

Sesión 11 (B) — Regresión simple con datos reales: ajuste y bondad de ajuste en hprice2

Marcos Bujosa

Descripción de la práctica

Retomamos `hprice2` (ya usado en las prácticas de correlación) y sus tres variables etiquetadas: `price`, `rooms`, `nox`. Hoy estimamos, por primera vez con datos *reales*, dos regresiones lineales simples: precio en función del número de habitaciones, y precio en función de la contaminación. Las comparamos entre sí, verificamos las dos condiciones de ortogonalidad y calculamos la bondad de ajuste, conectando con las correlaciones ya calculadas en la práctica de correlación (lección 5) y con el ángulo θ de la lección 8.

Objetivo

1. Estimar dos regresiones lineales simples con datos reales, interpretando el signo y la magnitud de la pendiente sin lenguaje causal.
2. Verificar, con datos reales, las dos condiciones de ortogonalidad de la lección 7.
3. Calcular R^2 “a mano” (STC/SEC/SRC) y compararlo con el que reporta Gretl y con ρ_{xy}^2 .
4. Comparar la bondad de ajuste de los dos modelos y relacionarla con el ángulo θ de la lección 8.
5. Visualizar ambos ajustes mediante diagramas de dispersión con la recta ajustada.
6. Intercambiar los papeles de las dos variables y comprobar que el R^2 no cambia —la advertencia de la lección 8—, aunque la recta sí.

Actividad 1 - Reapertura de hprice2 y etiquetas

en línea de comandos:

```
open hprice2.gdt --quiet
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)"
```

Actividad 2 - Regresión: price sobre rooms

en línea de comandos:

```
ols price const rooms
```

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

- GUI equivalente: Modelo -> Mínimos cuadrados ordinarios, con price como dependiente y const, rooms como regresores.

Modelo 1: MCO, usando las observaciones 1-506
Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p
const	-34796,2	2651,53	-13,12	4,74e-34 ***
rooms	9119,55	419,339	21,75	1,87e-74 ***
Media de la vble. dep.	22511,51	D.T. de la vble. dep.	9208,856	
Suma de cuad. residuos	2,21e+10	D.T. de la regresión	6620,865	
R-cuadrado	0,484110	R-cuadrado corregido	0,483086	
F(1, 504)	472,9521	Valor p (de F)	1,87e-74	
Log-verosimilitud	-5168,759	Criterio de Akaike	10341,52	
Criterio de Schwarz	10349,97	Crit. de Hannan-Quinn	10344,83	

Lo que debe observar

La pendiente es 9119,55: en el ajuste, a una habitación más le corresponden unos 9120 dólares más de precio mediano. La constante, -34796, no tiene lectura por sí sola (sería el precio de una vivienda con cero habitaciones, fuera del rango de los datos). Y $R^2 = 0,484$: el ajuste explica algo menos de la mitad de la variación de price. Como en la práctica anterior, ignore de momento las columnas Desv. típica, Estadístico t y valor p.

Interpretación (con cuidado): el coeficiente de rooms se lee como “por cada habitación adicional, el precio mediano cambia, en promedio, en esa cantidad de dólares, *manteniendo el resto igual* (aquí, sin ningún otro regresor)”. Evite decir que “una habitación adicional *causa*” ese cambio: como se discutió en la lección 8, la descomposición $STC = SEC + SRC$ es puramente algebraica y no certifica, por sí sola, ninguna relación causal ni su dirección.

Actividad 3 - Regresión: price sobre nox

en línea de comandos:

```
ols price const nox
```

Antes de mirar el resultado: ¿qué signo espera para el coeficiente de nox? Contraste su expectativa con lo obtenido.

Modelo 2: MCO, usando las observaciones 1-506
Variable dependiente: price

	coeficiente	Desv. típica	Estadístico t	valor p
const	41307,8	1816,18	22,74	2,53e-79 ***
nox	-3386,85	320,362	-10,57	9,93e-24 ***
Media de la vble. dep.	22511,51	D.T. de la vble. dep.	9208,856	
Suma de cuad. residuos	3,51e+10	D.T. de la regresión	8339,566	
R-cuadrado	0,181508	R-cuadrado corregido	0,179884	
F(1, 504)	111,7662	Valor p (de F)	9,93e-24	
Log-verosimilitud	-5285,537	Criterio de Akaike	10575,07	
Criterio de Schwarz	10583,53	Crit. de Hannan-Quinn	10578,39	

Lo que debe observar

La pendiente es negativa, $-3386,85$: a una unidad más de contaminación le corresponden unos 3400 dólares menos de precio mediano, como cabía esperar. El R^2 , 0,182, es bastante menor que el de **rooms**: la contaminación, por sí sola, explica menos de una quinta parte de la variación de los precios. Compare ambos R^2 con las correlaciones de la práctica de correlación: $0,6958^2 = 0,484$ y $(-0,4260)^2 = 0,182$.

Actividad 4 - Verificación de las dos condiciones de ortogonalidad

Repetimos ambas regresiones, guardando en cada caso los residuos, para verificar $\sum \hat{e}_i = 0$ y $\sum x_i \hat{e}_i = 0$.
en línea de comandos:

```
ols price const rooms --quiet
series ehat_rooms = $uhat
scalar sum_e_rooms = sum(ehat_rooms)
scalar sum_xe_rooms = sum(rooms*ehat_rooms)

ols price const nox --quiet
series ehat_nox = $uhat
scalar sum_e_nox = sum(ehat_nox)
scalar sum_xe_nox = sum(nox*ehat_nox)

printf "Modelo price-rooms: suma(ehat) = %.8f    suma(rooms*ehat) = %.8f\n", sum_e_rooms, sum_xe_rooms
printf "Modelo price-nox : suma(ehat) = %.8f    suma(nox*ehat) = %.8f\n", sum_e_nox, sum_xe_nox
```

```
Modelo price-rooms: suma(ehat) = -0,00000001    suma(rooms*ehat) = -0,00000019
Modelo price-nox : suma(ehat) = -0,00000001    suma(nox*ehat) = 0,00000005
```

(Los valores no serán cero por errores de redondeo en coma flotante, pero deben ser insignificantes.)

Actividad 5 - Bondad de ajuste: STC, SEC, SRC y R^2 en ambos modelos

en línea de comandos:

```
ols price const rooms --quiet
scalar mu_price = mean(price)
scalar STC_r = sum((price-mu_price)^2)
scalar SRC_r = sum($uhat^2)
scalar SEC_r = STC_r - SRC_r
scalar R2_r_manual = SEC_r/STC_r

ols price const nox --quiet
scalar STC_n = sum((price-mu_price)^2)
scalar SRC_n = sum($uhat^2)
scalar SEC_n = STC_n - SRC_n
scalar R2_n_manual = SEC_n/STC_n

printf "price-rooms: STC=%.2f SEC=%.2f SRC=%.2f R2(a mano)=%.4f\n", STC_r, SEC_r, SRC_r, R2_r_manual
printf "price-nox : STC=%.2f SEC=%.2f SRC=%.2f R2(a mano)=%.4f\n", STC_n, SEC_n, SRC_n, R2_n_manual

printf "\nComparacion con rho^2 (recuerde: en regresion simple R2 = rho_xy^2):\n"
printf "corr(price,rooms)^2 = %.4f\n", corr(price,rooms)^2
printf "corr(price,nox)^2 = %.4f\n", corr(price,nox)^2
```

```
price-rooms: STC=42825531146,45 SEC=20732259424,69 SRC=22093271721,76 R2(a mano)=0,4841
price-nox : STC=42825531146,45 SEC=7773157316,47 SRC=35052373829,98 R2(a mano)=0,1815
```

```
Comparacion con rho^2 (recuerde: en regresion simple R2 = rho_xy^2):
corr(price,rooms)^2 = 0,4841
corr(price,nox)^2 = 0,1815
```

El ángulo θ , en grados

El R^2 que acabamos de calcular *es* un coseno al cuadrado (lección 8). Nada nos impide, por tanto, deshacer el coseno y ver de cuántos grados estamos hablando:

en línea de comandos:

```
printf "price-rooms: R2 = %.4f  ->  theta = %.2f grados\n", R2_r_manual, acos(sqrt(R2_r_manual))*180/$pi
printf "price-nox  : R2 = %.4f  ->  theta = %.2f grados\n", R2_n_manual,  acos(sqrt(R2_n_manual))*180/$pi
```

```
price-rooms: R2 = 0,4841  ->  theta = 45,91 grados
price-nox  : R2 = 0,1815  ->  theta = 64,78 grados
```

Son los ángulos que forman, en $\mathbb{R}^{506}$, el vector de datos en desviaciones y el del ajuste. Un R^2 que parece modesto corresponde a un ángulo bastante abierto: conviene tener presente esta equivalencia al leer un R^2 , porque los dos números dicen lo mismo con escalas muy distintas.

Actividad 6 - Diagramas de dispersión con la recta ajustada

- GUI: seleccione **price** y **rooms** (con Ctrl), botón derecho, **Gráfico X-Y**; repita con **price** y **nox**.

o bien teclee en línea de comandos:

```
gnuplot price rooms
gnuplot price nox
```

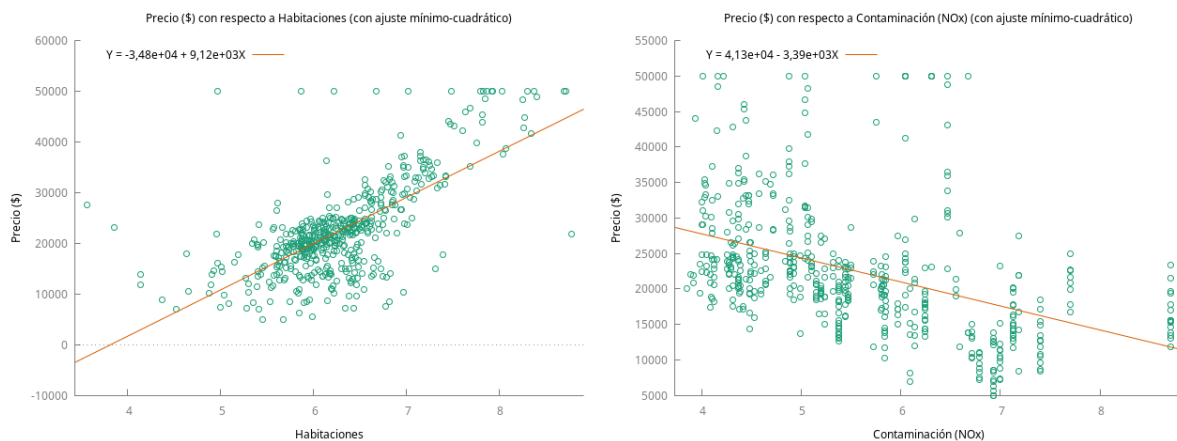


Figura 1: Diagramas de dispersión con la recta ajustada.

Compare cómo de “ceñidos” están los puntos que rodean cada recta con el R^2 obtenido en la Actividad 5. Recuerde (lección 8) que $R^2 = \cos^2 \theta$, con θ el ángulo entre los vectores en desviaciones de los datos y del ajuste: cuanto más ceñida la nube a la recta, más pequeño ese ángulo.

Actividad 7 - Intercambiar los papeles: el R^2 no distingue causa de efecto

La lección 8 advirtió que, si regresamos la superficie sobre el precio en lugar del precio sobre la superficie, el R^2 *no cambia*. Comprobémoslo con **price** y **rooms**: estimamos las dos regresiones, la habitual y la “del revés”.

en línea de comandos:

```
ols price const rooms --quiet
scalar b_pr = $coeff(rooms)
scalar r2_pr = $rsq

ols rooms const price --quiet      # los papeles, intercambiados
scalar b_rp = $coeff(price)
scalar r2_rp = $rsq

printf "R2 de price-rooms = %.6f      R2 de rooms-price = %.6f\n", r2_pr, r2_rp
printf "pendiente de price-rooms = %.4f      pendiente de rooms-price = %.8f\n", b_pr, b_rp
printf "inversa de la primera pendiente = %.8f      (no es la segunda)\n", 1/b_pr
printf "producto de las dos pendientes = %.6f      (= R2 = rho^2)\n", b_pr*b_rp
```

```
R2 de price-rooms = 0,484110      R2 de rooms-price = 0,484110
pendiente de price-rooms = 9119,5484      pendiente de rooms-price = 0,00005308
inversa de la primera pendiente = 0,00010965      (no es la segunda)
producto de las dos pendientes = 0,484110      (= R2 = rho^2)
```

Lo que debe observar

Los dos R^2 son el mismo número: en regresión simple $R^2 = \rho_{xy}^2$, y la correlación no sabe cuál variable hace de x y cuál de y . El ajuste, en cambio, sí cambia: la pendiente de **rooms-price** no es la inversa de la de **price-rooms** (las dos rectas son distintas; solo coinciden si $\rho = \pm 1$). El producto de las dos pendientes es $\frac{\sigma_{xy}}{\sigma_x^2} \cdot \frac{\sigma_{xy}}{\sigma_y^2} = \rho_{xy}^2$, que es otra vez el R^2 . La moraleja es la de la lección 8: el R^2 mide cuánto se alinean dos vectores en desviaciones, y esa alineación es simétrica; de ella no se sigue ninguna dirección causal.

Actividad 8 - Síntesis comparativa

Modelo	$\hat{\beta}_1$	$\hat{\beta}_2$	R^2
price~rooms			
price~nox			

Complete la tabla con los valores obtenidos. ¿Cuál de los dos modelos ajusta mejor, según R^2 ? ¿Es coherente con lo observado en los diagramas de dispersión y con las correlaciones ya calculadas en la práctica de correlación (lección 5)?

Preguntas de interpretación para la clase

1. ¿Por qué no podemos afirmar, a partir del modelo `price~rooms`, que “añadir una habitación causa” un aumento de precio?
2. Compare los dos modelos: ¿cuál tiene mayor R^2 ? ¿Qué relación tiene esto con los valores de correlación ya calculados en la práctica de correlación?
3. En los diagramas de dispersión con la recta ajustada, ¿qué modelo muestra una nube de puntos más “pegada” a la recta? ¿Cómo se traduce esto en términos del ángulo θ de la lección 8?
4. ¿Por qué la suma de los residuos no es cero en la salida de Gretl, sino un número muy pequeño (del orden de 10^{-8} o menor)?
5. Si estimáramos **price** en función de **rooms** y **nox** *simultáneamente* (dos regresores a la vez), ¿por qué NO cabría esperar que el R^2 de ese modelo fuera simplemente la suma de los R^2 de los dos modelos simples de hoy?

6. En la Actividad 7 las dos regresiones tienen el mismo R^2 pero distinta recta. ¿Por qué las dos pendientes no son inversas una de otra? ¿Cuándo lo serían?

Para profundizar:

- Véanse las lecciones 7 y 8 ([lección 7](#), [lección 8](#)) para las demostraciones completas de lo que hoy verificamos con datos reales.
- Wooldridge, J. M. (2020). *Introductory Econometrics*, cap. 2, sección 2.4 (advertencias sobre la interpretación causal de R^2).

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 --quiet
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)"

outfile --quiet olsPriceRooms.txt
  ols price const rooms
end outfile

outfile --quiet olsPriceNox.txt
  ols price const nox
end outfile

outfile --quiet ortogonalidadAmbosModelos.txt
  ols price const rooms --quiet
  series ehat_rooms = $uhat
  scalar sum_e_rooms = sum(ehat_rooms)
  scalar sum_xe_rooms = sum(rooms*ehat_rooms)

  ols price const nox --quiet
  series ehat_nox = $uhat
  scalar sum_e_nox = sum(ehat_nox)
  scalar sum_xe_nox = sum(nox*ehat_nox)

  printf "Modelo price-rooms: suma(ehat) = %.8f   suma(rooms*ehat) = %.8f\n", sum_e_rooms, sum_xe_rooms
  printf "Modelo price-nox : suma(ehat) = %.8f   suma(nox*ehat) = %.8f\n", sum_e_nox, sum_xe_nox
end outfile

outfile --quiet bondadDeAjusteAmbosModelos.txt
  ols price const rooms --quiet
  scalar mu_price = mean(price)
  scalar STC_r = sum((price-mu_price)^2)
  scalar SRC_r = sum($uhat^2)
  scalar SEC_r = STC_r - SRC_r
  scalar R2_r_manual = SEC_r/STC_r

  ols price const nox --quiet
  scalar STC_n = sum((price-mu_price)^2)
  scalar SRC_n = sum($uhat^2)
  scalar SEC_n = STC_n - SRC_n
  scalar R2_n_manual = SEC_n/STC_n

  printf "price~rooms: STC=%.2f SEC=%.2f SRC=%.2f R2(a mano)=%.4f\n", STC_r, SEC_r, SRC_r,
  ↳ R2_r_manual
```

```

printf "price-nox : STC=%.2f SEC=%.2f SRC=%.2f R2(a mano)=%.4f\n", STC_n, SEC_n, SRC_n,
↪ R2_n_manual

printf "\nComparacion con rho^2 (recuerde: en regresion simple R2 = rho_xy^2):\n"
printf "corr(price,rooms)^2 = %.4f\n", corr(price,rooms)^2
printf "corr(price,nox)^2 = %.4f\n", corr(price,nox)^2
end outfile

outfile --quiet anguloTheta.txt
printf "price-rooms: R2 = %.4f -> theta = %.2f grados\n", R2_r_manual,
↪ acos(sqrt(R2_r_manual))*180/$pi
printf "price-nox : R2 = %.4f -> theta = %.2f grados\n", R2_n_manual,
↪ acos(sqrt(R2_n_manual))*180/$pi
end outfile

gnuplot price rooms --output="price_rooms_fit.png"
gnuplot price nox --output="price_nox_fit.png"

outfile --quiet intercambiarPapeles.txt
ols price const rooms --quiet
scalar b_pr = $coeff(rooms)
scalar r2_pr = $rsq

ols rooms const price --quiet # los papeles, intercambiados
scalar b_rp = $coeff(price)
scalar r2_rp = $rsq

printf "R2 de price-rooms = %.6f R2 de rooms~price = %.6f\n", r2_pr, r2_rp
printf "pendiente de price-rooms = %.4f pendiente de rooms~price = %.8f\n", b_pr, b_rp
printf "inversa de la primera pendiente = %.8f (no es la segunda)\n", 1/b_pr
printf "producto de las dos pendientes = %.6f (= R2 = rho^2)\n", b_pr*b_rp
end outfile

```

Enlace al guión: [S11-Prct-B-hprice2-ajuste.inp](#)

Respuestas

1. ¿Por qué no podemos afirmar causalidad a partir de price~rooms? Porque la descomposición $STC = SEC + SRC$ y el propio ajuste MCO son consecuencias *puramente geométricas* de la ortogonalidad del residuo: se cumplen igual para cualquier par de vectores $\mathbf{x}, \mathbf{y}$, tengan o no una relación causal real entre sí (recuerde, de la lección 8, el ejemplo del queso y las muertes por enredarse en sábanas). Afirmar causalidad exige una teoría económica previa que justifique la dirección de la relación, algo que la sola regresión no proporciona.
2. Comparación de R^2 entre los dos modelos. El modelo con mayor $|\rho_{\mathbf{x}\mathbf{y}}|$ (en valor absoluto) entre la variable y el regresor tendrá, necesariamente, mayor R^2 , porque en regresión simple $R^2 = \rho_{\mathbf{x}\mathbf{y}}^2$ exactamente (lección 8). Por tanto, el orden de los R^2 de hoy debe coincidir con el orden, en valor absoluto, de las correlaciones $\rho_{\text{price,rooms}}$ y $\rho_{\text{price,nox}}$ ya calculadas en la práctica de correlación.
3. Nube más pegada a la recta y ángulo θ . El modelo con mayor R^2 es el que tiene la nube de puntos más ceñida a la recta ajustada, porque $R^2 = \cos^2 \theta$: cuanto menor el ángulo θ entre los vectores en desviaciones de los datos y del ajuste, más cerca de 1 está $\cos^2 \theta$, y más “alineados” (en el sentido geométrico de la lección 8) están esos dos vectores —lo que visualmente se traduce en una nube más estrecha alrededor de la recta.
4. Suma de residuos no cero. Por errores de redondeo en aritmética de coma flotante: los ordenadores no representan los números reales con precisión infinita, así que operaciones que algebraicamente deberían dar cero (como $\sum \hat{e}_i$) arrastran un pequeñísimo error de cálculo, típicamente del orden de 10^{-8} o menor. Esto no contradice la identidad teórica $\sum \hat{e}_i = 0$: es la misma situación que ya se comprobó en la práctica de verificación geométrica (con datos simulados).
5. R^2 de un modelo con dos regresores no es la suma de los R^2 simples. Porque, como se explicó al cierre de la lección 8, cuando hay más de un regresor no constante el subespacio $\mathcal{L}(\mathbf{1}, \mathbf{x}_2, \mathbf{x}_3)$ centrado tiene dimensión 2, no 1: los vectores en desviaciones de **rooms** y **nox** ya no están alineados en la misma recta, así que el ajuste conjunto $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ ya no es, en general, la suma de las contribuciones “aisladas” de cada regresor. Además, **rooms** y **nox** pueden estar correlacionadas entre sí, de modo que parte de lo que “explica” cada una por separado se solapa. Este fenómeno se estudiará con detalle en la regresión múltiple, más adelante en el curso.
6. *Dos rectas, un R^2* . La pendiente de **price~rooms** es $\sigma_{\mathbf{x}\mathbf{y}}/\sigma_{\mathbf{x}}^2$ y la de **rooms~price** es $\sigma_{\mathbf{x}\mathbf{y}}/\sigma_{\mathbf{y}}^2$: cada una proyecta un vector distinto sobre un plano distinto, y no hay razón para que una sea la inversa de la otra. Su producto vale $\rho_{\mathbf{x}\mathbf{y}}^2$, así que solo serían inversas ($b_{pr} b_{rp} = 1$) si $\rho = \pm 1$, es decir, si los puntos estuvieran exactamente sobre una recta. Como el R^2 es ρ^2 , no cambia al intercambiar los papeles; la recta sí, porque minimiza distancias *verticales*, y “vertical” depende de qué variable pongamos en el eje de ordenadas.