

Sesión 11 (D) — El cuarteto de Anscombe: cuatro rectas iguales, cuatro nubes distintas

Marcos Bujosa

Descripción de la práctica

Material adicional de la sesión, fuera del guion troncal: dura media hora y puede hacerse por su cuenta.

En el laboratorio que siguió a la lección 5 vimos cuatro empresas con el mismo salario medio y la misma desviación típica y estructuras salariales muy distintas: los momentos no bastan para describir *una* variable. Hoy repetimos la lección con *dos* variables y su recta de regresión. El estadístico Francis Anscombe construyó en 1973 cuatro conjuntos de once puntos (x, y) que comparten, cifra por cifra, todo lo que una tabla de regresión resume: medias, varianzas, correlación, recta ajustada y R^2 . Y sin embargo, dibujados, no se parecen en nada.

Objetivos

1. Estimar la misma regresión simple sobre cuatro conjuntos de datos y comprobar que las cuatro tablas de resultados son idénticas.
2. Dibujar los cuatro diagramas de dispersión y ver que solo uno de ellos es lo que la tabla sugiere.
3. Sacar la conclusión: la tabla de una regresión —coeficientes, R^2 , sumas de cuadrados— no dice qué forma tiene la nube de puntos; sin el gráfico, no se sabe qué se ha ajustado.

Actividad 1 - Los cuatro conjuntos de datos

Los cuatro comparten el mismo x salvo el último, en el que diez observaciones tienen el mismo valor.

en línea de comandos:

```
nulldata 11
setobs 1 1 --cross-section

series x1 = {10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5}'
series y1 = {8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68}'
series x2 = x1
series y2 = {9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74}'
series x3 = x1
series y3 = {7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73}'
series x4 = {8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8}'
series y4 = {6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89}'

summary x1 y1 x2 y2 x3 y3 x4 y4 --simple
```

	Media	Mediana	D. T.	Mín	Máx
x1	9,000	9,000	3,317	4,000	14,00
y1	7,501	7,580	2,032	4,260	10,84
x2	9,000	9,000	3,317	4,000	14,00
y2	7,501	8,140	2,032	3,100	9,260
x3	9,000	9,000	3,317	4,000	14,00

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

y3	7,500	7,110	2,030	5,390	12,74
x4	9,000	8,000	3,317	8,000	19,00
y4	7,501	7,040	2,031	5,250	12,50

Lo que debe observar

Las cuatro parejas tienen la misma media de x (9,0), la misma media de y (7,50) y las mismas desviaciones típicas (hasta el segundo decimal). Con lo visto en el laboratorio que siguió a la lección 5, esto ya debería ponerle en guardia.

Actividad 2 - Cuatro regresiones

en línea de comandos:

```
printf "conjunto  beta1  beta2  rho(x,y)  R2  SRC\n"
loop i=1..4 --quiet
  ols y$i const x$i --quiet
  printf " %d  %7.3f  %7.4f  %8.4f  %7.4f  %7.3f\n", i, $coeff(const), $coeff(x$i), corr(x$i, y$i), $rsq, $ess
endloop
```

conjunto	beta1	beta2	rho(x,y)	R2	SRC
1	3,000	0,5001	0,8164	0,6665	13,763
2	3,001	0,5000	0,8162	0,6662	13,776
3	3,002	0,4997	0,8163	0,6663	13,756
4	3,002	0,4999	0,8165	0,6667	13,742

Lo que debe observar

Las cuatro filas son la misma: recta $\hat{y} = 3 + 0,5x$, correlación 0,82, $R^2 = 0,67$, la misma suma de cuadrados de los residuos. Si solo tuviera estas tablas, diría que ha ajustado cuatro veces el mismo modelo a cuatro muestras equivalentes. El `loop` que hemos usado repite el mismo bloque para i de 1 a 4; su mecánica se explica en el laboratorio que sigue a la lección 11, aquí basta con leerlo.

Actividad 3 - Cuatro diagramas de dispersión

en línea de comandos (o, para cada pareja, marque y y x , botón derecho, /Gráfico X-Y)/:

```
gnuplot y1 x1 --output=display
gnuplot y2 x2 --output=display
gnuplot y3 x3 --output=display
gnuplot y4 x4 --output=display
```

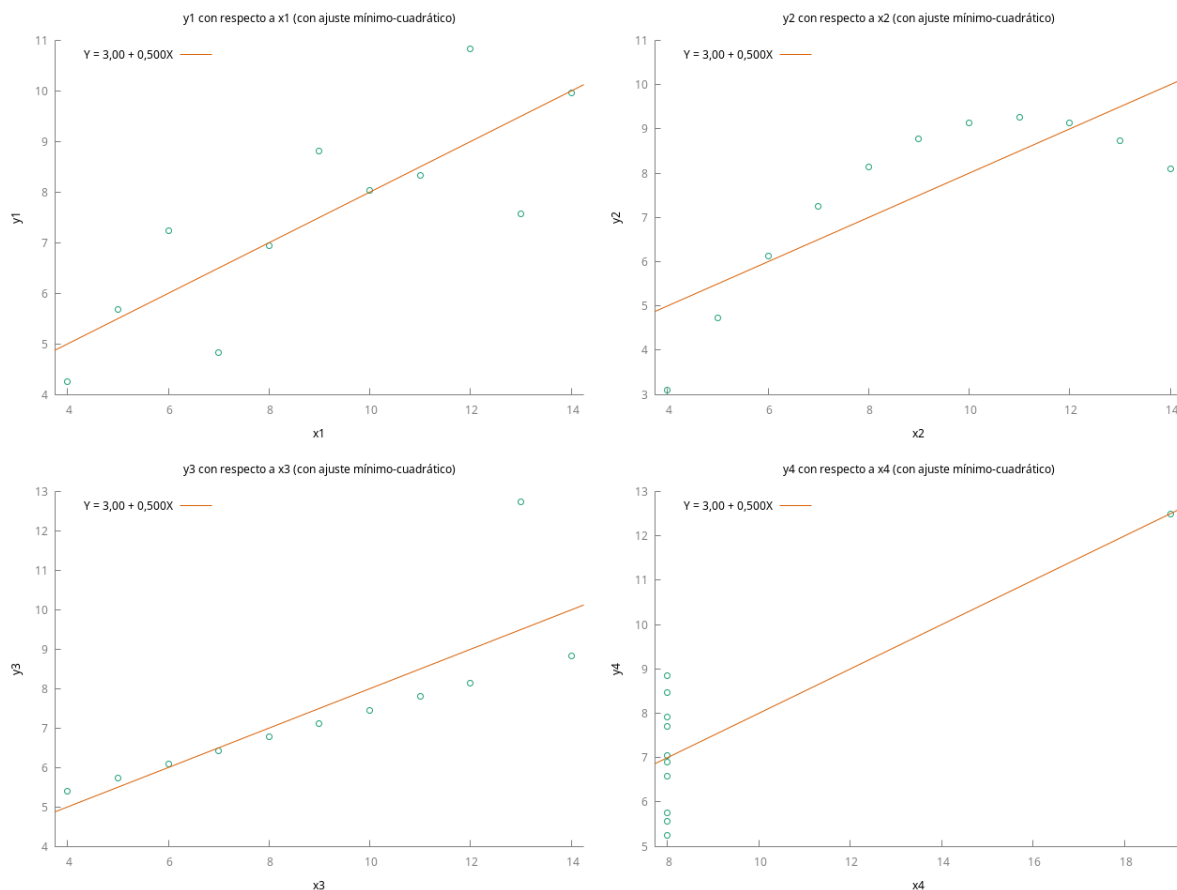


Figura 1: Los cuatro conjuntos de Anscombe, con la recta ajustada (la misma en los cuatro).

Lo que debe observar

Solo el primer conjunto es lo que la tabla sugiere: una nube alargada alrededor de la recta, con dispersión parecida en todo el rango. El segundo es una curva perfecta, sin ruido: la relación existe, pero no es lineal, y la recta la describe mal. El tercero es una recta exacta con una sola observación fuera de sitio, que tira de la pendiente. Y el cuarto no tiene relación alguna: diez puntos con el mismo x y uno solo, muy alejado, que decide por sí mismo la recta entera. Cuatro situaciones distintas, cuatro conclusiones distintas, la misma tabla.

Actividad 4 - Síntesis

Todo lo que una regresión simple imprime — $\hat{\beta}_1$, $\hat{\beta}_2$, R^2 , STC, SEC, SRC— se calcula con cinco números: dos medias, dos varianzas y una covarianza. Dos conjuntos de datos que compartan esos cinco números tendrán la misma tabla, tengan la forma que tengan. Es la versión bivalente de lo que vimos con los salarios de las cuatro empresas, y la razón práctica de la advertencia de la lección 8: un R^2 no certifica que el modelo sea sensato. El diagrama de dispersión no es un adorno del informe; es la única forma de saber qué se ha ajustado.

Preguntas de interpretación para la clase

1. ¿Por qué basta con que dos conjuntos compartan medias, varianzas y covarianza para que su tabla de regresión sea idéntica? Enumere qué cantidades de la tabla dependen de cada uno de esos cinco

números.

2. En el segundo conjunto no hay ruido: los puntos están exactamente sobre una parábola. ¿Por qué el R^2 no vale 1?
3. En el cuarto conjunto, ¿qué ocurriría con la recta si se quitara la observación con $x = 19$? ¿Y con el R^2 ?
4. ¿Cuál de los cuatro conjuntos ilustra la situación de la Actividad 6 del laboratorio que sigue a la lección 10 (una vivienda que tira del ajuste)?

Para profundizar:

- Anscombe, F. J. (1973). “Graphs in Statistical Analysis”. *The American Statistician*, 27(1), 17-21.
- Lección 8 ([lección 8](#)), sección “Una advertencia: el nombre *suma explicada* es engañoso”.

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

outfile --quiet datos.txt
  nulldata 11
  setobs 1 1 --cross-section

  series x1 = {10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5}'
  series y1 = {8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68}'
  series x2 = x1
  series y2 = {9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74}'
  series x3 = x1
  series y3 = {7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73}'
  series x4 = {8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8}'
  series y4 = {6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89}'

  summary x1 y1 x2 y2 x3 y3 x4 y4 --simple
end outfile

outfile --quiet regresiones.txt
  printf "conjunto  beta1      beta2      rho(x,y)      R2      SRC\n"
  loop i=1..4 --quiet
    ols y$i const x$i --quiet
    printf "    %d      %7.3f    %7.4f    %8.4f    %7.4f    %7.3f\n", i, $coeff(const), $coeff(x$i),
    ↪ corr(x$i, y$i), $rsq, $ess
  endloop
end outfile

gnuplot y1 x1 --output="@workdir/anscombe1.png"
gnuplot y2 x2 --output="@workdir/anscombe2.png"
gnuplot y3 x3 --output="@workdir/anscombe3.png"
gnuplot y4 x4 --output="@workdir/anscombe4.png"
```

Enlace al guión: [S11-Prc-D-anscombe.inp](#)

Respuestas

1. *Cinco números.* $\hat{\beta}_2 = \sigma_{xy}/\sigma_x^2$ (lección 7) depende de la covarianza y de la varianza de x ; $\hat{\beta}_1 = \mu_y - \hat{\beta}_2\mu_x$ añade las dos medias; $R^2 = \rho^2 = \sigma_{xy}^2/(\sigma_x^2\sigma_y^2)$ (lección 8) usa la covarianza y las dos varianzas; $STC = n\sigma_y^2$ y $SRC = (1 - R^2)STC$. Nada de la tabla mira a los datos uno a uno: todo pasa por esos cinco resúmenes, y dos nubes que los comparten son indistinguibles para la regresión.
2. *Parábola sin ruido.* Porque R^2 mide cuánto se acerca $\mathbf{y}$ al plano $\mathcal{L}(\mathbf{1}, \mathbf{x})$, es decir, a la mejor recta. Una parábola no está en ese plano: la distancia de $\mathbf{y}$ a la recta ajustada no es cero aunque la relación entre x e y sea exacta. El residuo mide lo que la recta no captura, y aquí no captura la curvatura. Es un caso de modelo mal especificado, no de datos ruidosos; la lección 9 lo llamaría un fallo del primer supuesto.
3. *Quitar el punto de $x = 19$.* Los diez puntos restantes tienen todos $x = 8$: el regresor en desviaciones es el vector nulo, $\sigma_x^2 = 0$, y la pendiente no está definida (Gretl se negaría a estimar). Esa única observación no “influye” en la recta: la *es*. El R^2 tampoco existe sin ella. Es la versión extrema de la Actividad 6 del laboratorio que sigue a la lección 10.
4. *El cuarto.* Allí una vivienda lejos del centro del rango y lejos de la recta movía el ajuste de forma visible; aquí una sola observación está tan lejos de las demás en el eje de x que determina por completo la recta. El tercer conjunto ilustra lo contrario: un punto lejos de la recta pero dentro del rango de x , que tira de la pendiente sin llegar a decidirla.