Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 4. Regresión y correlación II

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

- Tema 4: Regresión y correlación.

    - Regresión lineal múltiple
    - Regresión no lineal.
    - Regresión robusta

Modelo de regresión lineal múltiple

Múltiple. Mas de una variable independiente.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Se supone que las variables independientes son linealmente independientes entre si.

ε_i: término de error o perturbación: factores distintos a ***x´s*** que afectan a ***y*** (y que no observamos).

Interpretación de β_i: La variable dependiente Y cambiara en β_i unidades en promedio cuando la variable independiente (X_i) cambia en una unidad, bajo ceteris paribus.

Modelo de regresión lineal múltiple

Lineal Múltiple**.**

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Se llama lineal por los parámetros

Los parámetros del modelo son los $\beta_i' s$

Este modelo es lineal en los parámetros y en las variables.

$$Y_i = \beta_1 + \beta_2 X_i^2 + \varepsilon_i$$

No lineal en variables, se puede estimar con MMC o MCO

$$Y_i = \beta_1 X_i^{\beta_2} \varepsilon_i$$

No lineal en variables, se puede estimar con MMC o MCO

Transforma en lineal

$$LnY_i = Ln\beta_1 + \beta_2 LnX_i + \varepsilon_i$$

Modelo estimable bajo MCO

Algunos modelos de regresión no lineales

Modelo log-log

$$Y_i = \beta_1 X_i^{\beta_2} \varepsilon_i$$

$$LnY_i = Ln\beta_1 + \beta_2 LnX_i + \varepsilon_i$$

Interpretación de β_2: cuando X varia en un (1%), Y varia en β_2% en promedio.

Modelo log-lin

$$LnY_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Interpretación de β_2: cuando X varia en una unidad, Y varia en $(100\beta_2)$% en promedio.

Modelo lin-log

$$Y_i = \beta_1 + \beta_2 LnX_i + \varepsilon_i$$

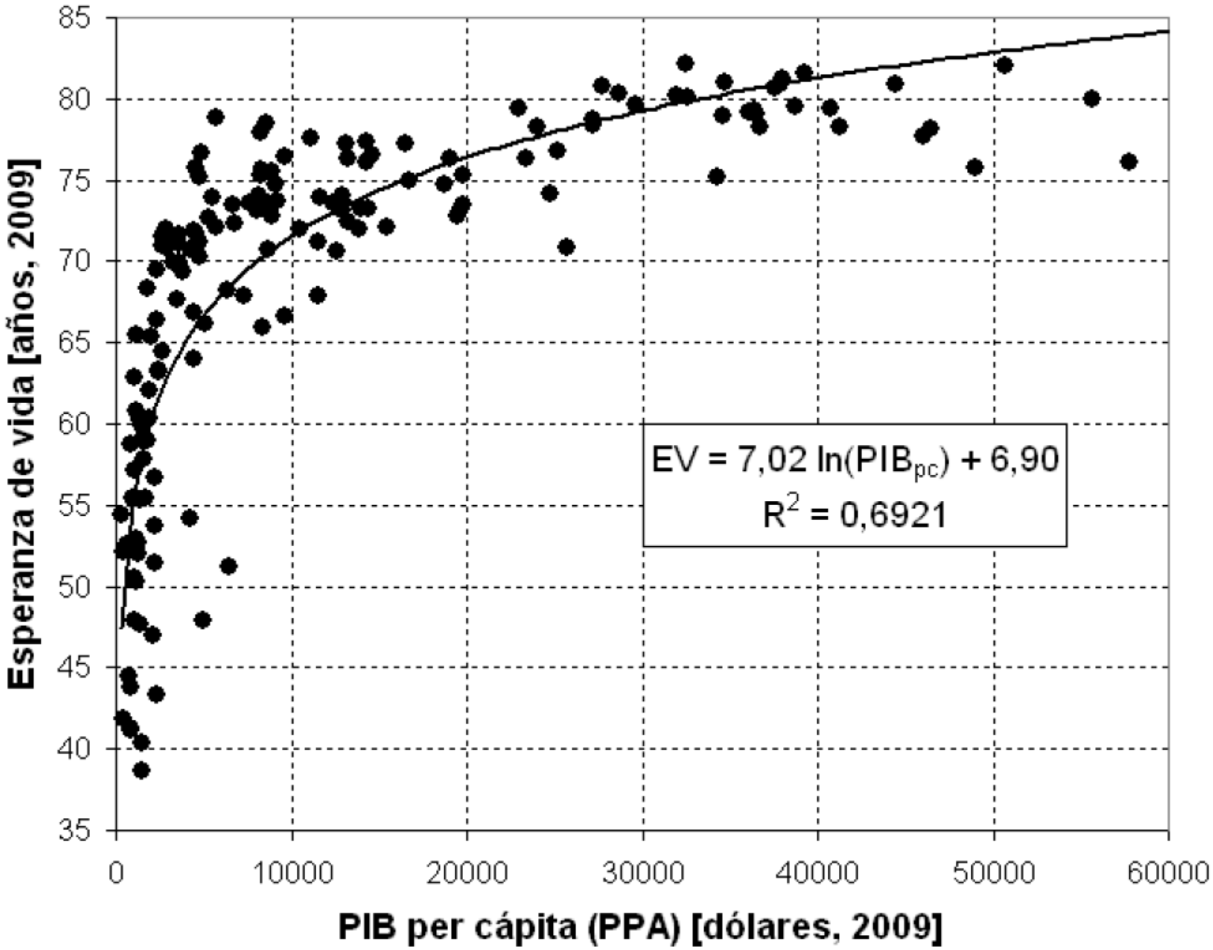
Interpretación de β_2: un incremento de un (1%) en X esta asociado con un cambio en Y de $\frac{\beta_2}{100}$ en promedio. (100% en $X -> \beta_2$ en Y).

Modelo lin-lin

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Interpretación de β_2: cuando X varia en una unidad, Y varia en β_2 unidades en promedio.

Ejemplo modelos no lineales:



Un crecimiento de un 100% en el PIB per cápita predice un aumento de 7,02 años de vida en promedio. 10% en PIB per capita, produce un aumento en 0.702 años.

Ejemplo data=mtcars

```
> print(mtcars[1:5,])
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

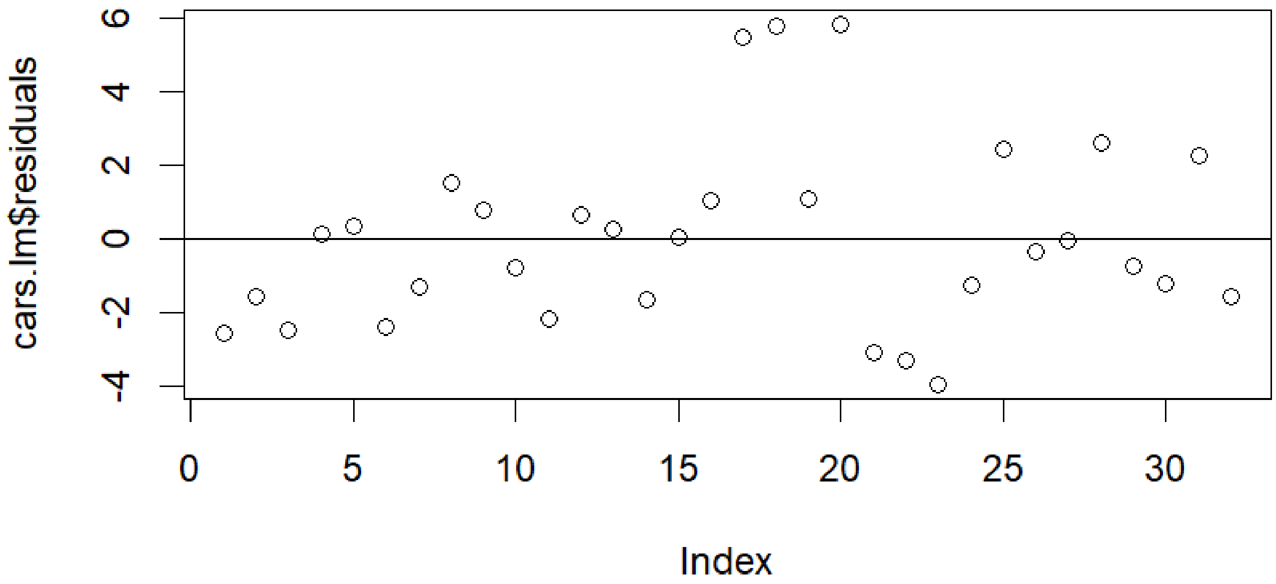
```
Call:
lm(formula = mpg ~ hp + wt)

Residuals:
   Min     1Q Median     3Q    Max 
-3.941 -1.600 -0.182  1.050  5.854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,	Adjusted R-squared:  0.8148 
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

#mpg:Miles/US Gallon
#hp:Gross horsepower
#wt:Weight (lb/1000)

Ejemplo data=mtcars

$$Mpg = 37.22727 - 0.03177 \cdot hp - 3.87783 \cdot wt$$

Gráfico de residuales:



Todo parece indicar que hay una relación lineal entre las variables. No parece haber signos de **heterocedasticidad** ni de ningún **patrón** que pudiera hacernos sospechar que la relación entre las variables fuera no lineal

LTS (Least Trimmed Squares):

Regresión de mínimos cuadrados recortados es una variación del método de regresión por mínimos cuadrados que **trata de reducir la influencia de los outliers (Regresíón robusta).**

Es un método iterativo sobre subconjuntos de puntos a los que se les va aplicando el método de mínimos cuadrados normales. Al final, nos quedamos con la versión que minimice los residuos.

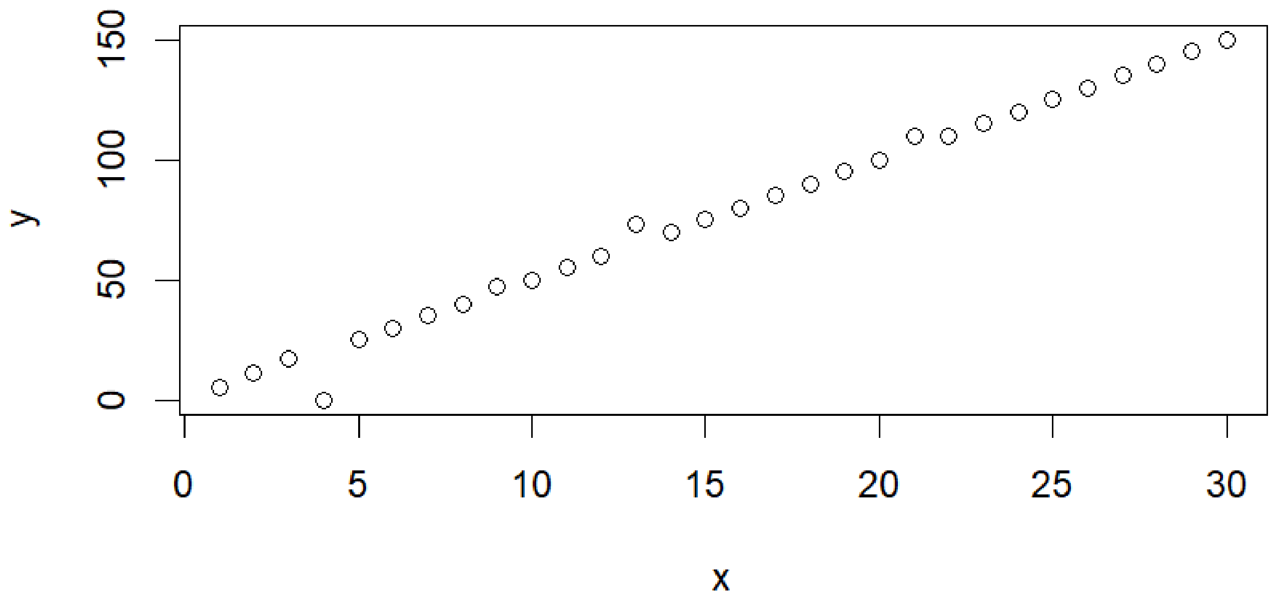
Pasos:

- Selección del número de puntos para la realización de la regresión
- Formación de los subconjuntos
- Aplicación de la regresión de mínimos cuadrados sobre cada subconjunto.
- Selección de la opción con menor error

https://www.rdocumentation.org/packages/robustbase/versions/0.95-0/topics/ltsReg
https://creates.au.dk/fileadmin/site_files/filer_oekonomi/subsites/creates/Seminar_Papers/2011/ELTS.pdf

LTS (Least Trimmed Squares):

Library(robustbase)
ltsReg: Least Trimmed Squares Robust



LTS (Least Trimmed Squares):

Library(robustbase)
ltsReg: Least Trimmed Squares Robust

```
> resultado=ltsReg(y ~ x, data=datos)
> print(resultado)

Call:
ltsReg.formula(formula = y ~ x, data = datos)

Coefficients:
Intercept          x
2.446e-14  5.000e+00
```

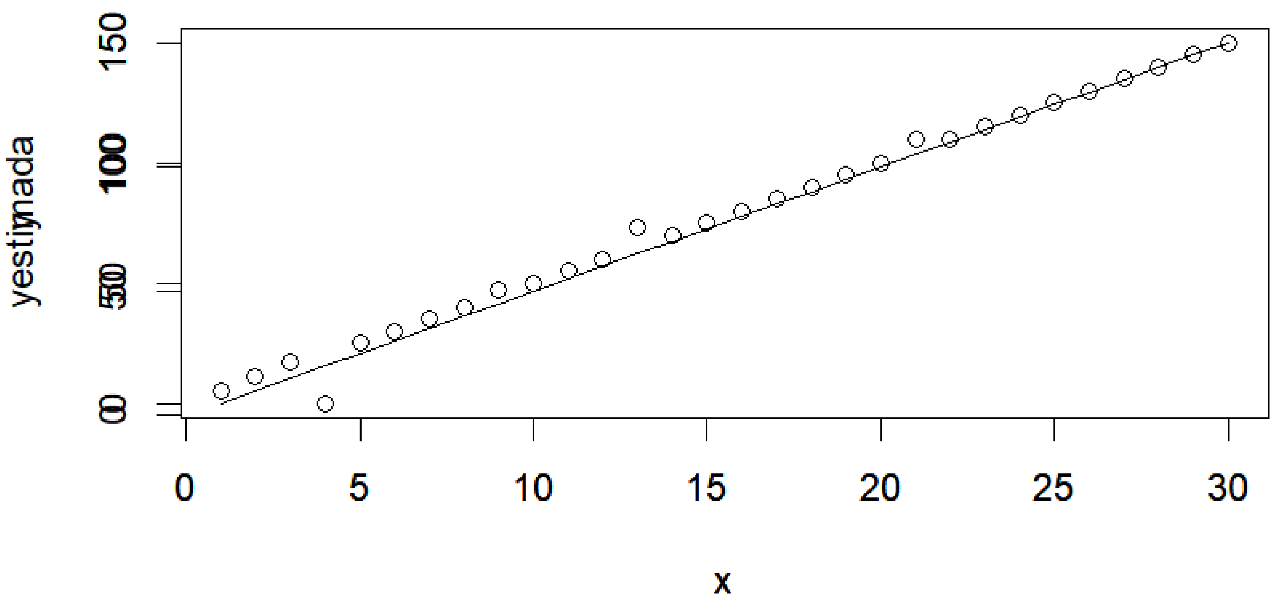
$$y = 5x + 2.446e - 14$$

A pesar de haber introducido inexactitudes en los datos, el **método es robusto y hace caso omiso de los valores outliers**.

LTS (Least Trimmed Squares):

Library(robustbase)
ltsReg: Least Trimmed Squares Robust

$$y = 5x + 2.446e - 14$$



Próxima sesión

- Tema 5: Probabilidad condicional y variables aleatorias.

  - Introducción a la teoría de la probabilidad.
  - Principios de la teoría de la probabilidad.
  - Probabilidad condicional e independencia

  - Presentación de la Actividad Grupal

  - Reflexiones sobre actividad Individual (12/01/2022) unir

LA UNIVERSIDAD
EN INTERNET

www.unir.net