

Sesión 14 (A) — Regresión múltiple con Gretl: precios de vivienda (Ramanathan)

Marcos Bujosa

Descripción de la práctica

Empezamos el bloque de regresión múltiple (lección 10) con el conjunto de datos reales más pequeño y manejable de todo el curso: 14 viviendas unifamiliares de University City (San Diego, California), tomado de Ramanathan (2002). Es un dataset de sección cruzada tan reducido que, si quisiera, podría casi reproducir a mano alguno de los cálculos; pero lo usaremos únicamente para lo que de verdad importa hoy: *aprender a leer* una regresión con más de un regresor, e interpretar sus parámetros.

Objetivo

1. Estimar por primera vez, con Gretl, una regresión con más de un regresor no constante.
2. Comparar el coeficiente de un regresor cuando se estima *solo* frente a cuando se estima *junto* a otro: la esencia del efecto parcial o *ceteris paribus* (lección 10).
3. Ampliar el modelo con un tercer regresor y observar cómo cambia (o no) la lectura de los anteriores.
4. Observar, sin todavía formalizarlo, que el R^2 no baja al añadir regresores (se formalizará, con el R^2 ajustado, en la práctica siguiente).
5. Practicar la lectura de una salida de regresión múltiple sin dejarse arrastrar a conclusiones causales injustificadas.
6. Ver cuánto se mueve el ajuste al quitar una sola vivienda, y que no todas pesan igual: el primer indicio de que una estimación tiene una *dispersión*, que la lección 11 cuantificará.

Actividad 1 - Carga y exploración de los datos

- GUI: Archivo -> Abrir archivo de datos -> Archivo de muestra, pestaña Ramanathan, fichero data4-1 (Prices of single-family homes).

o bien teclee en línea de comandos:

```
open data4-1.gdt --quiet
```

Si lo consideramos conveniente podemos castellanizar las etiquetas de las variables para tener gráficos con las leyendas en español. Además, siempre conviene echar un vistazo a los estadísticos descriptivos principales de las variables.

en línea de comandos:

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

```
setinfo price -d "Precio de la vivienda (miles de dólares)" -n "Precio (miles $)"
setinfo sqft -d "Superficie habitable (pies cuadrados)" -n "Superficie (sqft)"
setinfo bedrms -d "Número de dormitorios" -n "Dormitorios"
setinfo baths -d "Número de cuartos de baño" -n "Baños"

summary price sqft bedrms baths --simple
```

Son solo 14 observaciones: puede inspeccionarlas todas de un vistazo con

```
print price sqft bedrms baths --byobs
```

	price	sqft	bedrms	baths
1	199,9	1065	3	1,75
2	228,0	1254	3	2,00
3	235,0	1300	3	2,00
4	285,0	1577	4	2,50
5	239,0	1600	3	2,00
6	293,0	1750	4	2,00
7	285,0	1800	4	2,75
8	365,0	1870	4	2,00
9	295,0	1935	4	2,50
10	290,0	1948	4	2,00
11	385,0	2254	4	3,00
12	505,0	2600	3	2,50
13	425,0	2800	4	3,00
14	415,0	3000	4	3,00

Un conjunto de datos tan pequeño no es representativo de ningún mercado real, pero es perfecto para una primera toma de contacto con la regresión múltiple.

Actividad 2 - Recordatorio: la regresión simple $\text{price} \sim \text{sqft}$

Antes de añadir regresores, recordamos el punto de partida: la regresión simple ya vista en la lección 6. *en línea de comandos:*

```
ols price const sqft
```

Modelo 1: MCO, usando las observaciones 1-14

Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p

const	52,3509	37,2855	1,404	0,1857
sqft	0,138750	0,0187329	7,407	8,20e-06 ***
Media de la vble. dep.	317,4929	D.T. de la vble. dep.	88,49816	
Suma de cuad. residuos	18273,57	D.T. de la regresión	39,02304	
R-cuadrado	0,820522	R-cuadrado corregido	0,805565	
F(1, 12)	54,86051	Valor p (de F)	8,20e-06	
Log-verosimilitud	-70,08421	Criterio de Akaike	144,1684	
Criterio de Schwarz	145,4465	Crit. de Hannan-Quinn	144,0501	

De dicha salida, ya podemos reconocer algunos de los resultados obtenidos:

- Los números de la columna bajo el nombre **coeficiente** son los valores estimados para los parámetros β_1 (constante) y β_2 (sqft).

- El número a la derecha del nombre **Media de la vble. dep.** es μ_y donde y es el vector con los datos de los precios de las 14 viviendas unifamiliares.
- El número a la derecha del nombre **Suma de cuad. residuos** es SRC, es decir, $\|\hat{e}\|_e^2$.
- El número a la derecha del nombre **R-cuadrado** es el coeficiente de determinación R^2 del ajuste.

Veamos la recta de regresión correspondiente al ajuste MCO que hemos logrado:
en línea de comandos:

```
gnuplot price sqft --output="price_sqft_simple.png"
```

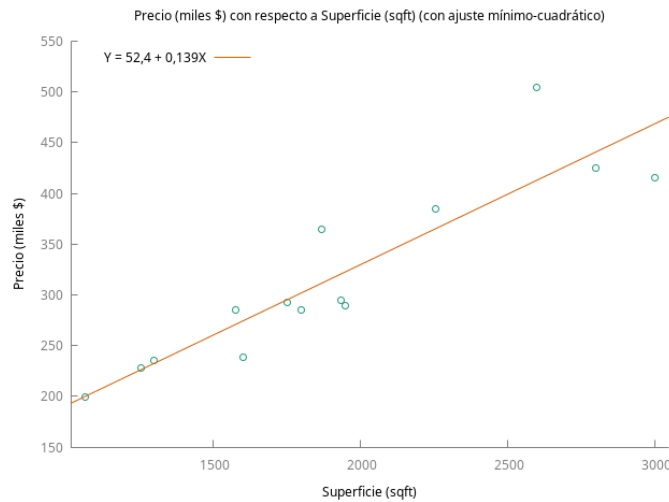


Figura 1: Precio frente a superficie, con la recta ajustada (regresión simple).

Anote el valor de $\hat{\beta}_2$ (coeficiente de **sqft**) y de R^2 : serán su punto de referencia para todo lo que sigue.

Actividad 3 - Más regresiones: con **bedrms**, y con **sqft + bedrms**

Estimamos ahora **price** en función únicamente de **bedrms**, y a continuación en función de **sqft** y **bedrms** a la vez.

Antes de mirar el resultado: ¿qué signo espera para el coeficiente de **bedrms** en el modelo simple? ¿Y en el modelo con **sqft** incluido? ¿Espera que sea *el mismo número* en ambos casos?

en línea de comandos:

```
ols price const bedrms
ols price const sqft bedrms
```

Modelo 2: MCO, usando las observaciones 1-14
 Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p
const	112,853	179,123	0,6300	0,5405
bedrms	56,1756	48,7511	1,152	0,2716
Media de la vble. dep.	317,4929	D.T. de la vble. dep.		88,49816
Suma de cuad. residuos	91671,71	D.T. de la regresión		87,40314

R-cuadrado	0,099625	R-cuadrado corregido	0,024594
F(1, 12)	1,327777	Valor p (de F)	0,271640
Log-verosimilitud	-81,37352	Criterio de Akaike	166,7470
Criterio de Schwarz	168,0252	Crit. de Hannan-Quinn	166,6287

Modelo 3: MCO, usando las observaciones 1-14
Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p
const	121,179	80,1778	1,511	0,1589
sqft	0,148314	0,0212080	6,993	2,29e-05 ***
bedrms	-23,9106	24,6419	-0,9703	0,3527
Media de la vble. dep.	317,4929	D.T. de la vble. dep.	88,49816	
Suma de cuad. residuos	16832,79	D.T. de la regresión	39,11846	
R-cuadrado	0,834673	R-cuadrado corregido	0,804613	
F(2, 11)	27,76736	Valor p (de F)	0,000050	
Log-verosimilitud	-69,50933	Criterio de Akaike	145,0187	
Criterio de Schwarz	146,9358	Crit. de Hannan-Quinn	144,8412	

Lo que debe observar

Compare el coeficiente de **bedrms** en ambas regresiones: solo, vale +56,2 (mil dólares por dormitorio) con un R^2 de apenas 0,10; junto a **sqft**, pasa a -23,9, y el R^2 sube a 0,83, casi el mismo que tenía **sqft** sola. En la lección 10 vimos que cada $\hat{\beta}_j$, en un modelo con varios regresores, mide un *efecto parcial*: “el efecto promedio en **price** ante un dormitorio más, *manteniendo fija la superficie*”. Eso es una pregunta distinta de “el efecto promedio en **price** ante un dormitorio más, sin ninguna información adicional” (lo que mide el coeficiente del modelo simple). No tienen por qué coincidir, ni siquiera en signo: dos viviendas con la misma superficie pero distinto número de dormitorios reparten ese mismo espacio en habitaciones más pequeñas o más grandes, lo cual puede pesar en el precio de forma diferente a como pesa “más dormitorios sin más información sobre otras características de la vivienda” (pues muchos dormitorios solo caben en viviendas grandes, por lo que posiblemente en la muestra habrá una correlación positiva entre el número de habitaciones y superficie... verifíquelo).

en línea de comandos:

```
corr sqft bedrms
```

```
corr(sqft, bedrms) = 0,46473046
Bajo la hipótesis nula de no correlación:
t(12) = 1,81814, con valor p a dos colas 0,0941
```

Actividad 4 - Añadimos baths

Ampliamos el modelo con otro regresor adicional: el número de cuartos de baño.

en línea de comandos:

```
ols price const sqft bedrms baths
```

Modelo 4: MCO, usando las observaciones 1-14
Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p
--	-------------	--------------	---------------	---------

const	129,062	88,3033	1,462	0,1746
sqft	0,154800	0,0319404	4,847	0,0007 ***
bedrms	-21,5875	27,0293	-0,7987	0,4430
baths	-12,1928	43,2500	-0,2819	0,7838

Media de la vble. dep.	317,4929	D.T. de la vble. dep.	88,49816
Suma de cuad. residuos	16700,07	D.T. de la regresión	40,86572
R-cuadrado	0,835976	R-cuadrado corregido	0,786769
F(3, 10)	16,98894	Valor p (de F)	0,000299
Log-verosimilitud	-69,45391	Criterio de Akaike	146,9078
Criterio de Schwarz	149,4641	Crit. de Hannan-Quinn	146,6712

Sin considerar la constante, el valor p más alto fue el de la variable 4 (baths)

Compare de nuevo el coeficiente de **bedrms** (y ahora también el de **sqft**) con los obtenidos en la Actividad 3: apenas se mueven ($-23,9 \rightarrow -21,6$ y $0,148 \rightarrow 0,155$). ¿Qué signo tiene el coeficiente de **baths**, $-12,2$, y le parece razonable? Si no le parece razonable ¿Cuál podría ser el motivo por el que el resultado obtenido es distinto de lo que usted esperaba?

Observe la matriz de correlaciones entre **sqft**, **bedrms** y **baths**:

en línea de comandos:

```
corr sqft bedrms baths
```

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

sqft	bedrms	baths
1,0000	0,4647	0,7873
	1,0000	0,5323
		1,0000

A la luz de los resultados, ¿qué signo esperaría ver para **baths** en la regresión lineal simple **ols price const baths**? **Verifique su intuición estimando el modelo.**

¿Cree que el valor estimado para **baths** se debe a la valoración de los baños en el mercado inmobiliario? ¿O a que realmente se está midiendo inadvertidamente otra cosa?

Actividad 5 - Síntesis: cuatro modelos, una tabla

Complete la tabla con los resultados de las cuatro regresiones estimadas en esta práctica.

Modelo	$\hat{\beta}_{\text{sqft}}$	$\hat{\beta}_{\text{bedrms}}$	$\hat{\beta}_{\text{baths}}$	R^2
price~sqft		—	—	
price~bedrms	—		—	
price~sqft+bedrms			—	
price~sqft+bedrms+baths				

en línea de comandos, para obtener los cuatro R^2 de golpe:

```
ols price const sqft --quiet
scalar r2_1 = $rsq
ols price const bedrms --quiet
scalar r2_2 = $rsq
ols price const sqft bedrms --quiet
scalar r2_3 = $rsq
ols price const sqft bedrms baths --quiet
scalar r2_4 = $rsq

printf "R2 price-sqft" = %.4f\n", r2_1
printf "R2 price-bedrms" = %.4f\n", r2_2
printf "R2 price-sqft+bedrms" = %.4f\n", r2_3
printf "R2 price-sqft+bedrms+baths" = %.4f\n", r2_4
```

```

R2 price-sqft          = 0,8205
R2 price-bedrms        = 0,0996
R2 price-sqft+bedrms   = 0,8347
R2 price-sqft+bedrms+baths = 0,8360

```

Observe el R^2 a medida que se añaden regresores: nunca *baja*. Esto no es casualidad ni una propiedad de estos datos en particular: es una propiedad puramente algebraica de MCO (cuya idea sembramos en la lección 8), que discutiremos con detalle —y con su corrección adecuada, el R^2 ajustado— en la próxima práctica.

Actividad 6 - ¿Cuánto depende el ajuste de las viviendas que nos han tocado?

Hasta aquí hemos cambiado el *modelo* y hemos visto moverse los coeficientes. Ahora vamos a hacer justo lo contrario: dejar el modelo quieto y cambiar los *datos*. Concretamente, vamos a quitar una vivienda de las catorce y reestimar la misma regresión.

La pregunta es sencilla y conviene plantearse la antes de ejecutar nada: *si una sola de estas catorce casas hubiera sido distinta —o sencillamente no estuviera en nuestra lista—, ¿cambiaría mucho la recta que hemos ajustado?*

Gretl permite restringir la muestra con `smpl`. Desde la interfaz gráfica, lo natural es ir acumulando los ajustes en la *tabla de modelos* (menú Archivo → Tabla de modelos, o el comando `modeltab add`); en un guion que se ejecuta por lotes, en cambio, resulta más cómodo guardar los coeficientes en escalares e imprimirlos juntos al final. Eso es lo que hace el código siguiente.

Estimamos el mismo modelo quitando, cada vez, una sola vivienda

Quitaremos siempre *una* vivienda, y elegiremos tres que se diferencian en dos rasgos distintos: dónde están dentro del rango de superficies, y a qué distancia quedan de la recta ajustada con las catorce. Anotamos ambas cosas junto a cada resultado, porque la comparación entre las tres filas es todo el contenido de la actividad.

o bien teclee en línea de comandos:

```

smpl full          # las 14 viviendas
ols price const sqft --quiet
scalar a_todas = $coeff(const)
scalar b_todas = $coeff(sqft)
series resid = $uhat # distancia de cada vivienda a la recta
scalar media_sqft = mean(sqft)

smpl full
smpl obs != 1 --restrict --quiet # fuera la vivienda 1: la MAS PEQUEÑA
ols price const sqft --quiet
scalar a_s1 = $coeff(const)
scalar b_s1 = $coeff(sqft)

smpl full
smpl obs != 12 --restrict --quiet # fuera la 12: NO es extrema en tamaño
ols price const sqft --quiet
scalar a_s12 = $coeff(const)
scalar b_s12 = $coeff(sqft)

smpl full
smpl obs != 14 --restrict --quiet # fuera la 14: la MAS GRANDE
ols price const sqft --quiet
scalar a_s14 = $coeff(const)
scalar b_s14 = $coeff(sqft)

smpl full          # recuperamos siempre la muestra completa

```

y ahora los comparamos:

```

printf "media de sqft en las 14 viviendas = %.1f\n\n", media_sqft
printf "se quita      sqft      residuo      constante      pendiente\n"
printf "-----\n"
printf "nadie (n=14)      -      -      %10.3f %10.4f\n", a_todas, b_todas
printf "obs 1 (n=13)      %5.0f %8.2f %10.3f %10.4f\n", sqft[1], resid[1], a_s1, b_s1
printf "obs 12 (n=13)     %5.0f %8.2f %10.3f %10.4f\n", sqft[12], resid[12], a_s12, b_s12
printf "obs 14 (n=13)     %5.0f %8.2f %10.3f %10.4f\n", sqft[14], resid[14], a_s14, b_s14

```

media de sqft en las 14 viviendas = 1910,9

se quita	sqft	residuo	constante	pendiente
nadie (n=14)	-	-	52,351	0,1388
obs 1 (n=13)	1065	-0,22	52,479	0,1387
obs 12 (n=13)	2600	91,90	78,380	0,1209
obs 14 (n=13)	3000	-53,60	18,962	0,1593

Lo que debe observar

Las tres viviendas que hemos ido quitando se comportan de tres maneras muy distintas, y esa es justamente la lección:

- *La obs. 1* es la más pequeña de todas (1065 pies cuadrados, el mínimo de la muestra): está en el borde mismo del rango. Y sin embargo quitarla no mueve nada: la pendiente pasa de 0,1388 a 0,1387. Mire su residuo: -0,22. Esa vivienda cae prácticamente *encima* de la recta.
- *La obs. 14* es la más grande (3000 pies cuadrados, el máximo): también está en el borde. Pero su residuo es -53,60, y quitarla mueve la pendiente de 0,1388 a 0,1593 y la constante de 52,351 a 18,962.
- *La obs. 12* (2600 pies cuadrados) *no* está en ningún extremo. Y aun así, con un residuo de +91,90, quitarla mueve la pendiente a 0,1209 y la constante a 78,380.

De donde se siguen tres conclusiones que conviene no confundir:

1. *No hemos cambiado el modelo.* La especificación es idéntica en las cuatro filas; lo único que cambia es qué viviendas entran en el cálculo. Todo el movimiento que vea procede exclusivamente de la muestra.
2. *No todas las observaciones pesan igual, y la posición no lo explica sola.* Estar en un extremo del rango de superficies no basta para tirar de la recta (la obs. 1 está en el extremo y no tira de nada), y no estar en un extremo no protege (la obs. 12 está en el interior y es de las que más mueven el ajuste). Lo que hace falta son *las dos cosas a la vez*: estar lejos del centro en el eje del regresor, *y además* quedar lejos de la recta. Volveremos sobre esta idea —con nombre propio, *apalancamiento* e *influencia*, y con una medida concreta— en el laboratorio integrador que cierra el curso.
3. *Con $n = 14$ esto es muy visible; con $n = 506$ lo sería mucho menos.* La fragilidad que acaba de observar no es un defecto de MCO: es lo que ocurre cuando se dispone de pocos datos.

Esta actividad no resuelve nada; deja una pregunta abierta. Hemos comprobado que el ajuste *se mueve* cuando cambian los datos, pero no tenemos todavía ninguna forma de decir *cuánto* se mueve, ni de anticiparlo antes de mirar. Para eso hace falta hablar de la variabilidad del estimador, y eso exige un aparato que aún no tenemos. Es lo que hará la lección 11: dar una fórmula —no una impresión— para la dispersión de $\hat{\beta}_2$, y mostrar que en ella aparecen el número de observaciones y cuán dispersas están las superficies de las viviendas.

Preguntas de interpretación para la clase

1. ¿Por qué el coeficiente de `bedrms` puede cambiar de signo (o de magnitud) al pasar del modelo simple al modelo con `sqft` incluido, si en ambos casos usamos los mismos 14 datos?
2. ¿Tiene sentido interpretar el coeficiente de `sqft` en el modelo `price~sqft+bedrms+baths` como “el efecto de un pie cuadrado más, manteniendo todo lo demás igual”? ¿Y como “lo que causa” ese cambio de precio?
3. Con solo 14 observaciones, ¿qué cautela adicional debería tener a la hora de generalizar cualquier conclusión de esta práctica a “el mercado de la vivienda” en general?
4. ¿Por qué el R^2 no puede bajar nunca al añadir un regresor, sea cual sea ese regresor (incluso uno completamente irrelevante)?
5. Si tuviera que explicarle a alguien sin formación en econometría la diferencia entre “el efecto de una habitación más” y “el efecto de una habitación más, manteniendo fija la superficie”, ¿cómo lo diría con sus propias palabras?

Para profundizar:

- Véase la lección 10 ([lección 10](#)) para la notación matricial y el enunciado del efecto parcial, cuyo desarrollo con datos reales es el objeto de esta práctica.
- Ramanathan, R. (2002). *Introductory Econometrics with Applications*, 5.^a ed. (fuente original del dataset `data4-1`).
- Wooldridge, J. M. (2020). *Introductory Econometrics*, cap. 3, sección 3.4 (interpretación *ceteris paribus*).

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 data4-1.gdt --quiet

setinfo price -d "Precio de la vivienda (miles de dólares)" -n "Precio (miles $)"
setinfo sqft -d "Superficie habitable (pies cuadrados)" -n "Superficie (sqft)"
setinfo bedrms -d "Número de dormitorios" -n "Dormitorios"
setinfo baths -d "Número de cuartos de baño" -n "Baños"

summary price sqft bedrms baths --simple

outfile --quiet vistazoDatos.txt
    print price sqft bedrms baths --byobs
end outfile

outfile --quiet OLS_Precio_Superficie.txt
    ols price const sqft
end outfile

gnuplot price sqft --output="price_sqft_simple.png"

outfile --quiet olsBedrms.txt
    ols price const bedrms
    ols price const sqft bedrms
end outfile

outfile --quiet corrSqftBedrms.txt
    corr sqft bedrms
end outfile

outfile --quiet olsBaths.txt
    ols price const sqft bedrms baths
end outfile

outfile --quiet corrTres.txt
    corr sqft bedrms baths
end outfile

outfile --quiet sintesisR2.txt
    ols price const sqft --quiet
    scalar r2_1 = $rsq
    ols price const bedrms --quiet
    scalar r2_2 = $rsq
    ols price const sqft bedrms --quiet
    scalar r2_3 = $rsq
    ols price const sqft bedrms baths --quiet
```

```

scalar r2_4 = $rsq

printf "R2 price~sqft          = %.4f\n", r2_1
printf "R2 price~bedrms       = %.4f\n", r2_2
printf "R2 price~sqft+bedrms  = %.4f\n", r2_3
printf "R2 price~sqft+bedrms+baths = %.4f\n", r2_4
end outfile

smpl full                                # las 14 viviendas
ols price const sqft --quiet
scalar a_todas = $coeff(const)
scalar b_todas = $coeff(sqft)
series resid   = $uhat                  # distancia de cada vivienda a la recta
scalar media_sqft = mean(sqft)

smpl full
smpl obs != 1 --restrict --quiet        # fuera la vivienda 1: la MAS PEQUEÑA
ols price const sqft --quiet
scalar a_s1 = $coeff(const)
scalar b_s1 = $coeff(sqft)

smpl full
smpl obs != 12 --restrict --quiet       # fuera la 12: NO es extrema en tamaño
ols price const sqft --quiet
scalar a_s12 = $coeff(const)
scalar b_s12 = $coeff(sqft)

smpl full
smpl obs != 14 --restrict --quiet       # fuera la 14: la MAS GRANDE
ols price const sqft --quiet
scalar a_s14 = $coeff(const)
scalar b_s14 = $coeff(sqft)

smpl full                                # recuperamos siempre la muestra completa
outfile --quiet submuestras.txt
printf "media de sqft en las 14 viviendas = %.1f\n\n", media_sqft
printf "se quita          sqft  residuo  constante  pendiente\n"
printf "-----\n"
printf "nadie (n=14)          -          -    %10.3f  %10.4f\n", a_todas, b_todas
printf "obs 1 (n=13)         %5.0f %8.2f  %10.3f  %10.4f\n", sqft[1], resid[1], a_s1, b_s1
printf "obs 12 (n=13)        %5.0f %8.2f  %10.3f  %10.4f\n", sqft[12], resid[12], a_s12, b_s12
printf "obs 14 (n=13)        %5.0f %8.2f  %10.3f  %10.4f\n", sqft[14], resid[14], a_s14, b_s14
end outfile

```

Enlace al guión: [S14-Pret-A-ramanathan.inp](#)

Respuestas

1. ¿Por qué puede cambiar el coeficiente de `bedrms`? Porque en el modelo simple, el coeficiente de `bedrms` recoge *todo* lo que varía con el número de dormitorios, incluida la superficie (las casas con más dormitorios suelen tener también más superficie): está, en cierto sentido, “contaminado” por esa variable omitida. En el modelo múltiple, al incluir `sqft` como regresor, el coeficiente de `bedrms` mide únicamente lo que aporta un dormitorio adicional *una vez fijada la superficie* —dos viviendas de igual tamaño, pero con distinto reparto en habitaciones—. No es sorprendente que ambos números difieran, incluso en signo: están respondiendo a preguntas distintas, aunque provengan de los mismos datos.
2. Interpretación del coeficiente de `sqft` en el modelo `price~sqft+bedrms+baths`. Sí cabe la lectura *ceteris paribus*: “manteniendo fijos `bedrms` y `baths`, un pie cuadrado adicional se asocia, en promedio, con ese cambio en el precio”. Lo que *no* cabe, sin más, es la lectura causal (“un pie cuadrado más *causa* ese cambio”): como se insistió en la lección 8 y en la lección 9 (“El puente”), la descomposición de MCO es puramente algebraica y no certifica, por sí sola, ninguna relación de causa a efecto, sea cual

sea el número de regresores. Aquí, además, con solo 14 observaciones, cualquier lectura —causal o no— debe tomarse con mucha cautela.

3. Cautela con solo 14 observaciones. Una muestra tan pequeña es extremadamente sensible a la presencia de una única vivienda atípica (un precio inusualmente alto o bajo, una combinación poco común de dormitorios y superficie) y ofrece muy poco margen para distinguir el efecto de `sqft` del de `bedrms` si ambas variables están, en esta muestra concreta, bastante correlacionadas entre sí. Generalizar cualquier coeficiente obtenido aquí a “el mercado de la vivienda” sin más matices sería, cuando menos, precipitado: 14 viviendas de un único barrio y un único año no permiten inferir nada sobre otros mercados o épocas.
4. ¿Por qué el R^2 nunca baja al añadir un regresor? Porque MCO elige siempre los coeficientes que minimizan la suma de residuos al cuadrado (SRC) dentro del subespacio disponible; al añadir un regresor, ese subespacio se amplía (contiene, como caso particular, la solución que ya teníamos antes de incluir el regresor nuevo), así que el mínimo de SRC en el subespacio ampliado nunca puede ser más grande que el mínimo anterior. Como $R^2 = 1 - \text{SRC}/\text{STC}$ y STC no cambia, R^2 no puede bajar. Esto es válido incluso si el regresor añadido no tiene ninguna relación real con la variable dependiente —una propiedad puramente mecánica, no una señal de que el nuevo regresor importe—. Si hubiéramos añadido como regresor adicional el número de hermanos del arquitecto que diseñó cada una de las casas, el R^2 probablemente aumentaría; lo que es seguro es que no disminuiría.
5. Explicar la diferencia sin jerga técnica. Una posible formulación: “si comparo, sin más, viviendas con distinto número de habitaciones, estoy comparando peras con manzanas, porque esas viviendas probablemente también difieren en su tamaño; en cambio, si comparo únicamente viviendas del mismo tamaño pero con distinto número de habitaciones, estoy aislando el efecto de repartir ese mismo espacio de una forma u otra”. Esa segunda comparación es lo que mide el coeficiente en el modelo con varios regresores a la vez.