

Sesión 4 (A) — Primer contacto con Gretl: el dataset hprice2

Marcos Bujosa

Descripción de la práctica

Esta es nuestra *primera práctica* con [Gretl](#). No haremos todavía ninguna regresión: el objetivo de hoy es *tomar contacto con datos reales* y con la interfaz del programa. Usaremos el conjunto `hprice2`, sobre precios de la vivienda en 506 zonas del área de Boston (Harrison & Rubinfeld, 1978), que será nuestro *dataset recurrente* durante todo el curso: lo reencontraremos al estudiar formas funcionales con logaritmos (el precio admite transformación logarítmica). En la siguiente práctica trabajaremos con una versión ampliada de estos datos que incluye además una variable dicotómica o dummy (`chas`, proximidad al río Charles).

Trabajamos con datos *reales* (y no simulados) porque queremos enfrentarnos desde el inicio a la dificultad propia de la realidad económica. Los datos simulados los reservaremos, más adelante, para *verificar* propiedades teóricas.

Objetivo

Tras cargar los datos, cada actividad persigue un objetivo:

1. *Actividad 1*: mostrar los datos y empezar a leer cada variable como un *vector* de $\mathbb{R}^n$.
2. *Actividad 2*: editar los atributos de las variables.
3. *Actividad 3*: obtener estadísticos descriptivos.
4. *Actividad 4*: generar un histograma.
5. *Actividad 5*: generar un diagrama de caja.

Comencemos cargando los datos:

Archivo -> Abrir archivo de datos -> Archivo de muestra y, en la pestaña Wooldridge, seleccione `hprice2`.

o bien teclee en línea de comandos:

```
open hprice2.gdt
```

Gretl no preguntará por la frecuencia: son datos de *sección cruzada* (cada observación es una zona, no un instante de tiempo), de modo que *el orden de las observaciones es irrelevante*, tal como vimos en la lección 1.

Actividad 1 - Mostrar los datos

Visualice los valores de los precios de la vivienda:

- En la ventana principal de [Gretl](#), marque con el ratón la variable: `price`.
- “Pinche” sobre ella con el botón derecho del ratón.

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

- Seleccione **mostrar valores** del menú desplegable que se ha abierto al pinchar.

o bien teclee en línea de comandos:

```
print -o price
```

La variable `price` es una *lista de 506 números*. Para nosotros, eso es un *vector de $\mathbb{R}^{506}$* . Lo mismo vale para `rooms`, `nox` y cualquier otra variable del fichero.

Para Acceder a la ayuda

Para consultar la documentación sobre cualquier comando, puede emplear el menú desplegable **Ayuda** que aparece arriba, a la derecha de la ventana principal de **Gretl**.

- **Ayuda ->Guía de Instrucciones** y “pinche” sobre `print`

o bien teclee en línea de comandos: help print

Actividad 2 - Editar los atributos de las variables

Las variables traen ya cierta información, pero podemos enriquecerla con descripciones y etiquetas que aparecerán después en gráficos y salidas.

Editar la descripción de price

- Marque la variable `price`, “pinche” sobre ella con el botón derecho y seleccione **Editar atributos**.
- Escriba como descripción **Precio mediano de la vivienda (\$)** y como nombre a mostrar en gráficos **Precio (\$)**.

o bien teclee en línea de comandos:

```
setinfo price -d "Precio mediano de la vivienda ($)" -n "Precio ($)"
```

Haga lo mismo con rooms y con nox

en línea de comandos:

```
setinfo rooms -d "Número medio de habitaciones" -n "Habitaciones"
setinfo nox -d "Concentración de óxidos de nitrógeno" -n "Contaminación (NOx)"
```

Compruebe las etiquetas de todas las variables

- Las etiquetas se pueden leer en la pantalla principal a la derecha del nombre de cada variable.

o bien teclee en línea de comandos:

```
labels
```

Para leer la información registrada del conjunto completo de datos

Emplee el menú desplegable **Datos ->Información del conjunto de datos**.

Actividad 3 - Estadísticos descriptivos

Queremos resumir numéricamente varias variables a la vez.

- Marque con el ratón las variables **price**, **rooms** y **nox** (manteniendo pulsada la tecla **Ctrl**).
- “pinche” con el botón derecho y seleccione **Estadísticos descriptivos**.
- Gretl preguntará si debe mostrar todos los estadísticos o solo los más importantes (explore ambas opciones).

o bien teclee en línea de comandos:

```
summary price rooms nox
```

	Media	Mediana	Mínimo	Máximo
price	22512	21200	5000,0	50001
rooms	6,2841	6,2100	3,5600	8,7800
nox	5,5498	5,3800	3,8500	8,7100
	Desv. Típica.	C.V.	Asimetría	Exc. de curtosis
price	9208,9	0,40907	1,1063	1,4594
rooms	0,70259	0,11181	0,40281	1,8595
nox	1,1584	0,20873	0,72258	-0,080785
	Perc. 5%	Perc. 95%	Rango IQ	Observaciones ausentes
price	10200	43695	8199,0	0
rooms	5,3000	7,6295	0,74250	0
nox	4,0900	7,4000	1,7500	0

Obtendrá, para cada variable, la *media*, la *mediana*, el *mínimo*, el *máximo*, la *desviación típica*, etc.

Semilla geométrica (que germinará en la práctica de correlaciones): la media de una variable no es un número cualquiera; veremos que es el resultado de proyectar el vector x sobre la dirección del vector de unos $\mathbf{1}$. Y la desviación típica medirá la distancia de ese vector respecto de su proyección; es decir, la /norma del vector en desviaciones $x - \mu_x \mathbf{1}$, donde μ_x es la media de x (notación que fijaremos en la lección 4). Hoy solo leemos los números; pronto les daremos interpretación geométrica./

Actividad 4 - Histograma

Para ver la *distribución* de una variable:

- Marque **price**, “pinche” con el botón derecho y seleccione **Distribución de frecuencias**.
- Gretl preguntará por el número de intervalos; acepte el valor por defecto.

o bien teclee en línea de comandos:

```
freq price --plot=display
```

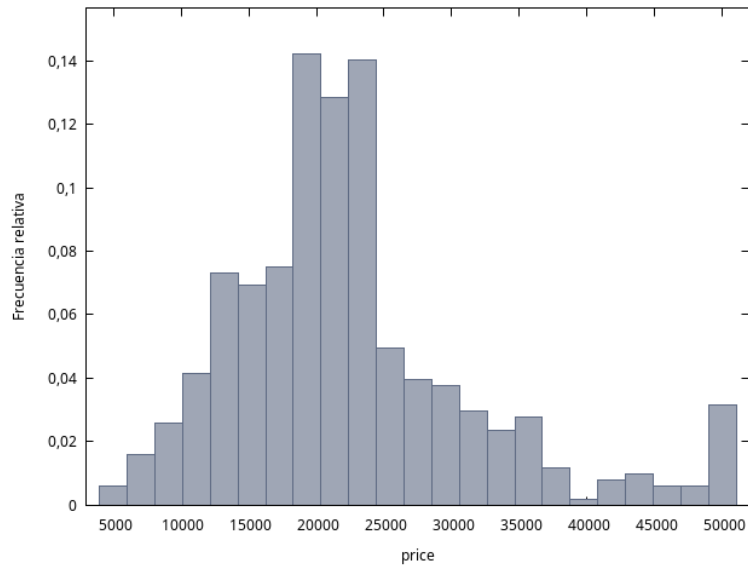


Figura 1: Distribución de frecuencias de **price**.

Observe la forma de la distribución de **price**: ¿es simétrica?, ¿se extiende la distribución hacia la derecha (cola larga de precios altos)? Repita el histograma para **rooms** y compare.

Actividad 5 - Diagrama de caja

El *diagrama de caja* (boxplot) resume la distribución mediante la mediana, los cuartiles y los valores extremos, y es útil para detectar *observaciones atípicas*.

- Marque **price**, “pinche” con el botón derecho y seleccione **Gráfico de caja**.

o bien teclee en línea de comandos:

```
boxplot price --output=display
```

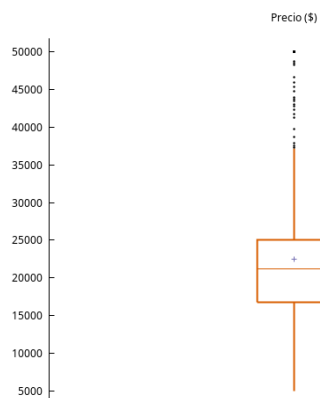


Figura 2: Diagrama de caja de **price**.

Los puntos por encima del “bigote” superior son zonas con precios anormalmente altos. *No los eliminamos: simplemente los observamos.* Más adelante (laboratorio integrador) hablaremos de *observaciones influyentes*.

Preguntas de interpretación para la clase

1. ¿Cuál es el precio medio de la vivienda en la muestra? ¿Y la mediana? ¿Por qué difieren?
2. El histograma de `price`, ¿sugiere simetría o asimetría? ¿Qué implicaría eso para una eventual transformación logarítmica (que veremos más adelante)?
3. La variable `price` es una lista de 506 números. ¿En qué espacio $\mathbb{R}^n$ “vive” como vector? ¿Y `rooms`?
4. ¿Por qué hemos usado datos *reales* y no simulados en esta práctica?

Para profundizar

- Cottrell, A. y Lucchetti, R. (2023). *Gretl User's Guide*. Capítulos introductorios sobre carga de datos y estadística descriptiva. Disponible en gretl.sourceforge.net.
- Adkins, L. C. (2018). *Using Gretl for Principles of Econometrics*, 4.^a ed. Capítulo 1 (introducción al entorno).
- Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. Apéndice sobre estadística descriptiva; descripción del dataset `hprice2` (Harrison y Rubinfeld, 1978).
- 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 hprice2.gdt  
  
print -o price  
  
setinfo price -d "Precio mediano de la vivienda ($)" -n "Precio ($)"  
  
setinfo rooms -d "Número medio de habitaciones" -n "Habitaciones"  
setinfo nox -d "Concentración de óxidos de nitrógeno" -n "Contaminación (NOx)"  
  
labels  
  
outfile --quiet estadisticos.txt  
    summary price rooms nox  
end outfile  
  
freq price --plot="histograma_price.png"  
  
boxplot price --output="boxplot_price.png"
```

Enlace al guión: [S04-Prct-A-hprice2.inp](#)

Respuestas

1. *Precio medio frente a mediana.* Con datos reales de vivienda es habitual que la media sea *mayor* que la mediana: unas pocas zonas con precios muy altos "tiran" de la media hacia arriba, mientras que la mediana —al ser un estadístico de orden, no de suma— apenas se ve afectada por esos valores extremos. Con estos datos ocurre exactamente eso: la media es 22 512 dólares y la mediana 21 200. Es la asimetría positiva que confirma el histograma de la Actividad 4.
2. *Histograma de **price**: ¿simetría o asimetría?* Muestra asimetría *positiva* (cola larga hacia la derecha; el coeficiente de asimetría que da Gretl es 1,11): la mayoría de las zonas tienen precios moderados, y unas pocas zonas "de lujo" alargan la cola derecha. Esto anticipa exactamente el motivo por el que, en la lección 15 (formas funcionales), se transforma **price** con logaritmos: el logaritmo comprime^{es} a cola larga y acerca la distribución a la simetría, lo cual suele mejorar el ajuste lineal.
3. *¿En qué $\mathbb{R}^n$ vive **price**? ¿Y **rooms**?* Ambas viven en $\mathbb{R}^{506}$: el espacio lo determina el número de observaciones ($n = 506$ zonas), no el "tipo" de variable. **price** y **rooms** son dos vectores distintos dentro del *mismo* espacio $\mathbb{R}^{506}$, lo cual es lo que permite compararlos, sumarlos, proyectarlos o medir el ángulo entre ellos (lecciones 4 y 5).
4. *¿Por qué datos reales y no simulados aquí?* Porque el objetivo de esta primera práctica es *tomar contacto* con la irregularidad genuina de la realidad económica (asimetría, escalas, posibles atípicos) desde el principio. Los datos simulados se reservan para cuando necesitemos *verificar* una propiedad teórica de la que conocemos de antemano la "verdad" (por ejemplo, insesgadez de un estimador, o —como en la práctica de verificación geométrica— las identidades de las lecciones 4-5).