

Sesión 4 (B) — Importar datos desde internet: el Boston Housing original

Marcos Bujosa

Descripción de la práctica

En la práctica anterior usamos `hprice2`, la versión de `manual` del conjunto de datos sobre la vivienda en Boston (Harrison & Rubinfeld, 1978). Esa versión, distribuida con Gretl, *no incluye todas las variables del estudio original*. En particular, le falta una variable *dicotómica* muy útil para el curso: `chas` (vale 1 si la zona limita con el río Charles, 0 en caso contrario).

En esta práctica aprenderemos a *importar datos desde internet* —una destreza imprescindible cuando trabajamos con datos propios— y cargaremos el *Boston Housing original*, que sí contiene `chas`. Este conjunto será nuestro *dataset integrador* cuando estudiemos variables dicotómicas y en el laboratorio final.

Objetivo:

1. Descargar un fichero de datos (formato CSV) desde internet.
2. Importarlo en Gretl y reconocer sus variables.
3. Declarar correctamente el conjunto como datos de sección cruzada.
4. Examinar la variable dicotómica `chas` y entender qué significa su *media*.
5. Comparar este dataset con el `hprice2` de la práctica anterior.

Actividad 1 - Descargar el fichero de datos

El conjunto original está disponible en varios repositorios públicos. Usaremos una copia en formato CSV (valores separados por comas), el formato de intercambio de datos más habitual.

- Descargue el fichero [boston.csv](#) del repositorio de la asignatura. Es la vía recomendada: no exige registrarse en ningún sitio y sus columnas vienen ya en minúsculas.
- Como alternativa, el mismo conjunto está en repositorios públicos de datos, por ejemplo el [Boston Housing Dataset en Kaggle](#) (allí hace falta una cuenta y los nombres de las columnas vienen en mayúsculas).
- Guarde el fichero en una carpeta de trabajo conocida (p. ej. su carpeta de la asignatura).

Un fichero CSV es simplemente texto: la primera fila suele contener los nombres de las variables y cada fila siguiente, una observación, con los valores separados por comas. Puede abrirlo con cualquier editor de texto para inspeccionarlo antes de importarlo.

Según la fuente, los nombres de las variables pueden variar (mayúsculas/minúsculas). Las variables del estudio original son: `crim`, `zn`, `indus`, `chas`, `nox`, `rm`, `age`, `dis`, `rad`, `tax`, `ptratio`, `b`, `lstat` y `medv` (valor mediano de la vivienda, en miles de dólares).

Si la fuente es [Boston Housing Dataset en Kaggle](#), los nombres aparecerán en mayúsculas (cámbielos a minúsculas, o use mayúsculas en la sesión de Gretl). Si la fuente es el repositorio del curso, los nombres estarán en minúsculas.

Licencia: Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

Actividad 2 - Importar el CSV en Gretl

- En la ventana principal de [Gretl](#), seleccione Archivo ->Abrir archivo de datos ->Archivo de usuario en el menú desplegable en la parte inferior de la pantalla escoja **texto/CSV** y elija el fichero descargado.
- Gretl detectará el separador (coma) y la fila de nombres. Acepte las opciones por defecto y confirme.
o bien teclee en línea de comandos (ajustando la ruta a su sistema):

```
open ../../data/boston.csv --quiet
```

Actividad 3 - Declarar los datos como sección cruzada

Al importar un CSV, Gretl trata las observaciones como una serie sin estructura temporal, lo que es correcto: estos son datos de *sección cruzada* (cada fila es una zona del área de Boston, no un instante de tiempo). Si Gretl preguntara, indique que *no* se trata de series temporales ni de panel.

en línea de comandos, para asegurarse:

```
setobs 1 1 --cross-section
```

Actividad 4 - La variable dicotómica chas

La variable **chas** es *dicotómica* (o *binaria*, o *dummy*): solo toma los valores 0 y 1. Vale 1 si la zona limita con el río Charles y 0 en caso contrario.

Mostrar sus valores

- Marque **chas**, “pinche” con el botón derecho y seleccione **mostrar valores**.

o bien teclee en línea de comandos:

```
print -o chas
```

Comprobará que solo contiene ceros y unos.

Calcular su media

- Marque **chas**, “pinche” con el botón derecho y seleccione **Estadísticos descriptivos**.

o bien teclee en línea de comandos:

```
summary chas
```

Estadísticos principales, usando las observaciones 1 - 506
para la variable 'chas' (506 observaciones válidas)

Media	0,069170
Mediana	0,0000
Mínimo	0,0000
Máximo	1,0000
Desviación típica	0,25399
C.V.	3,6720
Asimetría	3,3958
Exc. de curtosis	9,5315
Percentil del 5%	0,0000
Percentil del 95%	1,0000
Rango intercuartílico	0,0000
Observaciones ausentes	0

Observe que la media de una variable 0/1 es justamente la proporción de unos: el porcentaje de zonas que limitan con el río. Esta idea —la media de una dummy es una proporción— reaparecerá al interpretar los coeficientes de variables dicotómicas en la regresión.

Actividad 5 - Comparar con hprice2

Compare el valor mediano de la vivienda en ambos datasets. En el original, la variable es **medv** (en miles de dólares); en **hprice2**, **price** (en dólares).

- Obtenga los estadísticos descriptivos de **medv**.

en línea de comandos:

```
summary medv rm nox
```

	Media	Mediana	Mínimo	Máximo
medv	22,533	21,200	5,0000	50,000
rm	6,2846	6,2085	3,5610	8,7800
nox	0,55470	0,53800	0,38500	0,87100
	Desv. Típica.	C.V.	Asimetría	Exc. de curtosis
medv	9,1971	0,40817	1,1048	1,4686
rm	0,70262	0,11180	0,40241	1,8610
nox	0,11588	0,20890	0,72714	-0,075864
	Perc. 5%	Perc. 95%	Rango IQ	Observaciones ausentes
medv	10,200	43,695	8,0500	0
rm	5,3040	7,6327	0,74125	0
nox	0,40900	0,74000	0,17500	0

¿Coinciden el número de observaciones (506) y los órdenes de magnitud (recordando el cambio de unidades) con los de **hprice2**? Debería tratarse, esencialmente, de los mismos datos.

Preguntas de interpretación para la clase

1. ¿Qué ventaja tiene saber importar datos en formato CSV, frente a depender solo de los datasets de muestra de Gretl?
2. La media de **chas** es un número entre 0 y 1. ¿Qué proporción de zonas limita con el río Charles? ¿Por qué la media de una variable 0/1 es una proporción?
3. ¿Por qué decimos que estos datos son de *sección cruzada* y no una serie temporal?

4. Las variables `medv` (original) y `price` (`hprice2`) miden lo mismo en *unidades distintas*. ¿Cómo afectaría esto a la interpretación de un coeficiente de regresión, si usáramos una u otra?
5. La variable `chas` es una lista de 506 ceros y unos: un *vector de* $\mathbb{R}^{506}$. ¿Qué *dirección* tiene ese vector en relación con el vector de unos `1`?

Para profundizar

- Cottrell, A. y Lucchetti, R. (2023). *Gretl User's Guide*. Capítulo sobre importación de datos (Importing data). Disponible en gretl.sourceforge.net.
- Adkins, L. C. (2018). *Using Gretl for Principles of Econometrics*, 4.^a ed. Apéndice sobre lectura de datos externos.
- Harrison, D. and Rubinfeld, D.L. (1978), "[Hedonic Housing Prices and the Demand for Clean Air](#)", *Journal of Environmental Economics and Management* 5, 81-102.

Código completo de la práctica

```
# -----  
# Copyright (C) 2026 Marcos Bujosa  
# Licencia: GNU General Public License v3.0 o posterior  
# Este código es software libre y puede ser redistribuido y/o modificado bajo los términos de la GPL.  
# Ver el archivo LICENSE del repositorio para más detalles.  
# -----  
  
# Los dos primeros comandos son necesarios para que Gretl guarde los resultados de la práctica en el  
↪ directorio de trabajo  
# al ejecutar lo siguiente desde un terminal (use los nombres y ruta que correspondan)  
#  
#   DIRECTORIO="Nombre_Directorio_trabajo" gretlcli -b ruta/nombre_fichero_de_la_practica.inp  
#  
# Si esto no le funciona en su sistema, comente las siguientes dos líneas y sitúese en el directorio de  
↪ trabajo de gretl  
# que corresponda (configure dicho directorio de trabajo desde la ventana principal de Gretl).  
  
string directory = getenv("DIRECTORIO")  
set workdir "@directory"  
  
open ../../data/boston.csv --quiet  
  
setobs 1 1 --cross-section  
  
print -o chas  
  
outfile --quiet resumen_chas.txt  
    summary chas  
end outfile  
  
outfile --quiet resumen_medv.txt  
    summary medv rm nox  
end outfile
```

Enlace al guión: [S04-Prct-B-boston.inp](#)

Respuestas

1. *Ventaja de saber importar CSV.* Permite trabajar con *cualquier* fuente de datos propia (no solo los datasets de muestra empaquetados con Gretl), una destreza imprescindible en cuanto el alumno necesite datos de su propio interés (TFG, trabajos aplicados, datos institucionales, etc.).
2. *Media de chas: ¿qué proporción de zonas limita con el río?* En el Boston Housing original, 35 de las 506 zonas censales limitan con el río Charles, es decir, una proporción de $35/506 \approx 0,069$ (6,9%). La media de una variable 0/1 es siempre la proporción de unos porque $\mu_{\mathbf{x}} = \frac{1}{n} \sum x_i$ y, al ser cada x_i o bien 0 o bien 1, la suma $\sum x_i$ es sencillamente el *número* de unos; dividida por n , da la *proporción*.
3. *¿Por qué sección cruzada y no serie temporal?* Porque cada fila representa una *unidad distinta* (una zona geográfica), no un instante de tiempo dentro de la evolución de una misma unidad. No existe una relación de .antes/después. entre las filas: podríamos reordenarlas arbitrariamente sin perder ninguna información (a diferencia de una serie temporal, donde el orden es esencial).
4. *medv frente a price: lo mismo en unidades distintas.* Si *medv* está en miles de dólares y *price* en dólares, un coeficiente de regresión estimado usando *medv* como variable dependiente sería 1000 veces *menor* (en valor absoluto) que el estimado usando *price*, para la misma relación económica subyacente. El cambio de unidades no altera el fenómeno real, solo reescala numéricamente el coeficiente —es la misma idea de "homogeneidad ante escala" que se demuestra formalmente en la lección 5 para la covarianza ($\sigma_{\lambda \mathbf{xy}} = \lambda \sigma_{\mathbf{xy}}$)—. Por eso siempre hay que fijarse en las unidades antes de interpretar la magnitud de un coeficiente.
5. *chas como vector de $\mathbb{R}^{506}$: ¿qué dirección tiene respecto a $\mathbf{1}$?* (Esta respuesta se adelanta a las lecciones 4 y 5: hoy basta con quedarse con la idea; recobrará su sentido pleno al estudiar la media como proyección.) *chas* no es ortogonal a $\mathbf{1}$ en general (su media, $\approx 0,069$, no es cero), así que se descompone en una componente en $\mathcal{L}(\mathbf{1})$ —su vector de medias, $0,069\mathbf{1}$ — y una componente ortogonal a $\mathbf{1}$ —su vector en desviaciones—. Al ser una variable 0/1, esa componente en desviaciones tiene una estructura muy particular: solo toma dos valores, $1 - 0,069$ (en las zonas junto al río) y $-0,069$ (en el resto), y su norma al cuadrado (la varianza) es $p(1 - p)$ con p la proporción de unos. Es un buen ejemplo de que incluso una variable tan simple como una dummy admite la misma descomposición geométrica que cualquier otra.