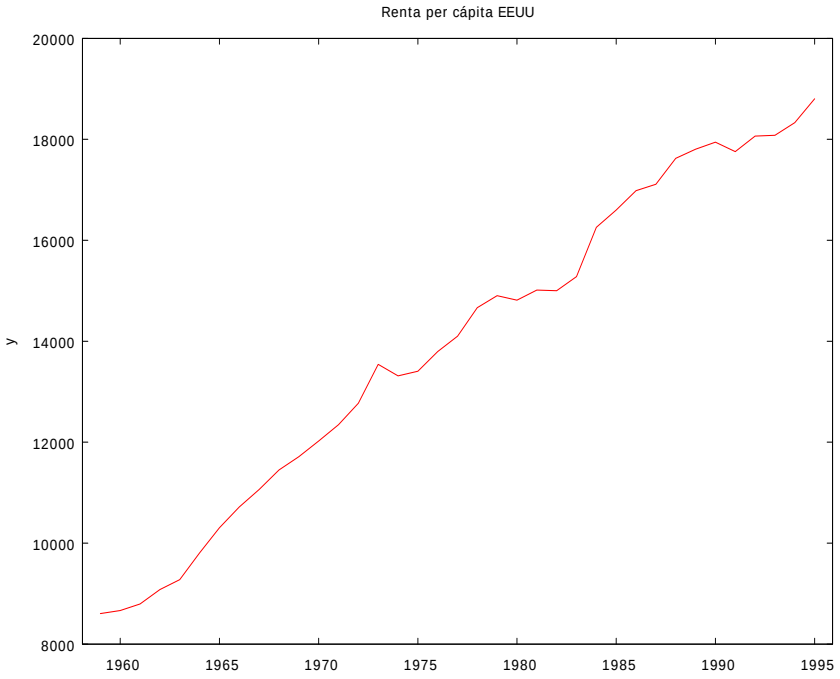
Tema 2: Análisis gráfico y estadístico de relaciones

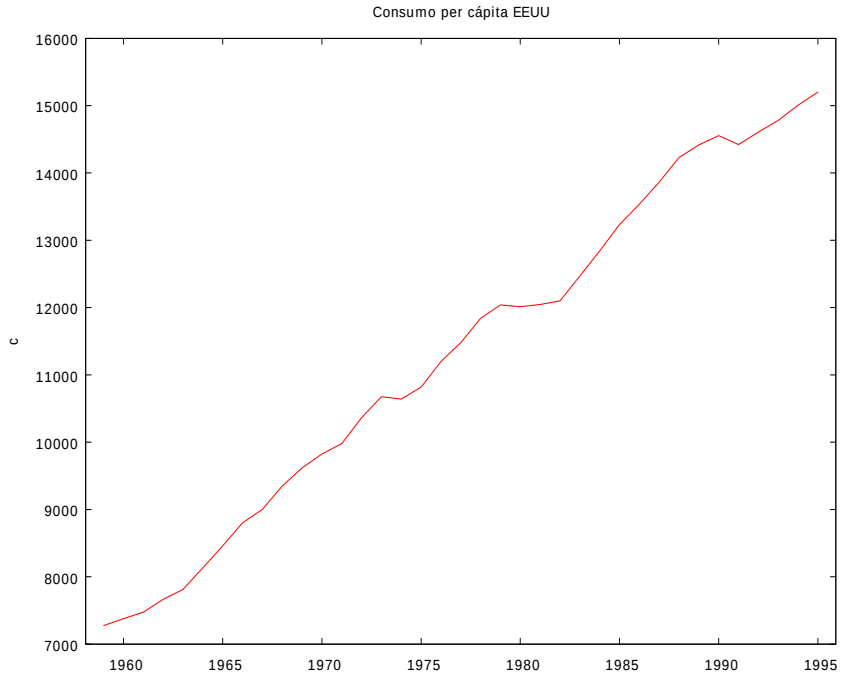
Universidad Complutense de Madrid

2013

Análisis gráfico y descriptivo de una variable (I)

- **Datos de series temporales:** Evolución anual de la renta y el Consumo per cápita en EEUU



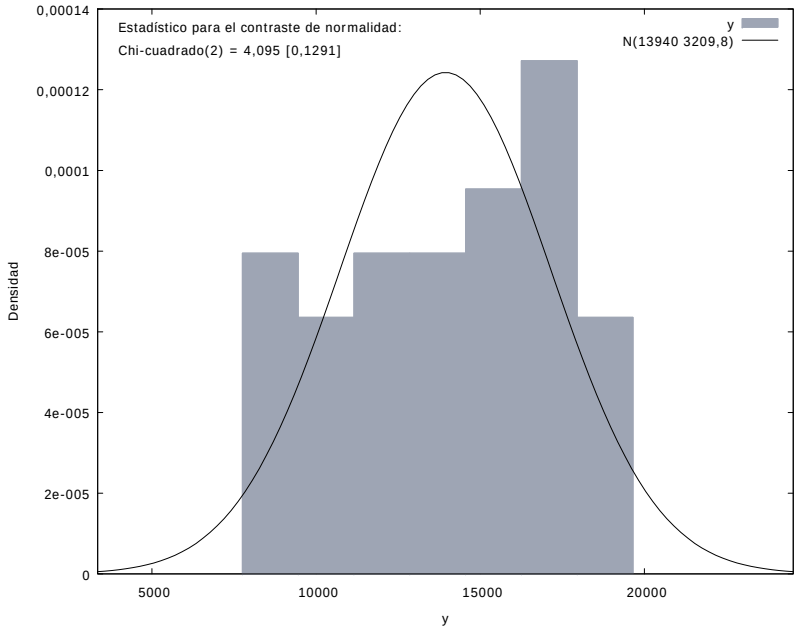


- Los dos gráficos muestran una clara tendencia creciente y común durante los años 1959 hasta 1995.

Análisis gráfico y descriptivo de una variable (II)

```
Renta per cápita en EEUU (y)
Estadísticos principales, usando las
observaciones 1959 - 1995
para la variable 'y' (37 observaciones
válidas)

  Media                          13940,
  Mediana                        14099,
  Mínimo                         8604,3
  Máximo                         18803,
  Desviación típica              3209,8
  C.V.                          0,23025
  Asimetría                    -0,17109
  Exc. de curtosis              -1,1960
```

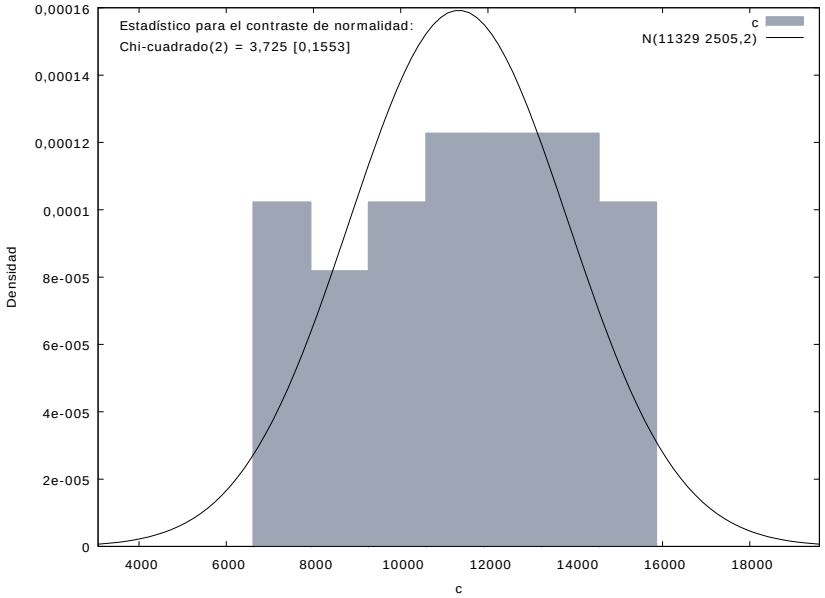


Cuanto más parecidas son la media y la mediana, más homogénea es la muestra. Como medidas de dispersión, además de la Desviación típica (DT), se calcula el Coeficiente de Variación (C.V) como el ratio entre la DT y la media (en valor absoluto). Este coeficiente es adimensional.

Análisis gráfico y descriptivo de una variable (III)

Consumo per cápita en EEUU (c)
Estadísticos principales, usando las observaciones 1959 - 1995
para la variable 'c' (37 observaciones válidas)

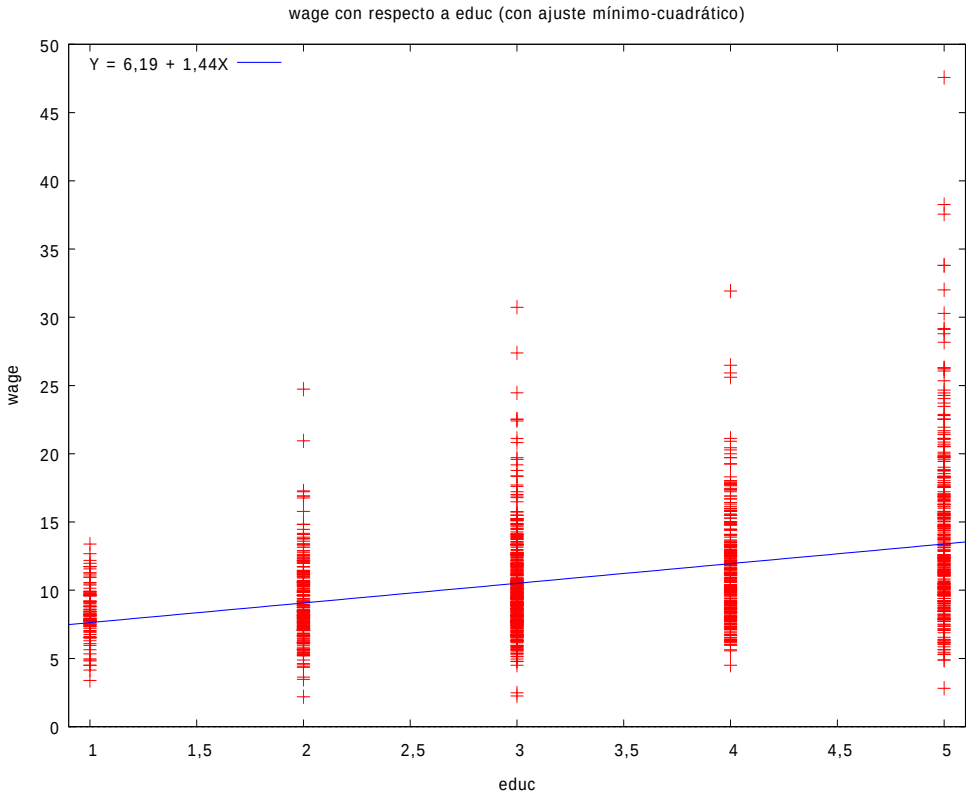
Media	11329,
Mediana	11481,
Mínimo	7274,9
Máximo	15203,
Desviación típica	2505,2
C.V.	0,22114
Asimetría	-0,073928
Exc. de curtosis	-1,2195



Los momentos de tercer y cuarto orden son la asimetría y el exceso de curtosis, sabiendo que la curtosis de una Normal es tres. En estos datos, hay defecto de curtosis. Se dibuja el histograma de los datos frente a la normal y se calcula un estadístico para contrastar normalidad.

Análisis gráfico y descriptivo de dos variables (IV)

Datos de sección cruzada: Salario (wage) en dólares por hora en función del nivel de educación del individuo



Se representa el salario (wage) de 1472 individuos con respecto a su educación (medida en 5 niveles). El nivel 1 es el de más baja educación y el 5 el más alto. Obsérvese que para un mismo nivel de educación, hay varios individuos con salarios muy diferentes.

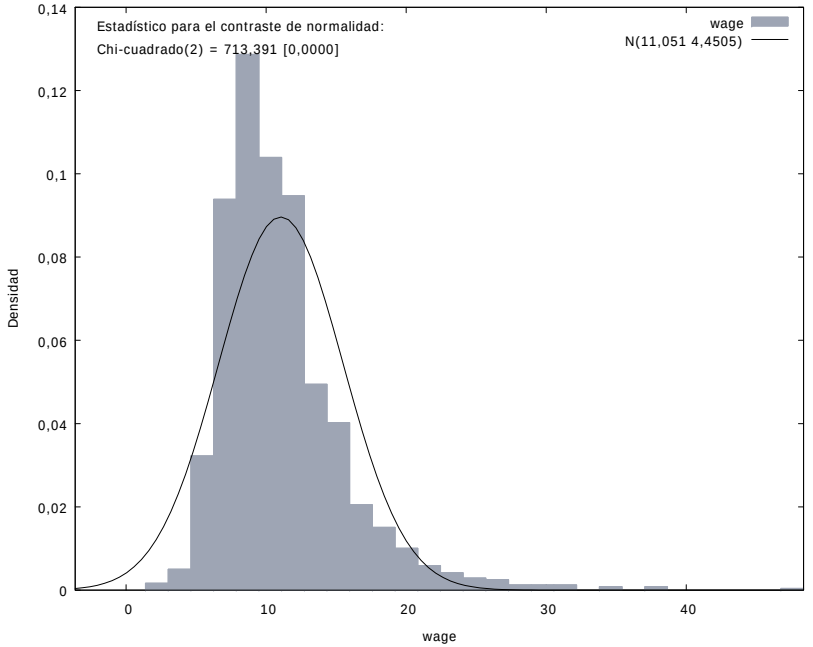
Se aprecia una clara asociación positiva entre salario y educación, pero no está clara una relación lineal entre ambas variables

Análisis gráfico y descriptivo de dos variables (V)

Salario en dólares por hora

Estadísticos principales, usando las observaciones 1 – 1472 para la variable 'wage' (1472 observaciones válidas)

Media	11,051
Mediana	10,127
Mínimo	2,1910
Máximo	47,576
Desviación típica	4,4505
C.V.	0,40274
Asimetría	1,9534
Exc. de curtosis	7,3180



El histograma de los datos de salarios muestra un elevado exceso de curtosis (7,318), es decir, una distribución mucho más apuntada que la distribución normal. El contraste rechaza la hipótesis de normalidad con total contundencia.

A veces, una variable en nivel no es normal, pero sí en logaritmos. El logaritmo contrae los valores numéricos grandes y expande los valores pequeños. Por ello, esta transformación induce normalidad.

Análisis gráfico y descriptivo de dos variables (VI)

Tablas cruzadas: otro instrumento descriptivo para secciones cruzadas

```
Tabulación cruzada de educ
(filas) contra male (columnas)

        [   0][   1]  TOT.

[   1]    23    76     99
[   2]    70   195    265
[   3]   162   258    420
[   4]   192   164    356
[   5]   132   200    332

TOTAL    579   893   1472
```

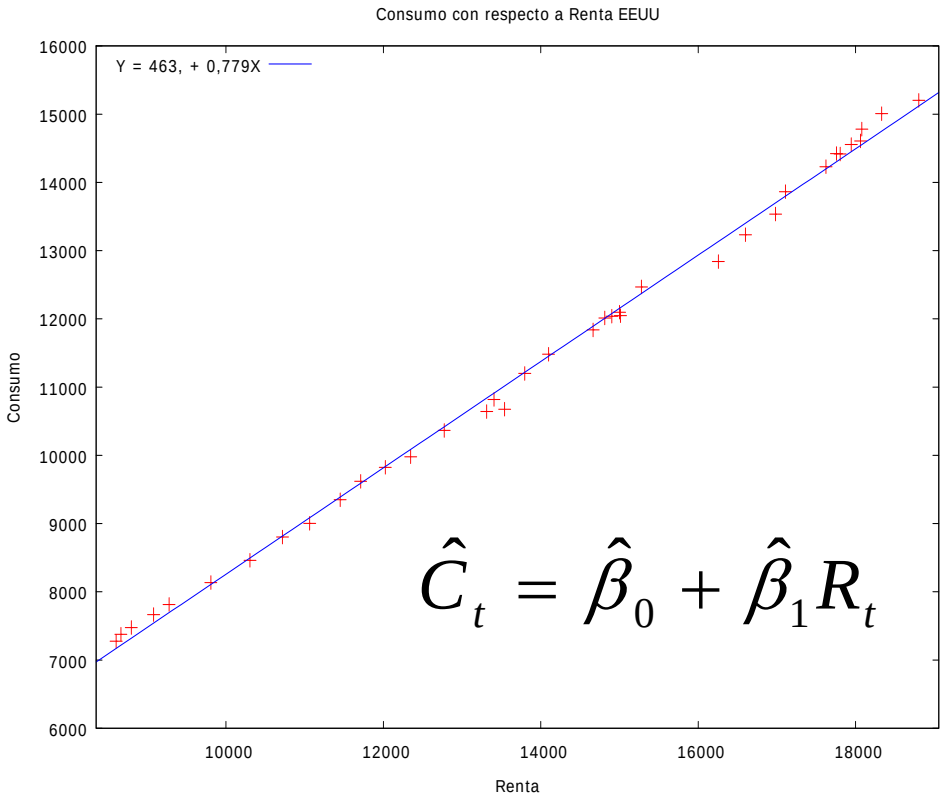
Suponga que además de la variable deeducación, se incluye información sobre el sexo del individuo (male: 1 si es hombre, 0 si es mujer)

La tabla cruzada de la izquierda informa que de un total de 1472 individuos 579 son mujeres y 893 son hombres. Además informa de cuántas mujeres y hombres tienen cada nivel de educación considerado (del 1 al 5)

Regresión lineal simple (I)

- Un paso adicional al análisis gráfico y/o descriptivo de los datos es construir un modelo lineal que relacione dos variables.
- El más sencillo es el llamado modelo de regresión lineal simple, en donde una variable de interés (endógena) viene explicada por la evolución de otra llamada variable explicativa (exógena).
- Ejemplo: la función de Consumo Keynesiana donde la variable endógena es el Consumo (*C*) y la variable explicativa es la Renta (*R*).

Regresión lineal simple (II)



Si el modelo es lineal, una estimación posible es una recta llamada RECTA DE AJUSTE. La distancia entre cada punto de la nube y la recta de ajuste es el residuo

Se quiere estimar la función de consumo

$$C_t = \beta_0 + \beta_1 R_t + \varepsilon_t$$

donde β_0 y β_1 son los parámetros de la regresión, interpretados como el consumo autónomo y la propensión marginal a consumir, respectivamente. El error ε_t es aleatorio y cualquier variable diferente a la Renta que explique el Consumo en ese momento estará recogido en él.

Se dibuja en el plano el par de valores de Consumo y Renta observados en cada año (NUBE DE PUNTOS REAL).

Regresión lineal simple (III)

- El residuo es medible, tiene signo y la misma unidad de medida que el consumo.

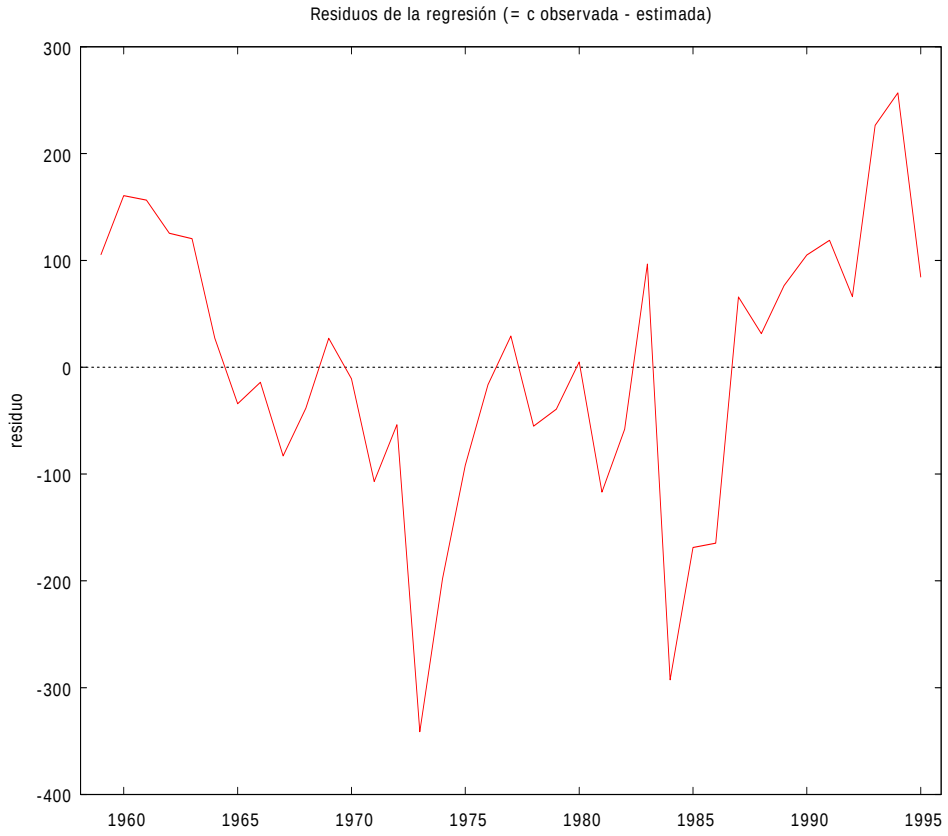
- Se calcula para cada instante de tiempo, como la distancia entre el valor del consumo observado C_t y el valor de consumo generado o ajustado por el modelo $\hat{C}_t$. Es decir:

$$\hat{\varepsilon}_t = C_t - \hat{C}_t$$

- Los residuos se pueden dibujar (en este caso, a lo largo del tiempo). Nótese que mientras que el error es no observable, el residuo se calcula.

Regresión lineal simple (IV)

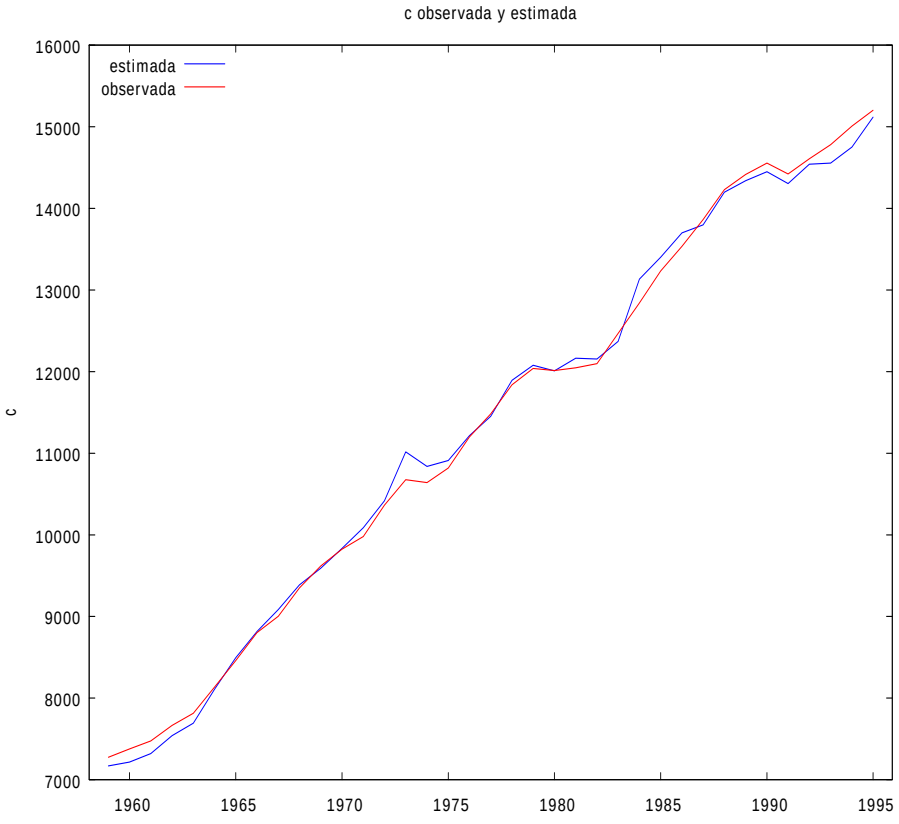
Gráfico temporal de los residuos de la regresión lineal simple del Consumo con respecto a la Renta en términos per cápita para EEUU



El gráfico de la izquierda muestra la evolución temporal de los residuos resultantes (en algunos años positivos, en otros cero y en otros negativos).

Si el residuo es positivo en ese año el Consumo observado supera al estimado por la recta, luego el modelo infraestima el verdadero valor del Consumo. Si el residuo es cero, en ese año la recta ajusta perfectamente y si es negativo, el modelo sobrestima el verdadero dato del Consumo.

Regresión lineal simple (V)



El gráfico de la izquierda muestra la evolución conjunta del Consumo observado y del Consumo ajustado (o estimado) por el modelo (recta de ajuste). La distancia en cada año de la muestra es el residuo MCO.

El residuo tiene la misma escala que el Consumo.

Este gráfico y el de los residuos resultantes, ofrecen la misma información.

Regresión lineal simple (VI)

Objetivo: Estimar los parámetros de una regresión simple de forma que se cumpla algún criterio de optimalidad. Si el criterio es minimizar la suma de los cuadrados de los residuos:

$$\min \sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \min \sum_{t=1}^{n} \left[C_t - \hat{C}_t \right]^2 = \min \sum_{t=1}^{n} \left[C_t - \hat{\beta}_0 - \hat{\beta}_1 R_t \right]^2$$

Estimación por MCO
(Mínimos Cuadrados Ordinarios)

Regresión lineal simple (VII)

$$\min \sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \min \sum_{t=1}^{n} (C_t - \hat{\beta}_0 - \hat{\beta}_1 R_t)^2$$

Condiciones de primer orden:

$$\frac{\partial \sum_{t=1}^{n} \hat{\varepsilon}_t^2}{\partial \hat{\beta}_0} = -2 \sum_{t=1}^{n} (C_t - \hat{\beta}_0 - \hat{\beta}_1 R_t) = 0$$

$$\frac{\partial \sum_{t=1}^{n} \hat{\varepsilon}_t^2}{\partial \hat{\beta}_1} = -2 \sum_{t=1}^{n} (C_t - \hat{\beta}_0 - \hat{\beta}_1 R_t) R_t = 0$$

Regresión lineal simple (VIII)

- Este es un sistema de dos ecuaciones con dos incógnitas $\hat{\beta}_0$ y $\hat{\beta}_1$ Resolviendo:

$$-\sum_{t=1}^{n} C_t + n\hat{\beta}_0 + \hat{\beta}_1 \sum_{t=1}^{n} R_t = 0 \Rightarrow \hat{\beta}_0 = \bar{C} - \hat{\beta}_1 \bar{R}$$

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} C_t R_t - n\bar{C}\bar{R}}{\sum_{t=1}^{n} R_t^2 - n\bar{R}^2} = \frac{\sum_{t=1}^{n} (C_t - \bar{C})(R_t - \bar{R})}{\sum_{t=1}^{n} (R_t - \bar{R})^2}$$

donde $\bar{C}$ y $\bar{R}$ son las medias muestrales de Consumo y Renta

Regresión lineal simple (IX)

Recapitulando, dado siguiente el modelo lineal simple

$$C_t = \beta_0 + \beta_1 R_t + \varepsilon_t$$

y una muestra de tamaño *n* de las variables C_t y R_t la estimación puntual por MCO de los dos parámetros se lleva a cabo estimando primero la pendiente y luego, la constante:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (C_t - \bar{C})(R_t - \bar{R})}{\sum_{t=1}^{n} (R_t - \bar{R})^2}$$

$$\hat{\beta}_0 = \bar{C} - \hat{\beta}_1 \bar{R}$$

Regresión lineal simple (X)

- Relación de la estimación de la pendiente con el coeficiente muestral de correlación lineal entre las dos variables:

$$\rho_{CR} = \frac{\hat{\text{cov}}[C_t R_t]}{\sqrt{\hat{\text{var}}[C_t]}\sqrt{\hat{\text{var}}[R_t]}}$$

$$\hat{\beta}_1 = \frac{\hat{\text{cov}}[C_t R_t]}{\hat{\text{var}}[R_t]}$$

Regresión lineal simple (XI)

- Por tanto:

$$\rho_{CR} = \frac{\hat{\text{cov}}[C_t R_t]}{\sqrt{\hat{\text{var}}[C_t]}\sqrt{\hat{\text{var}}[R_t]}} \frac{\sqrt{\hat{\text{var}}[R_t]}}{\sqrt{\hat{\text{var}}[R_t]}}$$

$$\rho_{CR} = \hat{\beta}_1 \frac{\sqrt{\hat{\text{var}}[R_t]}}{\sqrt{\hat{\text{var}}[C_t]}} = \hat{\beta}_1 \frac{se[R_t]}{se[C_t]}$$

- donde *s e [ ]* denota la desviación típica muestral de la variable.
- No son iguales, sino directamente proporcionales y tienen el mismo signo. El coeficiente de correlación es adimensional y está acotado entre -1 y 1.

Regresión lineal simple (XII)

Coeficiente de correlación entre Consumo y Renta

corr(c, y) = 0,99862710

Modelo: MCO, usando las observaciones 1959-1995 (T = 37)
Variable dependiente: c

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	463,177	98,7912	4,688	4,10e-05	***
y	0,779419	0,00691064	112,8	1,99e-046	***

Media de la vble. dep.	11328,65	D.T. de la vble. dep.	2505,241
Suma de cuad. residuos	619971,4	D.T. de la regresión	133,0920
R-cuadrado	**0,997256**	R-cuadrado corregido	0,997178
F(1, 35)	12720,51	Valor p (de F)	1,99e-46
Log-verosimilitud	-232,4412	Criterio de Akaike	468,8824
Criterio de Schwarz	472,1042	Crit. de Hannan-Quinn	470,0182

En la regresión lineal simple, el coeficiente de correlación lineal simple al cuadrado coincide con el R-cuadrado (0.997256)

Transformación logarítmica y semilogarítmica

Modelo teórico	Interpretación matemática	Interpretación conceptual
$y_t = \beta x_t + \varepsilon_t$	$\beta = \frac{\Delta y_t}{\Delta x_t}$	Cambio esperado en y_t cuando x_t aumenta en una unidad
$\ln y_t = \beta \ln x_t + \varepsilon_t$	$\beta = \frac{\%\ \Delta\ y_t}{\%\ \Delta\ x_t}$	Elasticidad. Cambio porcentual en y_t cuando x_t aumenta en un 1%
$\ln y_t = \beta x_t + \varepsilon_t$	$100\beta = \frac{\%\Delta y_t}{\Delta x_t}$	Semielasticidad. Cambio porcentual en y_t cuando x_t aumenta en 1 unidad
$y_t = \beta \ln x_t + \varepsilon_t$	$\frac{\beta}{100} = \frac{\Delta y_t}{\%\ \Delta x_t}$	Semielasticidad. Cambio en y_t en unidades cuando x_t aumenta en un 1%

Data Set 1: Datos de Anscombe (I)

- Usando los datos de Anscombe disponibles en los archivos de datos de muestra de Gretl, se pide estimar por MCO las cuatro regresiones simples siguientes:

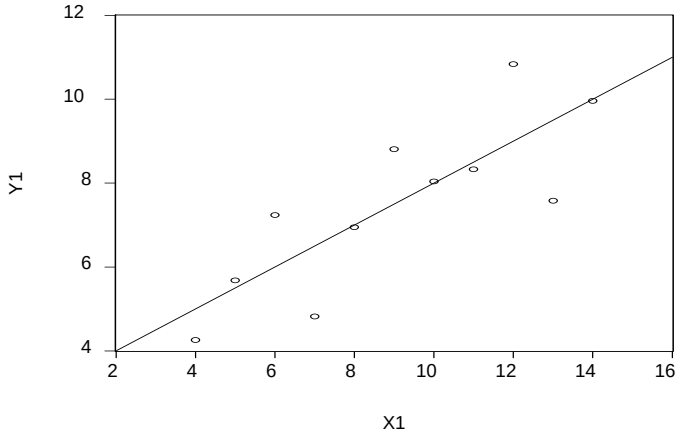
$$y_{t1} = \beta_{11} + \beta_{12} x_{t1} + \varepsilon_{t1}$$

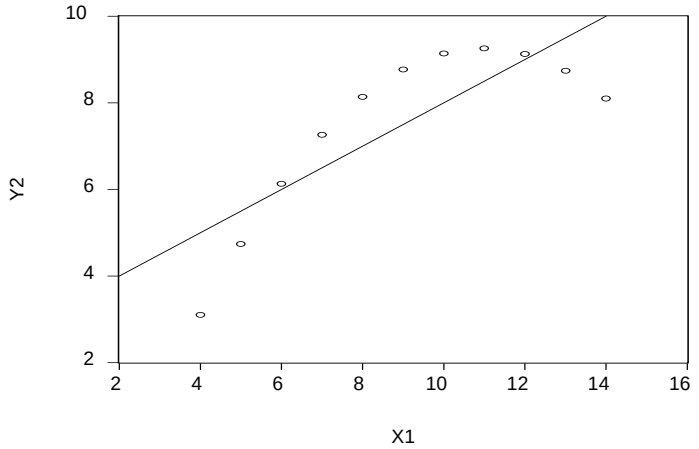
$$y_{t2} = \beta_{21} + \beta_{22} x_{t1} + \varepsilon_{t2}$$

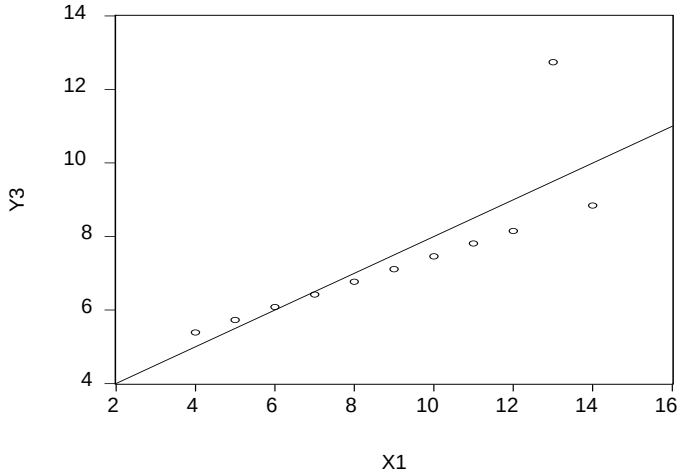
$$y_{t3} = \beta_{31} + \beta_{32} x_{t1} + \varepsilon_{t3}$$

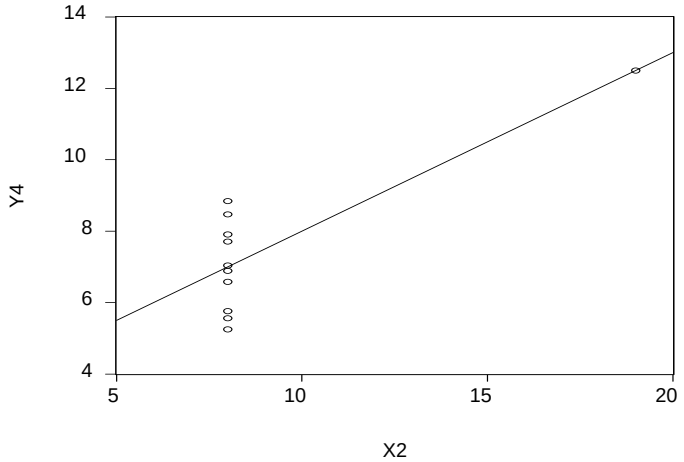
$$y_{t4} = \beta_{41} + \beta_{42} x_{t2} + \varepsilon_{t4}$$

Data Set 1: Datos de Anscombe (II)









Tareas a realizar por el alumno

(1) Estimar por MCO las cuatro regresiones lineales usando Gretl.

(2) Especificar y estimar otras relaciones entre *y2* y *x1* de forma que el ajuste de los datos mejore. Por ejemplo, pruebe a introducir como regresores *x1* y su cuadrado, o bien, sustituir *x1* por su logaritmo neperiano.

(3) Reestime la regresión de *y3* sobre *x1*, eliminando el tercer par de valores de ambas variables. ¿Cómo cambian los resultados?

(4) ¿Es posible estimar la regresión de *y4* sobre *x2*, eliminado el octavo par de valores de ambas variables? Calcule la varianza muestral de *x2* y la varianza de la pendiente de la regresión.