

Sesión 26 (A) — Laboratorio integrador II: palanca, distancia de Cook y el diccionario geometría $\leftrightarrow$ probabilidad

Marcos Bujosa

Descripción de la práctica

La [actividad 2 \(html\)](#) de la práctica A de la sesión 25 dejó estimado el modelo del Boston original: el logaritmo del precio mediano sobre `rm`, `lstat`, `dis` y la indicadora `chas`. Esta práctica pregunta cuánto depende ese ajuste de barrios concretos. La [actividad 6 \(html\)](#) de la práctica A de la sesión 14 hizo la misma pregunta con catorce viviendas: las quitó una a una y miró cómo se movían los coeficientes. Esta práctica usa dos medidas que resumen esa comprobación para las 506 observaciones a la vez: la *palanca* de cada barrio, que es geometría de la lección 10, y la *distancia de Cook*, que combina la palanca con el residuo.

La práctica cierra con un texto sin código: el diccionario que traduce la geometría de $\mathbb{R}^n$ de la primera mitad del curso al lenguaje de la probabilidad de la segunda. Es una analogía explícita, no una demostración, y no es materia de examen.

Objetivos

1. Calcular la palanca de cada barrio con `leverage` y leerla como la longitud al cuadrado de una proyección: cuánto pesa cada fila en su propio valor ajustado.
2. Calcular la distancia de Cook a partir de la palanca y el residuo, localizar el barrio más influyente y comprobar qué cambia al quitarlo.
3. Reconocer un grupo de observaciones influyentes, los 16 barrios con el precio censurado en 50, y ver qué cambia sin ellos.
4. Leer el diccionario geometría $\leftrightarrow$ probabilidad del curso.

Comandos nuevos de esta práctica. Dos:

leverage tras estimar un modelo con `ols`, calcula para cada observación su palanca y dos medidas más que esta práctica no usa, `infl` y `dffits` (la segunda, en Belsley, Kuh y Welsch); con `--save` guarda las tres como series, `lever`, `infl` y `dffits`, y con `--quiet` no imprime la tabla de las 506 filas.

imaxc dada una matriz, devuelve la fila en que cada columna alcanza su máximo; con `{lever}`, la serie convertida en matriz de una columna, da el número de la observación con la palanca máxima.

Los demás ya se han usado: `open`, `series`, `ols` y sus accesorios (`$uhat`, `$sigma`, `$coeff`, `$stderr`, `$nobs`), `summary --simple`, `smpl --restrict` y `smpl full`, `printf`, y las funciones `sum`, `min`, `max`, `mean` y `median`.

Actividad 1 - La palanca de cada barrio

Importe `boston.csv`, estime el modelo con `chas` (Modelo $\rightarrow$ Mínimos cuadrados ordinarios) y pida la palanca de cada observación (en la ventana del modelo, `Análisis  $\rightarrow$  Observaciones influyentes`).

El valor ajustado es la proyección de $\mathbf{y}$ sobre el subespacio generado por las columnas de $\mathbf{X}$ (lección 10), y la componente i de esa proyección es una combinación lineal de todas las componentes de $\mathbf{y}$. El peso de y_i

en su propio valor ajustado $\hat{y}_i$ es la palanca h_i : la longitud al cuadrado de la proyección sobre ese subespacio del vector canónico e_i , el que tiene un 1 en la fila i y ceros en las demás. Es un número entre 0 y 1, y las n palancas suman k , la dimensión del subespacio; su media es k/n . Una palanca grande dice que la fila i está lejos de las demás en el espacio de los regresores: el ajuste pasa cerca de ella porque pocas filas más ocupan esa zona de los regresores.

`leverage --quiet --save` guarda las tres series. `summary lever --simple` da sus estadísticos; `sum(lever)` comprueba que suman $k = 5$ y `median(lever)` da la mediana, que `summary --simple` no imprime. El `printf` compara con la convención de los manuales, $2k/n$, y cuenta cuántos barrios la superan. `imaxc({lever})` devuelve la fila de la palanca máxima, y el último `printf` imprime los datos de ese barrio.

o bien teclee en línea de comandos:

```
open ../../data/boston.csv --quiet
series lmedv = log(medv)
```

```
ols lmedv const rm lstat dis chas --quiet
scalar k = 5
leverage --quiet --save
summary lever --simple
printf "Suma de las palancas: %.2f ; media k/n = %.4f ; mediana %.4f ; convencion 2k/n = %.4f ; barrios por encima: %d\n",
↪ sum(lever), k/$nobs, median(lever), 2*k/$nobs, sum(lever > 2*k/$nobs)
scalar i = imaxc({lever})
printf "Palanca maxima %.4f en el barrio %d: medv %.1f  rm %.2f  lstat %.2f  dis %.2f  chas %d\n", lever[i], i, medv[i],
↪ rm[i], lstat[i], dis[i], chas[i]
```

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

Media	0,0098814
Mínimo	0,0021583
Máximo	0,075518
Desviación típica	0,0094198
Observaciones ausentes	0

Suma de las palancas: 5,00 ; media k/n = 0,0099 ; mediana 0,0061 ; convencion 2k/n = 0,0198 ; barrios por encima: 62
Palanca maxima 0,0755 en el barrio 366: medv 27,5 rm 3,56 lstat 7,12 dis 1,61 chas 0

Lo que debe observar

Las palancas suman 5, el número de columnas de **X**, y su media es 0,0099: en promedio, cada barrio pesa un 1% en su propio valor ajustado. La distribución es asimétrica: la mediana, 0,0061, queda por debajo de la media, y 62 barrios superan la convención $2k/n = 0,0198$. La palanca máxima, 0,0755, es la del barrio 366: tiene `rm` igual a 3,56, el mínimo del fichero, lejos de la media de 6,29. La palanca es grande donde los regresores están lejos de su media, que es también el motivo por el que la constante no se interpreta en un punto sin datos: allí el ajuste es una extrapolación.

Actividad 2 - La distancia de Cook y el barrio más influyente

Una palanca grande dice que la fila *puede* mover el ajuste; si además su residuo es grande, lo mueve. La distancia de Cook combina las dos cosas:

$$D_i = \frac{\hat{e}_i^2 h_i}{k s^2 (1 - h_i)^2},$$

con $\hat{e}_i$ el residuo, h_i la palanca y s^2 la cuasivarianza de los residuos. Mide cuánto se desplaza el vector de valores ajustados al quitar la observación i , en unidades de s^2 y por dimensión. Gretl no la imprime, pero se calcula en una línea con las series que `leverage` guardó.

`$uhat` es la serie de residuos del último modelo y `$sigma` la desviación típica de la regresión, `s`, de modo que `$sigma^2` es `s^2`. La serie `cook` aplica la fórmula fila a fila. El `printf` cuenta cuántos barrios superan la convención $4/n$ y localiza el máximo; el siguiente imprime el residuo y la distancia de Cook del barrio con la palanca máxima, el 366 de la Actividad 1. Después, `smpl cook <max(cook) --restrict` deja fuera ese barrio, se vuelve a estimar y `smpl full` recupera la muestra entera.

```
series e = $uhat
scalar s2 = $sigma^2
series cook = e^2 * lever / (k * s2 * (1 - lever)^2)
printf "Convencion 4/n = %.4f ; barrios por encima: %d\n", 4/$nobs, sum(cook > 4/$nobs)
scalar j = imaxc({cook})
printf "Cook maxima %.3f en el barrio %d: medv %.1f  rm %.2f  lstat %.2f  dis %.2f  chas %d  palanca %.4f  residuo %.3f\n",
↪ cook[j], j, medv[j], rm[j], lstat[j], dis[j], chas[j], lever[j], e[j]
printf "Barrio %d, el de palanca maxima: residuo %.3f  Cook %.4f ; barrios con Cook mayor: %d\n", i, e[i], cook[i], sum(cook
↪ > cook[i])
printf "Con todos:      chas %.4f  rm %.4f  lstat %.4f  dis %.4f\n", $coeff(chas), $coeff(rm), $coeff(lstat), $coeff(dis)
smpl cook < max(cook) --restrict
ols lmedv const rm lstat dis chas --quiet
printf "Sin el barrio %d: chas %.4f  rm %.4f  lstat %.4f  dis %.4f  (n = %d)\n", j, $coeff(chas), $coeff(rm), $coeff(lstat),
↪ $coeff(dis), $nobs
smpl full
```

```
Convencion 4/n = 0,0079 ; barrios por encima: 40
Cook maxima 0,113 en el barrio 365: medv 21,9  rm 8,78  lstat 5,29  dis 1,90  chas 1  palanca 0,0542  residuo -0,699
Barrio 366, el de palanca maxima: residuo 0,387  Cook 0,0506 ; barrios con Cook mayor: 6
Con todos:      chas 0,1597  rm 0,1203  lstat -0,0394  dis -0,0062
Sin el barrio 365: chas 0,1785  rm 0,1292  lstat -0,0391  dis -0,0068  (n = 505)
```

Lo que debe observar

Cuarenta barrios superan la convención $4/n = 0,0079$. La distancia máxima, 0,113, es la del barrio 365, un barrio junto al río con `rm` igual a 8,78, el máximo del fichero, y un precio mediano de 21,9 mil dólares, bajo para ese número de habitaciones: su residuo es $-0,70$ en logaritmos, el valor ajustado dobla el precio observado ($e^{0,70} = 2,0$), y su palanca, 0,054, es cinco veces y media la media. Palanca alta y residuo grande a la vez. El barrio 366, el de palanca máxima, tiene residuo 0,39 y distancia de Cook 0,051, la séptima del fichero. Al quitarlo, el coeficiente de `chas` pasa de 0,160 a 0,179 y el de `rm` de 0,120 a 0,129; un solo barrio entre 506 mueve la diferencia entre grupos en un 12 % de su valor. Que una sola observación mueva el ajuste es lo que la práctica A de la sesión 14 mostró quitando viviendas una a una; la distancia de Cook señala de antemano qué observación mirar.

Quitar el barrio no es la conclusión. La distancia de Cook dice dónde mirar; qué hacer con lo que se encuentra depende de los datos: si el barrio 365 es un error de registro, se corrige o se quita; si es un barrio real con un precio bajo, se queda, y el informe dice que el resultado depende de él.

Actividad 3 - Un grupo influyente: los precios censurados en 50

En el Boston original el precio mediano está censurado: el valor máximo, 50, aparece 16 veces, y en el fichero significa “50 o más”. Esos 16 barrios tienen un precio real desconocido y un residuo del mismo signo en todos ellos. Compare su distancia de Cook con la de los demás y estime el modelo sin ellos.

`sum(medv = 50)` cuenta los barrios censurados, porque la comparación vale 1 cuando es cierta. Las dos medias de `cook` se calculan como la media de grupo de la lección 16 (“Una indicadora desplaza el intercepto: el río Charles”): la media del producto de `cook` por la indicadora del grupo, dividida por la media de la indicadora. El primer `smpl` deja solo los 16 barrios para imprimir el menor y el mayor de sus residuos, sus medias de `rm`, `lstat` y palanca, cuántos tienen palanca mayor que la media k/n y cuántos superan la convención $4/n$ de la distancia de Cook; `smpl medv <50 --restrict` los deja fuera.

```
printf "Barrios con medv = 50: %d\n", sum(medv == 50)
printf "Cook media de esos barrios: %.4f ; de los demas: %.4f\n", mean(cook * (medv == 50)) / mean(medv == 50), mean(cook *
↪ (medv < 50)) / mean(medv < 50)
```

```

smpl medv == 50 --restrict
printf "Residuos de esos barrios: minimo %.3f ; maximo %.3f\n", min(e), max(e)
printf "Medias de esos barrios: rm %.2f lstat %.2f palanca %.4f ; con palanca mayor que k/n: %d ; con Cook mayor que 4/n: %d\n", mean(rm), mean(lstat), mean(lever), sum(lever > k/506), sum(cook > 4/506)
smpl full
smpl medv < 50 --restrict
ols lmedv const rm lstat dis chas
printf "Sin los 16: chas %.4f (ee %.4f) rm %.4f lstat %.4f dis %.4f\n", $coeff(chas), $stderr(chas), $coeff(rm), $coeff(lstat), $coeff(dis)
smpl full

```

Barrios con medv = 50: 16
Cook media de esos barrios: 0,0161 ; de los demas: 0,0026
Residuos de esos barrios: minimo 0,099 ; maximo 0,757
Medias de esos barrios: rm 7,48 lstat 4,36 palanca 0,0279 ; con palanca mayor que k/n: 15 ; con Cook mayor que 4/n: 8

Modelo 1: MCO, usando las observaciones 1-490
Variable dependiente: lmedv

	coeficiente	Desv. típica	Estadístico t	valor p
const	2,79539	0,144073	19,40	1,72e-62 ***
rm	0,107222	0,0193339	5,546	4,82e-08 ***
lstat	-0,0369454	0,00204660	-18,05	4,23e-56 ***
dis	0,00286834	0,00563626	0,5089	0,6110
chas	0,123365	0,0422265	2,922	0,0036 ***

Media de la vble. dep.	3,005859	D.T. de la vble. dep.	0,382794
Suma de cuad. residuos	23,33702	D.T. de la regresión	0,219357
R-cuadrado	0,674309	R-cuadrado corregido	0,671623
F(4, 485)	251,0351	Valor p (de F)	1,2e-116
Log-verosimilitud	50,58940	Criterio de Akaike	-91,17880
Criterio de Schwarz	-70,20677	Crit. de Hannan-Quinn	-82,94233

Log-verosimilitud para medv = -1422,28

Sin considerar la constante, el valor p más alto fue el de la variable 8 (dis)

Sin los 16: chas 0,1234 (ee 0,0422) rm 0,1072 lstat -0,0369 dis 0,0029

Lo que debe observar

La distancia de Cook media de los 16 barrios censurados es 0,016, seis veces la de los demás, 0,003: como grupo, desplazan el ajuste. Sus residuos son todos positivos, entre 0,10 y 0,76 en logaritmos: el ajuste queda por debajo de 50 en los 16, y más por debajo aún del precio real, que es 50 o más. Sin ellos, el coeficiente de **chas** baja de 0,160 a 0,123, con error estándar 0,042, y el de **rm** de 0,120 a 0,107; el de **dis** cambia de signo, de -0,006 a +0,003, con $t = -1,10$ y $t = 0,51$. La diferencia entre barrios junto al río y sin río sigue siendo positiva y su t sigue por encima de 2, pero el 17 % de la sesión 25 pasa a un 13 % ($e^{0,123} - 1 = 0,13$).

Con los barrios censurados, a diferencia del barrio 365, la decisión sí está clara: el valor 50 no es el precio de esos barrios, sino una cota, y un modelo del logaritmo del precio no debería tratar esa cota como un precio. La estimación sin los 16 barrios es la que corresponde a los datos conocidos, y el informe dice que se han excluido y por qué. Lo que la distancia de Cook añade es que el grupo se habría señalado sin leer la documentación del fichero: la mitad de los 16 supera la convención $4/n$, y sus residuos son todos positivos.

Cierre - El diccionario geometría ↔ probabilidad

El curso ha presentado la misma construcción dos veces. Las lecciones 3 a 8 la presentaron con vectores de $\mathbb{R}^n$: la media es el coeficiente de la proyección sobre la constante, la regresión es una proyección sobre el subespacio de los regresores, el residuo es ortogonal a ese subespacio y el R^2 es el cuadrado de un coseno. Las lecciones 9 a 12 la presentaron con variables aleatorias: el valor esperado es el coeficiente de la proyección sobre

la constante, la esperanza condicional es una proyección sobre las funciones de los regresores, la perturbación es ortogonal a esas funciones. Las dos construcciones coinciden porque los dos espacios tienen un producto escalar y una proyección, y la tabla recoge la traducción, ampliando la tabla-diccionario con la que cerró la lección 9. Es una analogía explícita, que el apéndice de geometría desarrolla en sus secciones 1 a 4, desde el [apéndice \(sección 1\) \(html\)](#), y que aquí no se demuestra.

Geometría de los datos ($\mathbb{R}^n$)	Probabilidad (variables aleatorias)
vector de datos $\mathbf{x}$	variable aleatoria X , función sobre Ω
$\langle \mathbf{x}, \mathbf{y} \rangle_s = \frac{1}{n} \sum_i x_i y_i$	$\langle X, Y \rangle_P = E(XY)$
media $\mu_{\mathbf{x}}$ = coeficiente de la proyección sobre $\mathbf{1}$	valor esperado $E(X)$ = coeficiente de la proyección sobre $\mathbf{1}$
vector en desviaciones $\mathbf{x} - \mu_{\mathbf{x}} \mathbf{1}$	variable centrada $X - E(X)$
desviación típica = norma del vector en desviaciones	desviación típica = norma de la variable centrada
correlación = coseno del ángulo entre vectores en desviaciones	correlación = coseno del ángulo entre variables centradas
ajuste $\hat{\mathbf{y}} = \text{proyección sobre } \mathcal{L}(\mathbf{1}, \mathbf{x}_2, \dots)$	esperanza condicional $E[Y X]$ = proyección sobre las funciones de X
residuo ortogonal a los regresores	perturbación con $E[U X] = 0$
$R^2 = \cos^2$ del ángulo entre datos y ajuste, en desviaciones	$\text{Corr}(Y, E[Y X])^2$
la muestra: un punto de $\mathbb{R}^n$	una columna de la matriz de copias; la muestra aleatoria son sus n filas

La columna de la izquierda es lo que Gretl calcula; la de la derecha es lo que se quiere conocer. El puente entre las dos es el muestreo aleatorio de la lección 11: los datos son una columna de la matriz de copias, y cada fórmula de la izquierda, evaluada sobre la muestra aleatoria, es un estadístico que estima la cantidad de la derecha.¹

Preguntas de interpretación para la clase

1. Las palancas suman 5 y su media es 0,0099. ¿Por qué suman el número de columnas de $\mathbf{X}$ y no el de filas?
2. El barrio 366 tiene la palanca máxima y el 365 la distancia de Cook máxima. ¿Por qué no coinciden?
3. ¿Qué cambia al quitar el barrio 365 y qué se concluye de ello? ¿Es una razón para quitarlo?
4. Los 16 barrios con `medv` igual a 50 tienen una distancia de Cook media seis veces mayor que los demás. ¿Qué tienen en común que explique esa influencia?
5. En la tabla del diccionario, la media de los datos y el valor esperado de una variable aleatoria ocupan la misma fila. ¿En qué sentido son el mismo objeto y en qué sentido no lo son?

Para profundizar

- Wooldridge, J. M. (2020), *Introductory Econometrics*, cap. 9, sección 9.5 (observaciones influyentes).
- Belsley, D. A., Kuh, E. y Welsch, R. E. (1980), *Regression Diagnostics*, Wiley: la palanca, DFFITS y las convenciones $2k/n$ y $2\sqrt{k/n}$.
- Cook, R. D. (1977), “Detection of influential observation in linear regression”, *Technometrics* 19, 15–18.

¹En la fila del R^2 , $\text{Corr}(Y, E[Y | X])$ es la correlación entre Y y su esperanza condicional, definida en la lección 9 con el mismo convenio que E y Var : $\text{Cov}(Y, E[Y | X]) / \sqrt{\text{Var}(Y) \text{Var}(E[Y | X])}$. Su cuadrado es $\text{Var}(E[Y | X]) / \text{Var}(Y)$, la fracción de la varianza de Y que la descomposición $\text{Var}(Y) = \text{Var}(E[Y | X]) + \text{Var}(U)$ de la lección 9 atribuye a la esperanza condicional, porque $\text{Cov}(Y, E[Y | X]) = \text{Var}(E[Y | X])$: U es ortogonal a las funciones de X y tiene esperanza cero.

- Apéndice de geometría del curso, secciones 1 a 4: el espacio de las variables aleatorias, las copias idénticas e independientes, la matriz de copias y el puente entre los dos espacios.

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"

set verbose off

open ../data/boston.csv --quiet
series lmedv = log(medv)
outfile --quiet palanca.txt
    ols lmedv const rm lstat dis chas --quiet
    scalar k = 5
    leverage --quiet --save
    summary lever --simple
    printf "Suma de las palancas: %.2f ; media k/n = %.4f ; mediana %.4f ; convencion 2k/n = %.4f ; barrios
↪ por encima: %d\n", sum(lever), k/$nobs, median(lever), 2*k/$nobs, sum(lever > 2*k/$nobs)
    scalar i = imaxc({lever})
    printf "Palanca maxima %.4f en el barrio %d: medv %.1f rm %.2f lstat %.2f dis %.2f chas %d\n",
↪ lever[i], i, medv[i], rm[i], lstat[i], dis[i], chas[i]
end outfile

outfile --quiet cook.txt
    series e = $uhat
    scalar s2 = $sigma^2
    series cook = e^2 * lever / (k * s2 * (1 - lever)^2)
    printf "Convencion 4/n = %.4f ; barrios por encima: %d\n", 4/$nobs, sum(cook > 4/$nobs)
    scalar j = imaxc({cook})
    printf "Cook maxima %.3f en el barrio %d: medv %.1f rm %.2f lstat %.2f dis %.2f chas %d palanca
↪ %.4f residuo %.3f\n", cook[j], j, medv[j], rm[j], lstat[j], dis[j], chas[j], lever[j], e[j]
    printf "Barrio %d, el de palanca maxima: residuo %.3f Cook %.4f ; barrios con Cook mayor: %d\n", i,
↪ e[i], cook[i], sum(cook > cook[i])
    printf "Con todos:      chas %.4f rm %.4f lstat %.4f dis %.4f\n", $coeff(chas), $coeff(rm),
↪ $coeff(lstat), $coeff(dis)
    smpl cook < max(cook) --restrict
    ols lmedv const rm lstat dis chas --quiet
    printf "Sin el barrio %d: chas %.4f rm %.4f lstat %.4f dis %.4f (n = %d)\n", j, $coeff(chas),
↪ $coeff(rm), $coeff(lstat), $coeff(dis), $nobs
    smpl full
end outfile

outfile --quiet censura.txt
    printf "Barrios con medv = 50: %d\n", sum(medv == 50)
    printf "Cook media de esos barrios: %.4f ; de los demas: %.4f\n", mean(cook * (medv == 50)) / mean(medv
↪ == 50), mean(cook * (medv < 50)) / mean(medv < 50)
    smpl medv == 50 --restrict
    printf "Residuos de esos barrios: minimo %.3f ; maximo %.3f\n", min(e), max(e)
    printf "Medias de esos barrios: rm %.2f lstat %.2f palanca %.4f ; con palanca mayor que k/n: %d ; con
↪ Cook mayor que 4/n: %d\n", mean(rm), mean(lstat), mean(lever), sum(lever > k/506), sum(cook >
↪ 4/506)
```

```
smpl full
smpl medv < 50 --restrict
ols lmedv const rm lstat dis chas
printf "Sin los 16: chas %.4f (ee %.4f) rm %.4f lstat %.4f dis %.4f\n", $coeff(chas), $stderr(chas),
↪ $coeff(rm), $coeff(lstat), $coeff(dis)
smpl full
end outfile
```

Respuestas

1. La palanca h_i es la longitud al cuadrado de la proyección de $\mathbf{e}_i$ sobre el subespacio de los regresores. Los n vectores canónicos forman una base ortonormal de $\mathbb{R}^n$, y la suma de las longitudes al cuadrado de sus proyecciones sobre un subespacio es la dimensión de ese subespacio, k , porque con una base ortonormal $\mathbf{b}_1, \dots, \mathbf{b}_k$ del subespacio,

$$\sum_{i=1}^n \|P\mathbf{e}_i\|_e^2 = \sum_{i=1}^n \sum_{j=1}^k \langle \mathbf{e}_i, \mathbf{b}_j \rangle_e^2 = \sum_{j=1}^k \sum_{i=1}^n b_{ji}^2 = \sum_{j=1}^k \|\mathbf{b}_j\|_e^2 = k;$$

esa suma es la traza de la matriz de proyección.² Si el subespacio fuese todo $\mathbb{R}^n$, cada palanca valdría 1 y el ajuste reproduciría los datos.

2. La palanca solo depende de los regresores, y la distancia de Cook multiplica la palanca por el residuo al cuadrado. El barrio 366 está lejos de los demás en **rm**, y su residuo, 0,39, no es pequeño: su distancia de Cook es 0,051, la séptima del fichero. El 365 tiene menos palanca, 0,054 frente a 0,076, pero su residuo, -0,70, es casi el doble en valor absoluto y, al cuadrado, más del triple: por eso su distancia de Cook, 0,113, es la mayor.
3. El coeficiente de **chas** pasa de 0,160 a 0,179: la diferencia entre barrios junto al río y sin río depende en un 12 % de un solo barrio. Esa dependencia no es una razón para quitarlo. La distancia de Cook señala qué observación mirar; quitarla exige una razón sobre los datos, como un error de registro. Si el barrio es real, se queda, y el informe dice de qué depende el resultado.
4. Su precio registrado, 50, es una cota inferior del precio real, y sus residuos tienen todos el mismo signo: son positivos, el ajuste queda por debajo de 50 en los 16, y más por debajo aún del precio verdadero. Que los 16 queden por debajo es sistemático, no azar. Además son barrios con **rm** alto y **lstat** bajo, con medias de 7,48 habitaciones y un 4,36 % de población de estatus bajo, y 15 de los 16 tienen palanca mayor que la media k/n . Un grupo de observaciones puede ser influyente aunque ninguna lo sea mucho por separado.
5. Los dos son el coeficiente de la proyección de un objeto sobre la constante, en un espacio con producto escalar: $\mu_{\mathbf{x}} = \langle \mathbf{x}, \mathbf{1} \rangle_s$ en $\mathbb{R}^n$ y $E(X) = \langle X, 1 \rangle_P$ en el espacio de las variables aleatorias. No son el mismo objeto: la media es un número calculado con los datos observados, y el valor esperado es un número de la población, que no se observa. La media muestral, la fórmula de la media evaluada sobre la muestra aleatoria, es la variable aleatoria que estima el valor esperado; que lo estime sin sesgo es la lección 11.

²La matriz de proyección sobre el subespacio de las columnas de $\mathbf{X}$ es $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$: multiplicada por cualquier vector de $\mathbb{R}^n$ devuelve su proyección sobre ese subespacio. En particular $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, porque $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ (lección 10). Es simétrica y $\mathbf{P}\mathbf{P} = \mathbf{P}$, así que su elemento (i, i) es $\mathbf{e}_i^\top \mathbf{P} \mathbf{e}_i = \|\mathbf{P} \mathbf{e}_i\|_e^2 = h_i$, la palanca; y su traza, la suma de los elementos de su diagonal, es $\sum_i h_i = k$.