(F > c)$. Vamos a *Herramientas →Tablas estadísticas →F*.

   Los grados de libertad del numerador son *gln 30* y los del denominador, *gld 18*. Finalmente, la *probabilidad en la cola derecha* es 0,05. El resultado es:

```
F(30, 18)     probabilidad en la cola derecha = 0.05
              probabilidad complementaria = 0.95
              Valor crítico = 2.10714
```

   Por tanto, si $\alpha = 5\,\%$, entonces $c = 2,107$.

8. Conclusión del contraste: $F^m = 1,10 < 2,11$, por tanto, al nivel de significación del $5\,\%$ no rechazamos la hipótesis de igualdad de varianzas entre los dos tipos de viviendas.

Ejemplo 3: Contraste de igualdad de medias. Vamos a contrastar la hipótesis de que el precio medio del piso es mayor en los pisos reformados. Suponiendo que el precio por m^2 de los dos tipos de pisos son variables independientes, ambas con distribución normal de igual varianza, σ^2 y medias diferentes, μ_1 y μ_2.

Para contrastar la hipótesis anterior, planteamos H_0: $\mu_1 = \mu_2$ frente a H_a: $\mu_1 > \mu_2$.

El procedimiento de contraste se basa en la comparación de las dos medias muestrales, $\bar{y}_1$ y $\bar{y}_2$. Pequeñas diferencias entre ellas apoyan la H_0. El estadístico de contraste y su distribución bajo H_0 son:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S\sqrt{1/N_1 + 1/N_2}} \overset{H_0}{\sim} t(N_1 + N_2 - 2)$$

donde S^2 es el estimador de la varianza común utilizando todos los datos:

$$S = \frac{1}{N_1 + N_2 - 2}\left(\sum_{i=1}^{N_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{N_2}(y_{2i} - \bar{y}_2)^2\right)$$

Dada H_a, rechazamos H_0 si la diferencia $\bar{y}_1 - \bar{y}_2$ es *grande*. La región crítica, por tanto, está definida por $t > c$, siendo c el valor crítico.

Aplicamos el procedimiento de contraste a los datos en Gretl. Las dos series de datos $y1$ e $y2$ se crean según lo descrito en el ejemplo 2. A continuación debemos:

1. Calcular el valor muestral del estadístico t^m en *Herramientas →Calculadora de estadísticos de contraste →2 medias*. En la siguiente ventana rellenamos los datos:
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y1$. Aparecen los estadísticos de $y1$: $\bar{y}_1 = 4,3040$, $S_1^* = 0,88150675$ y $N_1 = 31$
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y2$. Aparecen los estadísticos de $y2$: $\bar{y}_2 = 3,278717$, $S_2^* = 0,83869$ y $N_2 = 19$
   - Marcar *Suponer desviación típica poblacional común.*
   - Marcar *Mostrar el gráfico de la distribución muestral* y pinchar en *Aplicar.*

   El resultado es una tabla y un gráfico con la distribución $t(50-2)$ y el valor muestral del estadístico.

```
Hipótesis nula: Diferencia de medias = 0
Muestra 1: n = 31, media = 4,304, d.t. = 0,881507
        desviación típica de la media = 0,158323
        Intervalo de confianza 95% para la media: 3,98066 a 4,62734

Muestra 2: n = 19, media = 3,27872, d.t. = 0,838691
        desviación típica de la media = 0,192409
        Intervalo de confianza 95% para la media: 2,87448 a 3,68295

Estadístico de contraste: t(48)= (4,304-3,27872)/0,252229 =4,0649
valor p a dos colas = 0,0001774 (a una cola = 8,871e-005)
```

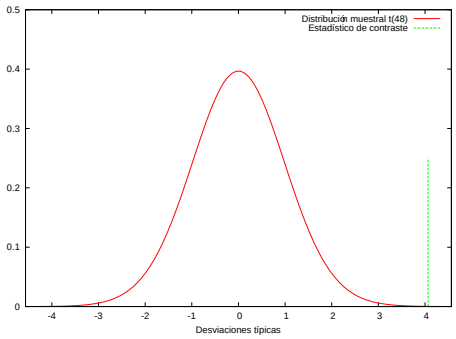


Gráfico A.14: Ejemplo 3: Resultado y distribución del estadístico bajo H_0

2. Definir la región crítica: se trata de un contraste a una cola, por tanto, buscamos c tal que $0,05 = Prob(t > c)$. Vamos a *Herramientas* $\rightarrow$*Tablas estadísticas* $\rightarrow t$, grados de libertad *gl 48* y para $\alpha = 5\,\%$, obtenemos $c = 1,229$.
3. Resultado del contraste: $4,06496 > 1,229$, por tanto, al nivel de significación del $5\,\%$ rechazamos la hipótesis nula de igualdad de medias. Es decir, los datos apoyan la hipótesis de que el precio del m^2 es mayor en los pisos reformados.

Bibliografía

Peña, D. y J. Romo (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill.

Bibliografía

Alonso, A., Fernández, F. & Gallastegui, I. (2005), *Econometría*, Prentice-Hall, Madrid.

Davidson, D. & Mackinnon, J. (2004), *Econometric Theory and Methods*, Oxford University Press, New York.

Engle, R. (1982), 'A general approach to lagrangian multiplier modelo diagnostics', *Journal of Econometrics* **20**, 83–104.

Greene, W. (2008), *Econometric Analysis*, 6th edn, Prentice-Hall, Englewood Cliffs, New Jersey.

Gujarati, D. . (1997), *Econometría*, 4ª edn, McGraw-Hill, México.

Heij, C., de Boer, P., Franses, P., Kloek, T. & Dijk, H. V. (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press, Oxford.

Neter, J., Wasserman, W. & Kutner, M. (1990), *Applied Linear Statistical Models*, 3ª edn, M.A: Irwin, Boston.

Peña, D. & Romo, J. (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill, Madrid.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn, South-Western, Mason, Ohio.

Stock, J. & Watson, M. (2003), *Introduction to Econometrics*, Addison-Wesley, Boston.

Verbeek, M. (2004), *A Guide to Modern Econometrics*, 2ª edn, John Wiley, England.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2ª edn, South-Western, Mason, Ohio.