

Sesión 7 (A) — Correlación como coseno: verificación geométrica con hprice2

Marcos Bujosa

Descripción de la práctica

En las lecciones 4 y 5 vimos que la media se obtiene mediante una proyección ortogonal, la desviación típica es una norma, la covarianza es un producto escalar de vectores centrados (es decir, vectores en desviaciones respecto a su media) y —el resultado central de la lección 5— que la correlación ρ_{xy} es el *coseno del ángulo* entre los vectores en desviaciones. Todo eso se demostró con álgebra. Hoy lo *verificamos con datos reales*: calcularemos ρ con el comando de Gretl y, por otro lado, “a mano”, construyendo el producto escalar y las normas con instrucciones de Gretl, para comprobar que ambos caminos dan el mismo número.

Retomamos **hprice2** (práctica de la sesión con Gretl anterior) y las mismas tres variables ya etiquetadas: **price**, **rooms**, **nox**.

Objetivo

1. Repasar estadísticos descriptivos de varias variables a la vez.
2. Calcular covarianzas y correlaciones entre pares de variables.
3. Visualizar la relación mediante diagramas de dispersión.
4. Verificar numéricamente que $\rho = \cos \theta$ entre los vectores *centrados* (es decir, en desviaciones).
5. Interpretar el signo y la magnitud de ρ en términos del ángulo entre vectores en desviaciones.

Actividad 1 - Recordatorio: datos y estadísticos descriptivos

Abrimos de nuevo **hprice2** y recuperamos las etiquetas de la práctica anterior.

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)"
```

- Recuerde los estadísticos descriptivos:

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

Actividad 2 - Covarianzas y correlaciones

- En la ventana principal de Gretl, marque **price**, **rooms** y **nox** (con Ctrl), “pinche” con el botón derecho y seleccione **Matriz de correlación**.

o bien teclee en línea de comandos:

```
corr price rooms nox
```

Antes de mirar el resultado: pensando en términos económicos, ¿qué signo esperaría en cada uno de los tres pares?

Resultado

Debe obtener una salida como esta:

Coefficientes de correlación, usando las observaciones 1 - 506
 Valor crítico al 5% (a dos colas) = 0,0872 para n = 506

```

price      rooms      nox
1,0000      0,6958     -0,4260 price
           1,0000     -0,3028 rooms
                   1,0000 nox
```

- Calculemos también, por separado, la covarianza y la correlación de cada par de variables (funciones **cov** y **corr** de Gretl)¹:

en línea de comandos:

```

printf "Cov(price,rooms) = %.4f    Corr(price,rooms) = %.4f\n", cov(price,rooms), corr(price,rooms)
printf "Cov(price,nox)   = %.4f    Corr(price,nox)   = %.4f\n", cov(price,nox), corr(price,nox)
printf "Cov(rooms,nox)   = %.4f    Corr(rooms,nox)   = %.4f\n", cov(rooms,nox), corr(rooms,nox)
```

```

Cov(price,rooms) = 4501,7557    Corr(price,rooms) = 0,6958
Cov(price,nox)   = -4544,7477   Corr(price,nox)   = -0,4260
Cov(rooms,nox)   = -0,2465      Corr(rooms,nox)   = -0,3028
```

¹Fíjese que **corr(x,y)** de Gretl es la implementación numérica del mismo coeficiente de correlación ρ_{xy} definido en la lección 5; aunque **cov(x,y)** no es la implementación exacta de la covarianza σ_{xy} pues Gretl divide internamente por $n - 1$ (en lugar de por n).

Actividad 3 - Diagramas de dispersión

- Marque dos variables a la vez (p. ej. price y rooms), “pinche” con el botón derecho y seleccione **Gráfico X-Y**.

o bien teclee en línea de comandos:

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

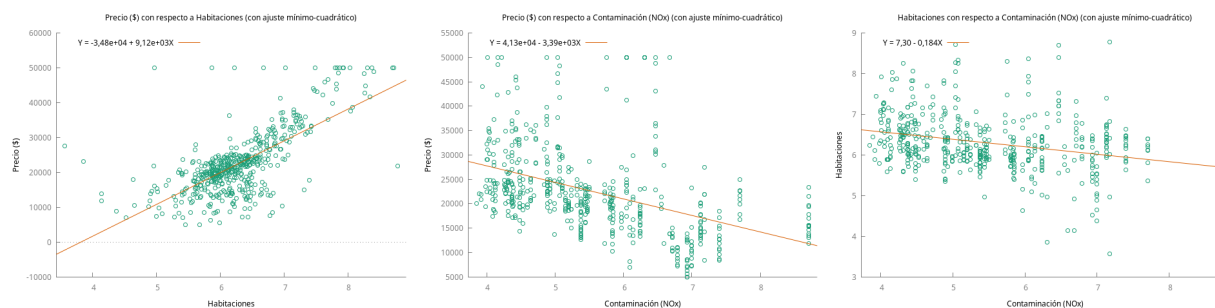


Figura 1: Diagramas de dispersión.

Compare la “forma” de cada nube de puntos (más o menos alargada y “alineada” alrededor de la recta mostrada en cada figura²) con el valor numérico de ρ obtenido en la Actividad 2.

Actividad 4 - Verificación geométrica: la correlación como coseno

Vamos a reconstruir, con instrucciones de Gretl, exactamente lo que hicimos con lápiz y papel en la lección 5: centrar los vectores (i.e., restarles su media), comprobar que son ortogonales a **1**, calcular su producto escalar y sus normas, y formar el coseno del ángulo.

Vectores centrados

en línea de comandos:

```
series price_c = price - mean(price)
series rooms_c = rooms - mean(rooms)
series nox_c = nox - mean(nox)
```

Comprobación de ortogonalidad con 1

La media de cada vector centrado debe ser (esencialmente) cero, es decir, sus valores deben sumar cero (salvo por errores de redondeo). Por tanto, “cada vector centrado debe ser ortogonal a **1**”:

en línea de comandos:

```
printf "Suma price_c = %.10f\n", sum(price_c)
printf "Suma rooms_c = %.10f\n", sum(rooms_c)
printf "Suma nox_c = %.10f\n", sum(nox_c)
```

²que, como veremos en la próxima lección, se llama “recta de regresión”

```
Suma price_c = 0,0000000152
Suma rooms_c = 0,0000000000
Suma nox_c   = 0,0000000000
```

(Los valores no serán cero por errores de redondeo en coma flotante, pero deben ser insignificantes; del orden de 10^{-8} o menores.)

El coseno del ángulo entre los datos en desviaciones (par price-rooms)

Producto escalar y normas “a mano” (sin dividir por ningún n , es decir, con el producto escalar y la norma *euclídeos*):

en línea de comandos:

```
scalar dp_pr      = sum(price_c * rooms_c)
scalar norma_price = sqrt(sum(price_c^2))
scalar norma_rooms = sqrt(sum(rooms_c^2))
scalar coseno_pr   = dp_pr / (norma_price * norma_rooms)

scalar rad2deg = 180/3.14159265358979
scalar theta_pr = acos(coseno_pr) * rad2deg

printf "cos(theta) price-rooms (a mano) = %.6f\n", coseno_pr
printf "corr(price,rooms) de Gretl      = %.6f\n", corr(price,rooms)
printf "Ángulo price-rooms = %.2f grados\n", theta_pr
```

Compruebe que `coseno_pr` y `corr(price,rooms)` coinciden.

```
cos(theta) price-rooms (a mano) = 0,695780
corr(price,rooms) de Gretl      = 0,695780
Ángulo price-rooms = 45,91 grados
```

(Hemos usado `acos`, la función arcocoseno de Gretl, que devuelve el ángulo en radianes; por eso multiplicamos por $\frac{180}{\pi}$ para expresarlo en grados.)

Repetimos para los otros dos pares

en línea de comandos:

```
scalar norma_nox = sqrt(sum(nox_c^2))

scalar dp_pn = sum(price_c * nox_c)
scalar coseno_pn = dp_pn / (norma_price * norma_nox)
scalar theta_pn = acos(coseno_pn) * rad2deg
printf "cos(theta) price-nox = %.6f (ángulo = %.2f grados)\n", coseno_pn, theta_pn

scalar dp_rn = sum(rooms_c * nox_c)
scalar coseno_rn = dp_rn / (norma_rooms * norma_nox)
scalar theta_rn = acos(coseno_rn) * rad2deg
printf "cos(theta) rooms-nox = %.6f (ángulo = %.2f grados)\n", coseno_rn, theta_rn
```

```
cos(theta) price-nox = -0,426037 (ángulo = 115,22 grados)
cos(theta) rooms-nox = -0,302828 (ángulo = 107,63 grados)
```

Actividad 5 - Síntesis: signo y magnitud como ángulo

Recordando la lección 5: $\rho \approx 1$ corresponde a vectores en desviaciones casi *alineados* (ángulo pequeño); $\rho \approx -1$, a vectores casi alineados pero en *sentidos opuestos* (ángulo cercano a 180°); $\rho \approx 0$, a vectores casi *ortogonales* (ángulo cercano a 90°).

De los tres pares calculados, ¿cuál tiene el ángulo más cercano a 90° ? ¿Qué le dice eso sobre la relación (lineal) entre esas dos variables?

Apéndice - ¿Importa dividir entre n , entre $n - 1$, o no dividir?

En la Actividad 4 no dividimos por ningún n : usamos directamente el producto escalar y la norma *euclídeos* (sin el factor $1/n$ de la lección 4), y aun así el coseno obtenido coincidió con **corr**. Esto no es casualidad: al ser un *cociente*, cualquier factor común que multiplique por igual al numerador y ambas normas se cancela. Comprobémoslo explícitamente, calculando ρ de tres formas distintas:

en línea de comandos:

```
scalar n = $nobs

scalar cov_hat_n = sum(price_c*rooms_c)/n
scalar var_price_n = sum(price_c^2)/n
scalar var_rooms_n = sum(rooms_c^2)/n
scalar rho_dividiendo_n = cov_hat_n / sqrt(var_price_n*var_rooms_n)

scalar cov_hat_n1 = sum(price_c*rooms_c)/(n-1)
scalar var_price_n1 = sum(price_c^2)/(n-1)
scalar var_rooms_n1 = sum(rooms_c^2)/(n-1)
scalar rho_dividiendo_n1 = cov_hat_n1 / sqrt(var_price_n1*var_rooms_n1)

printf "rho sin dividir (Actividad 4) = %.6f\n", coseno_pr
printf "rho dividiendo por n = %.6f\n", rho_dividiendo_n
printf "rho dividiendo por (n-1) = %.6f\n", rho_dividiendo_n1
printf "rho de Gretl (corr) = %.6f\n", corr(price,rooms)
```

```
rho sin dividir (Actividad 4) = 0,695780
rho dividiendo por n = 0,695780
rho dividiendo por (n-1) = 0,695780
rho de Gretl (corr) = 0,695780
```

Los cuatro números deben coincidir. Esto no es un capricho de Gretl: es consecuencia de que ρ es un coseno, y el coseno no cambia si escalamos ambos vectores por la misma constante (homogeneidad de la norma, lección 2; invarianza de la correlación ante reescalados, lección 5). Aquí no nos limitamos a creerlo: lo hemos programado y comprobado nosotros mismos con instrucciones que Gretl no ofrece “de fábrica”.

Preguntas de interpretación para la clase

1. En los tres pares calculados (price-rooms, price-nox, rooms-nox), ¿coincide el signo de la correlación con lo que esperaríamos por argumentos económicos?
2. La covarianza y la correlación entre un mismo par de variables tienen siempre el mismo signo, pero sus magnitudes pueden ser muy distintas de un par a otro. ¿Por qué la correlación es preferible para *comparar* la fuerza de la relación entre distintos pares de variables?
3. ¿Por qué el valor de ρ no depende de si se divide por n , por $n - 1$, o no se divide en absoluto?
4. De los tres pares de variables (en desviaciones), ¿cuál tiene el ángulo más cercano a 90° ? ¿Qué interpretación geométrica tiene ese resultado?
5. Observando los tres diagramas de dispersión, ¿la “forma” de la nube (más o menos alargada/diagonal) se corresponde con la magnitud de ρ calculada?

Para profundizar:

- Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. Apéndice C.3: covarianza y correlación.
- Cottrell, A. y Lucchetti, R. (2023). *Gretl User's Guide*. Sección sobre funciones estadísticas (**corr**, **cov**) y sobre gráficos X-Y.
- Véanse las lecciones 4 y 5 para las demostraciones completas de lo que hoy verificamos numéricamente.

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 estadisticos.txt
    summary price rooms nox
end outfile

outfile --quiet matrizCorrelaciones.txt
    corr price rooms nox
end outfile

outfile --quiet covarianzasYcorrelaciones.txt
    printf "Cov(price,rooms) = %.4f    Corr(price,rooms) = %.4f\n", cov(price,rooms), corr(price,rooms)
    printf "Cov(price,nox) = %.4f    Corr(price,nox) = %.4f\n", cov(price,nox), corr(price,nox)
    printf "Cov(rooms,nox) = %.4f    Corr(rooms,nox) = %.4f\n", cov(rooms,nox), corr(rooms,nox)
end outfile

gnuplot price rooms --output="price_rooms.png"
gnuplot price nox --output="price_nox.png"
gnuplot rooms nox --output="rooms_nox.png"

series price_c = price - mean(price)
series rooms_c = rooms - mean(rooms)
series nox_c = nox - mean(nox)

outfile --quiet comprobacionOrtogonalidad.txt
    printf "Suma price_c = %.10f\n", sum(price_c)
    printf "Suma rooms_c = %.10f\n", sum(rooms_c)
    printf "Suma nox_c = %.10f\n", sum(nox_c)
end outfile

outfile --quiet coseno-price-rooms.txt
    scalar dp_pr = sum(price_c * rooms_c)
    scalar norma_price = sqrt(sum(price_c^2))
    scalar norma_rooms = sqrt(sum(rooms_c^2))
    scalar coseno_pr = dp_pr / (norma_price * norma_rooms)

    scalar rad2deg = 180/3.14159265358979
    scalar theta_pr = acos(coseno_pr) * rad2deg

    printf "cos(theta) price-rooms (a mano) = %.6f\n", coseno_pr
    printf "corr(price,rooms) de Gretl = %.6f\n", corr(price,rooms)
```

```

    printf "Ángulo price-rooms = %.2f grados\n", theta_pr
end outfile

outfile --quiet cosenosOtrosPares.txt
    scalar norma_nox = sqrt(sum(nox_c^2))

    scalar dp_pn = sum(price_c * nox_c)
    scalar coseno_pn = dp_pn / (norma_price * norma_nox)
    scalar theta_pn = acos(coseno_pn) * rad2deg
    printf "cos(theta) price-nox = %.6f (ángulo = %.2f grados)\n", coseno_pn, theta_pn

    scalar dp_rn = sum(rooms_c * nox_c)
    scalar coseno_rn = dp_rn / (norma_rooms * norma_nox)
    scalar theta_rn = acos(coseno_rn) * rad2deg
    printf "cos(theta) rooms-nox = %.6f (ángulo = %.2f grados)\n", coseno_rn, theta_rn
end outfile

outfile --quiet invarianzarespectodivisor.txt
    scalar n = $nobs

    scalar cov_hat_n = sum(price_c*rooms_c)/n
    scalar var_price_n = sum(price_c^2)/n
    scalar var_rooms_n = sum(rooms_c^2)/n
    scalar rho_dividiendo_n = cov_hat_n / sqrt(var_price_n*var_rooms_n)

    scalar cov_hat_n1 = sum(price_c*rooms_c)/(n-1)
    scalar var_price_n1 = sum(price_c^2)/(n-1)
    scalar var_rooms_n1 = sum(rooms_c^2)/(n-1)
    scalar rho_dividiendo_n1 = cov_hat_n1 / sqrt(var_price_n1*var_rooms_n1)

    printf "rho sin dividir (Actividad 4) = %.6f\n", coseno_pr
    printf "rho dividiendo por n = %.6f\n", rho_dividiendo_n
    printf "rho dividiendo por (n-1) = %.6f\n", rho_dividiendo_n1
    printf "rho de Gretl (corr) = %.6f\n", corr(price,rooms)
end outfile

```

Enlace al guión: [S07-Prct-A-hprice2.inp](#)

Respuestas

1. Signos esperados y observados. Por argumentos económicos cabe esperar $\rho_{\text{price,rooms}} > 0$ (más habitaciones, casas más grandes o mejores, mayor precio) y $\rho_{\text{price,nox}} < 0$ (más contaminación, zona menos deseable, menor precio). El signo de $\rho_{\text{rooms,nox}}$ es menos evidente a priori; con estos datos sale *negativo*, $\rho_{\text{rooms,nox}} = -0,3028$ (las zonas con más contaminación —más industriales/urbanas— tienden a tener viviendas algo más pequeñas), y de magnitud menor que los otros dos pares (0,6958 y $-0,4260$). Lo importante pedagógicamente es que el alumno formule su expectativa *antes* de mirar el número, y después la contraste.

2. Covarianza vs. correlación para comparar fuerza de relación. La covarianza depende de las *unidades* de ambas variables (dólares, número de habitaciones, partes por cien millones de NOx), así que su magnitud no es comparable entre pares distintos: una covarianza “grande” puede deberse simplemente a que una variable se mide en una escala con números grandes. La correlación, al ser un coseno, es adimensional y está siempre en $[-1, 1]$; por eso sí permite comparar, de forma homogénea, qué par de variables está “más alineado” en sentido geométrico.

3. Invarianza de ρ al divisor. Porque ρ es un *cociente*: tanto el numerador (covarianza o “producto escalar”) como el denominador (producto de las dos desviaciones típicas o “normas”) llevan el mismo factor de escala (ya sea $1/n$, $1/(n-1)$, o ningún factor en absoluto), y ese factor se cancela exactamente en la división. Es la misma propiedad de homogeneidad de la norma y de invarianza de la correlación ante reescalados demostrada en la lección 5, aplicada aquí a un caso extremo: reescalar por *cualquier* constante positiva común no cambia un coseno.

4. Ángulo más cercano a 90° . Es el par con menor $|\rho|$: **rooms-nox**, cuyo ángulo es de $107,63^\circ$, frente a los $115,22^\circ$ de **price-nox** y los $45,91^\circ$ de **price-rooms**. Un ángulo cercano a 90° significa que los vectores en desviaciones son casi *ortogonales*: la variación de una variable, una vez descontada su media, apenas “se parece” (en dirección) a la variación de la otra. Esto no implica que no exista *ninguna* relación entre ambas variables, sino específicamente que no hay relación *lineal* apreciable entre sus componentes centradas.

5. Forma de la nube vs. magnitud de ρ . Sí: cuanto mayor es $|\rho|$, más “alargada” y alineada en torno a una diagonal aparece la nube de puntos (con pendiente positiva si $\rho > 0$, negativa si $\rho < 0$); cuanto más cercano a 0, más “redonda” o dispersa en todas direcciones se ve la nube, sin una dirección diagonal dominante. Esta es la lectura visual exacta del ángulo entre los vectores centrados.