Lección 3

Marcos Bujosa

4 de octubre de 2023

Índice

1. Datos de Anscombe

Guión: anscombe.inp

En esta práctica con Gretl trabajaremos con los datos diseñados por F.J. Anscombe para ilustrar la importancia de *ver* los diagramas de dispersión entre distintas variables antes de realizar ninguna regresión y así identificar deficiencias en el planteamiento de los modelo, pues si únicamente se analizan los resultados numéricos de las estimaciones dichas deficiencias quedan ocultas.

1.1. La importancia de *ver* los datos

1. Objetivo Cuando se ajusta un modelo a los datos, es NECESARIO comenzar observando gráficamente los datos. F.J. Anscombe diseñó este conjunto de datos para ilustrar la importancia de representar gráficamente los datos antes de realizar un análisis empírico.

1.2. Los datos

$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{y}_3$	$\boldsymbol{y}_4$	$\boldsymbol{x}$	$\boldsymbol{x}_4$
8.04	9.14	7.46	6.58	10	8
6.95	8.14	6.77	5.76	8	8
7.58	8.74	12.74	7.71	13	8
8.81	8.77	7.11	8.84	9	8
8.33	9.26	7.81	8.47	11	8
9.96	8.10	8.84	7.04	14	8
7.24	6.13	6.08	5.25	6	8
4.26	3.10	5.39	12.50	4	19
10.84	9.13	8.15	5.56	12	8
4.82	7.26	6.42	7.91	7	8
5.68	4.74	5.73	6.89	5	8

1.3. Los estadísticos de los datos

Estadísticos de los datos	
Media de cada una de las variables $\boldsymbol{x}$:	9.0
Varianza de cada una de las variables $\boldsymbol{x}$:	11.0
Media de cada una de las variables $\boldsymbol{y}$:	7.5
Varianza de cada una de las variables $\boldsymbol{y}$:	4.12
Correlación entre las variables $\boldsymbol{x}$ e $\boldsymbol{y}$ de cada regresión:	0.816

Rectas de regresión	R^2
$\widehat{\boldsymbol{y}_1} = 3 \cdot \mathbf{1} + 0{,}5 \cdot \boldsymbol{x}$	0.67
$\widehat{\boldsymbol{y}_2} = 3 \cdot \mathbf{1} + 0{,}5 \cdot \boldsymbol{x}$	0.67
$\widehat{\boldsymbol{y}_3} = 3 \cdot \mathbf{1} + 0{,}5 \cdot \boldsymbol{x}$	0.67
$\widehat{\boldsymbol{y}_4} = 3 \cdot \mathbf{1} + 0{,}5 \cdot \boldsymbol{x}_4$	0.67

1.4. Actividad 1 - Estadísticos descriptivos

1. Carga de datos *Archivo -->Abrir datos -->Archivo de muestra* y en la pestaña *Gretl* seleccione `anscombe`.

   o bien teclee en linea de comandos:

   ```
   open anscombe
   ```

2. Visualice los estadísticos descriptivos de las datos

- Marque una variable (o varias) y "Pinche" con el botón derecho. Elija ***Estadísticos principales***

  o bien teclee en linea de comandos `summary` seguido de las series. Por ejemplo:

  ```
  summary --simple y1 y2 y3 y4 x x4
  ```

3. Observar la correlación entre variables

   - ***Ver -->Matriz de correlación*** y elija variables

     o, por ejemplo, teclee en linea de comandos:

     ```
     corr y1 y2 y3 y4 x x4
     ```

1.5. Actividad 2 - Cuatro regresiones

Realice los siguientes cuatro ajustes MCO

- Modelo 1: $\widehat{\boldsymbol{y}_1} = \widehat{a}\mathbf{1} + \widehat{b}\boldsymbol{x}$
- Modelo 2: $\widehat{\boldsymbol{y}_2} = \widehat{a}\mathbf{1} + \widehat{b}\boldsymbol{x}$
- Modelo 3: $\widehat{\boldsymbol{y}_3} = \widehat{a}\mathbf{1} + \widehat{b}\boldsymbol{x}$
- Modelo 4: $\widehat{\boldsymbol{y}_4} = \widehat{a}\mathbf{1} + \widehat{b}\boldsymbol{x}_4$

y compare los resultados.

- Para cada modelo:

  Modelo ->Mínimos Cuadrados Ordinarios; indique regresando y regresores. Pulse ***Aceptar*** (guarde el modelo como icono).

  o bien teclee en linea de comandos:

  ```
  Modelo1 <- ols y1 0 x
  Modelo2 <- ols y2 0 x
  Modelo3 <- ols y3 0 x
  Modelo4 <- ols y4 0 x4
  ```

1.6. Actividad 3 - Discusión

1. Compare los estadísticos de los distintos modelos
   - el coeficiente de determinación y en el coeficiente de determinación ajustado
   - la suma de cuadrados de los residuos,
   - la desviación típica de los errores de ajuste (`D.T. de la regresión`),
   - los estadísticos $\widehat{\mathcal{T}}$
   - los p-valores.
2. A la luz de los estadísticos de las cuatro regresiones ¿Qué modelo es mejor?
3. Observe el diagrama de dispersión XY en cada modelo
   - "pinche" en ***Ver -->Gráficos -->Gráfico XY (scatter)*** Elija la variable para el eje X (regresor) y la variable Y (regresando)

   o, por ejemplo, teclee en linea de comandos:

```
gnuplot y1 x
```

podemos pintar varios diagramas juntos con:

```
gnuplot y1 y2 y3 x
```

o varios diagramas separados con:

```
scatters y1 y2 y3 ; x
scatters y4 ; x4
```

4. De los cuatro modelos... ¿cuáles parecen razonables?

Código completo de la práctica `anscombe.inp`

```
open anscombe

summary --simple y1 y2 y3 y4 x x4

corr y1 y2 y3 y4 x x4

Modelo1 <- ols y1 0 x
Modelo2 <- ols y2 0 x
Modelo3 <- ols y3 0 x
Modelo4 <- ols y4 0 x4

gnuplot y1 x

gnuplot y1 y2 y3 x

scatters y1 y2 y3 ; x
scatters y4 ; x4
```

2. Los errores de ajuste MCO son perpendiculares a los regresores

Guión: TextilTheil.inp

1. Carga de datos

   Vamos a usar el conjunto de datos de consumo percápita de textiles, de Henri Theil, Principios de Econometría, Nueva York: Wiley, 1971, p. 102. El conjunto de datos consta de 17 observaciones anuales de series de tiempo para el periodo 1923–1939 del consumo de textiles en los Países Bajos. Todas las variables son expresadas como índices con base 100 en 1925.

   Archivo -->Abrir datos -->Archivo de muestra y en la pestaña *Gretl* seleccione `theil`.

   o bien teclee en linea de comandos:

   ```
   open theil
   ```

2. Ajuste por MCO el modelo

   $$\boldsymbol{y} = \widehat{\beta}_1\mathbf{1} + \widehat{\beta}_2\boldsymbol{x}_2 + \widehat{\beta}_3\boldsymbol{x}_3 + \widehat{\boldsymbol{e}}$$

   donde $\boldsymbol{y}$ es el consumo de textiles $\boldsymbol{x}_2$ la renta y $\boldsymbol{x}_3$ los precios relativos: *Modelo ->Mínimos Cuadrados Ordinarios*; indique `consume` como regresando y `const`, `income` y `relprice` como regresores. Pulse *Aceptar* (guarde el modelo como icono).

   o bien teclee en linea de comandos:

   ```
   AjusteMCO <- ols consume const income relprice
   ```

3. Guardado de datos ajustados y de los errores

   En la ventana del modelo ajustado: *Guardar ->Valores estimados* e indique un nombre, por ejemplo `yhat`. Lo mismo para los errores: *Guardar ->Residuos* y como nombre por ejemplo `ehat`

   o bien teclee en linea de comandos:

   ```
   series ehat = $uhat
   series yhat = $yhat
   ```

4. Verificación de que los residuos son ortogonales a los regresores y al ajuste, pero no al regresando

   Compruebe que

   $$\mu_{\widehat{\boldsymbol{e}}} = 0, \quad \mu_{(\widehat{\boldsymbol{e}}\odot\boldsymbol{x}_2)} = 0, \quad \mu_{(\widehat{\boldsymbol{e}}\odot\boldsymbol{x}_3)} = 0, \quad \mu_{(\widehat{\boldsymbol{e}}\odot\widehat{\boldsymbol{y}})} = 0 \quad \text{pero} \quad \mu_{(\widehat{\boldsymbol{e}}\odot\boldsymbol{y})} \neq 0.$$

   - En la ventana principal: *Añadir ->Definir nueva variable* y en cada caso escribir la formula y pulsar en *Aceptar*
     a) ei = ehat*income
     b) er = ehat*relprice
     c) ey = ehat*yhat
     d) ec = ehat*consume

   o bien teclee en linea de comandos:

   ```
   series ei = ehat*income
   series er = ehat*relprice
   series ey = ehat*yhat
   series ec = ehat*consume
   ```

- Observe los valores medios de los productos, es decir, de `ehat`, `ei`, `er`, `ey` y `ec` marcando las variables haciendo click sobre ellas con el botón derecho y eligiendo ***Estadísticos principales***

o bien teclee en linea de comandos:

```
summary --simple ehat ei er ey ec
```

5. Explique los resultados.

Código completo de la práctica `TextilTheil.inp`

```
open theil

AjusteMCO <- ols consume const income relprice

series ehat = $uhat
series yhat = $yhat

series ei = ehat*income
series er = ehat*relprice
series ey = ehat*yhat
series ec = ehat*consume

summary --simple ehat ei er ey ec
```

3. El coeficiente de determinación como cuadrado de la correlación entre valores observados y ajustados

Guión: EjPviviendaR2.inp

Calcule el coeficiente de determinación R^2 para el ejemplo del precio de las viviendas, pero empleando el coeficiente de correlación entre los precios y los precios ajustados. (**Pista:** calcule el coeficiente de correlación lineal simple entre $\widehat{\boldsymbol{y}}$ y $\boldsymbol{y}$ y elévelo al cuadrado.) Puede hacerlo mediante los menús desplegables de las ventanas de Gretl, *o bien en linea de comandos*:

```
open data3-1
ols price const sqft
genr phat  = $yhat
Coef_detR2 = corr(price,phat)^2
```

4. La importancia a los criterios de ajuste es muy relativa

Guión: PesoEdad.inp

- Cargue los datos del ejemplo del peso y edad de ocho niños.
  - Puede descargar el fichero `PesoEdad.gdt` del subdirectorio `datos` del directorio con el material del curso,
  - o introducir los datos manualmente siguiendo *Archivo ->Nuevo conjunto de datos*. Indique que hay 8 observaciones de sección cruzada, y marque "empezar a introducir los valores de los datos". Introduzca el nombre de la primera variable y luego los datos del peso de cada niño. Pulsando en + puede añadir la segunda variable.
- Genere la serie de edades al cuadrado y de la de edades al cubo.
- Ajuste el modelo $\widehat{\boldsymbol{peso}} = \widehat{\beta}_1\mathbf{1} + \widehat{\beta}_2\boldsymbol{edad}$ y añádalo a la tabla de modelos.
- Guarde el modelo como icono y pulse sobre su icono con el botón derecho. Seleccione *Añadir a la tabla de modelos*
  - o bien, tras estimar el modelo teclee `modeltab add`
- Ajuste $\widehat{\boldsymbol{peso}} = \widehat{\beta}_1\mathbf{1} + \widehat{\beta}_2\boldsymbol{edad} + \widehat{\beta}_3(\boldsymbol{edad})^{\mathbf{2}}$ y añádalo a la tabla de modelos.
- Ajuste $\widehat{\boldsymbol{peso}} = \widehat{\beta}_1\mathbf{1} + \widehat{\beta}_2\boldsymbol{edad} + \widehat{\beta}_3(\boldsymbol{edad})^{\mathbf{2}} + \widehat{\beta}_3(\boldsymbol{edad})^{\mathbf{3}}$ y añádalo a la tabla de modelos.
- Compare los ajustes: pinchando sobre el icono de `Tabla de modelos`; o bien tecleando `modeltab show`.
- Genere las figuras de los distintos ajustes.

Fíjese que que en los dos últimos ajustes las potencias de la edad son regresores, pero que en los tres modelos la única variable *explicativa* del peso es la edad (variable explicativa y regresor no son sinónimos).

El siguiente guión realiza todos los pasos

```
open datos/PesoEdad.gdt

series Edad2=Edad^2
series Edad3=Edad^3

Modelo1 <- ols Peso_Kg const Edad
modeltab add
Modelo2 <- ols Peso_Kg const Edad Edad2
modeltab add
Modelo3 <- ols Peso_Kg const Edad Edad2 Edad3
modeltab add
modeltab show

genr yhat1=Modelo1.$yhat
Modelo1 <- gnuplot Peso_Kg yhat1 Edad --with-lp=yhat1 --output="display"
genr yhat2=Modelo2.$yhat
Modelo2 <- gnuplot Peso_Kg yhat2 Edad --with-lp=yhat2 --output="display" --fit=quadratic
genr yhat3=Modelo3.$yhat
Modelo3 <- gnuplot Peso_Kg yhat3 Edad --with-lp=yhat3 --output="display" --fit=cubic
```

5. ¿Tiene sentido llamar variable explicativa a cualquier regresor?

Guión: cigfecfr.inp

Una regresión infantil: *exploremos la teoría que "Dumbo" ofrece a los niños sobre la natalidad y las cigüeñas.*

¿Podemos encontrar relación entre la tasa de fecundidad de las mujeres francesas (`fec`) y la densidad de cigüeñas (`cig`) en Alsacia para el período 1945-1986 (`annee`)? La tasa de fecundidad está calculada como número de niños por 10000 mujeres (Indicateur conjucturel de fécondité en 2004 par l'INSEE http://www.insee.fr). Las cifras de cigüeñas proceden de The Global Population Database: NERC Centre for Population Biology (http://www3.imperial.ac.uk/cpb/research/patternsandprocesses/gpdd) y se trata del número de parejas de cigüeñas que anidan en la región de Alsacia.

- Cargue el conjunto de datos `cigfecfr.inp`.
- Realice un diagrama de dispersión entre `fec` y `cig` y calcule el coeficiente de correlación.
- Ajuste por MCO la tasa de fecundidad `fec` con la constante y `cig`
- Realice un gráfico de series temporales de ambas variables. Observe que parece haber un retardo entre la aparición de las cigüeñas y la variación en la tasa de natalidad.
- Cree una nueva serie `cig6` que sea la serie `cig` retardada 6 meses y repita los pasos anteriores. Observe que el ajuste mejora.
- A la luz de los resultados ¿"explica" el número de parejas de cigüeñas casi el 90 % de la variabilidad en la natalidad de la región de Alsacia en esos años?

El siguiente guión realiza todos los pasos

```
open datos/cigfecfr.gdt

Diagrama  <- gnuplot fec cig  --output="display"
rho=corr(fec,cig)
ols fec 0 cig

FecCig <- gnuplot fec cig --time-series --with-lines --output="display"
genr cig6=cig(-6)

Diagrama6 <- gnuplot fec cig6 --output="display"
rho=corr(fec,cig6)
ols fec 0 cig6
```