Trabajo de Evaluación Continua de Modelos de Regresión

Curso 2021/2022

Xiana Carrera Alonso, Pablo Díaz Viñambres

R Markdown

Poner introducción aquí

o tablas:

celda1	celda2	celda3
celda4	celda5	celda6
celda7	celda8	celda9

Introducción

En el siguiente informe se hará un estudio estadístico en el que se analizará la influencia de la variable X sobre la variable Y, en el marco de la regresión lineal.

Librerías utilizadas

```
library(ggplot2)      # Para diagrama de dispersión con región de confianza
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sm)
```

```
## Warning: package 'sm' was built under R version 4.1.3
```

```
## Package 'sm', version 2.2-5.7: type help(sm) for summary information
```

```
library(rpanel)
```

```
## Warning: package 'rpanel' was built under R version 4.1.3
```

```
## Loading required package: tcltk
```

```
## Package 'rpanel', version 1.1-5: type help(rpanel) for summary information
```

```
library(viridis)      # Para gradiente de colores en gráfica de normalidad
```

```
## Warning: package 'viridis' was built under R version 4.1.3
```

```
## Loading required package: viridisLite
```

```
## Warning: package 'viridisLite' was built under R version 4.1.3
```

```
library(nortest)      # Necesario para lillie.test
library(car)          # Necesario para QQPlot
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

Lectura de datos

Asimismo, leemos el número de datos n.

En primer lugar, leemos los datos del archivo proporcionado, que cuenta con 76 variables respuesta, Y1, . . . , Y76, y una variable explicativa común, X. En nuestro caso, limitaremos el estudio a Y47, que denotaremos sencillamente como Y de aquí en adelante.

Nada más importar el archivo (para lo cual es necesario que el usuario cambie el directorio actual, empleando, por ejemplo, *setwd* o *Ctrl + May + H*), realizamos un pequeño análisis estadístico de los datos empleando las funciones estándar *head*, *class*, *names*, *str* y *summary*.

Por comodidad para cálculos posteriores, también guardamos el número de datos, n.

```
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path)) # Configurar wd a la carpeta actual (solo
# Ejemplo de uso de setwd para cambiar el directorio actual:
#setwd("C:\\Users\\Pablo\\Desktop\\IE_Regresion")

# Leemos los datos empleando read.table (por la extensión .txt)
# Indicamos que existe una cabecera, que las columnas están separadas por espacios y que el signo decim
datos <- read.table("datos_trabajo_temas6y7.txt", header=T, sep=" ", dec=".")
# Comprobamos la estructura de las primeras filas
head(datos)
```

```
##       Y1      Y2      Y3      Y4     Y5      Y6       Y7      Y8      Y9    Y10
## 1 2.7936  1.4752 -1.4537  0.1194 0.8976  2.2824  -0.3421  8.3449  2.6926 0.2066
## 2 9.7720 -5.6849  5.9324 -5.8971 6.7202 12.8063  -6.8206 -6.8406 24.8270 4.9308
## 3 7.0580 -6.6688  6.5421 -4.4595 6.8422  1.9432  -1.8182 -2.2344 29.5625 6.0275
## 4 6.5642 -4.7374  3.6894 -4.6314 7.8882  8.2429  -3.6053 -6.6413 20.7592 7.6040
## 5 9.6015 -7.7211  8.4552 -7.6432 9.4562 13.0013  -5.1338 -1.8076 38.6192 7.6951
## 6 9.1802 -5.6075  5.7340 -6.1928 5.9942  7.3444 -13.0044 -3.5044 18.2656 6.2518
##      Y11     Y12     Y13     Y14     Y15     Y16     Y17     Y18      Y19
## 1 3.0180  1.1751  8.4093  9.8550 -0.4541  0.4048 -2.0111  3.8622  -0.8111
## 2 6.3933 -5.9441 11.7228 23.4858 -5.9780  8.1232 -4.6495  7.7169 -10.2109
## 3 7.3119 -4.5747 13.3034 22.5277 -7.4146  9.0688 -5.5802  8.7706 -11.8967
## 4 8.0416 -4.2717 10.9460 20.9227 -7.8371  8.6477 -3.6450  9.3550  -6.1172
## 5 9.9560 -6.9841 14.3346 16.1194 -8.0503 11.5215 -7.0518 14.5981 -17.8672
## 6 8.8719 -6.4898 11.9024 14.1177 -4.8284 10.3868 -3.5018 10.0539 -10.5218
##      Y20     Y21     Y22     Y23     Y24    Y25     Y26      Y27     Y28
## 1 3.2525  3.6851  1.4066 -1.5791 23.1095 1.0566  2.4127  -2.3566  8.6631
## 2 7.0330 11.3887 -8.0260 -5.4715 43.5835 6.6822 14.4174 -13.7691 20.1870
## 3 6.0840  9.8021 -4.5552  0.7662 40.1477 5.7555 13.4886 -10.7275 17.4628
## 4 6.3954  6.7942 -2.0075 -7.7817 41.9463 5.9666 16.0072 -11.0726 16.0018
## 5 7.6931 15.2457 -2.0352 -7.6959 58.3724 8.6102 19.6299 -13.0778 21.9418
## 6 8.6100 10.4684 -2.0878 -0.1977 38.1193 5.4019 15.6906 -11.0560 17.2514
##       Y29      Y30     Y31     Y32     Y33      Y34    Y35     Y36     Y37
## 1 22.7469  -2.4220  3.5571  1.8961  0.0066   1.7376 2.1588  4.0657  1.3298
## 2 30.2249 -16.1472  8.2697 -4.0607 21.0551 -14.1993 7.6442 11.3337 -5.4907
## 3 24.0192 -13.2535  8.9133 -3.2104 15.3934 -20.2062 6.4781  0.8821 -6.3081
## 4 20.5644 -12.6253 10.2436 -3.9753 16.9404 -12.0525 6.1412 16.9086  1.6349
## 5 30.3645 -15.7150 11.6328 -6.0175 24.9262 -25.0584 8.0722 20.3899 -1.2564
## 6 28.3363 -11.9880  8.7541 -3.7809 18.2787 -13.3801 5.0839 11.0259 -1.1167
##       Y38      Y39     Y40     Y41      Y42     Y43     Y44      Y45     Y46
## 1  6.4494  -1.2899 -0.6144  5.8854   1.0759  6.5516 12.2680  -3.2194  3.9002
## 2 -4.5418  62.1693  6.6262 20.0729 -16.4415 25.0951 39.1833 -21.9718 10.5971
## 3 -3.6508  66.1270  6.7749 12.7528 -16.8969 20.6230 31.8153 -21.7585  9.9867
## 4  0.0321  63.9750  5.8612 22.9456 -21.7783 25.9051 35.9875 -18.2197 10.3999
## 5 -3.6769 109.1664  8.4507 26.7306 -25.0174 37.9203 47.3905 -23.7512 13.2672
## 6 -8.0320  60.2133  6.2397 15.6328 -20.7705 25.9365 32.7206 -16.5978 10.5171
##       Y47     Y48      Y49    Y50     Y51      Y52     Y53      Y54    Y55
## 1  1.4567  8.8645  -1.1170 2.3795  5.0323   3.3162  0.6895  -5.9556 2.4969
## 2 -1.2109 17.1167 -22.5165 6.4156 10.7858  -0.7594 -5.0630  77.7399 6.0267
## 3 -2.1529 17.6692 -11.5372 7.3901  1.3646   0.3152 -8.4656  77.1477 7.8546
## 4 -2.1343 25.0218 -21.7804 6.5836  4.5108  -0.3765 -2.9016  90.0764 5.7428
## 5 -4.9584 27.1666 -29.8614 9.1737  4.6134 -13.0519 -8.6802 126.6687 8.4139
## 6 -1.3663 17.7839 -20.2871 5.5370 13.4254  -2.6063 -4.0223  49.2404 6.3436
##       Y56      Y57     Y58     Y59      Y60     Y61     Y62     Y63      Y64
## 1  3.7127  -2.4819  4.5862 44.3813  -9.5253  6.5714  3.2692 -1.7145  -1.6168
## 2 23.2494 -22.3551 38.2988 57.6196 -22.5788 12.5265  0.2058 20.2873 -19.4924
## 3 28.1697 -22.5717 28.9487 34.4986 -23.8544 10.4411 -0.3878 24.2690 -23.6455
## 4 25.2633 -20.9713 26.6047 41.0510 -21.0315 11.2169 -2.9589 33.4181 -25.6653
## 5 37.0260 -31.3867 34.4175 68.6210 -28.3239 13.0415 -2.6899 47.3101 -33.9286
## 6 30.3660 -24.0677 27.4447 37.6048 -25.3962 10.2760 -1.1854 24.9994 -31.4601
##      Y65     Y66     Y67     Y68      Y69    Y70     Y71      Y72     Y73
## 1 1.8847  5.5941  3.0926  4.1399 -19.1810 1.0801  8.6350  -8.5187  3.9191
## 2 6.2252 13.4508 -6.0011 -4.8800  83.7734 4.9470 26.9301 -33.3999 34.6876
## 3 6.8574  7.1768 -0.2933 -3.8193 123.9641 6.0652 33.6126 -27.2912 29.2082
## 4 6.5760  5.3684  1.0325 -6.8934  83.6935 5.2763 30.2013 -32.7018 40.2327
```

```
## 5 8.8142 15.7666 -8.4158 -5.8817 186.4314 9.6777 44.7807 -35.8710 42.3833
## 6 6.7634 16.3575 -2.3546 -5.1990  95.2365 6.2355 21.9198 -30.3876 32.5226
##       Y74      Y75     Y76      X
## 1 21.5076   1.6372  5.8789 1.1370
## 2 43.5076 -30.4974 12.7668 6.2230
## 3 45.2820 -32.0879 11.5578 6.0927
## 4 64.9294 -33.1266 11.0957 6.2338
## 5 70.0942 -40.3019 12.8412 8.6092
## 6 49.3564 -31.8837 11.6881 6.4031
```

```
# Comprobamos que el objeto resultante es un data.frame
class(datos)
```

```
## [1] "data.frame"
```

```
# Vemos los nombres de las variables
names(datos)
```

```
##  [1] "Y1"  "Y2"  "Y3"  "Y4"  "Y5"  "Y6"  "Y7"  "Y8"  "Y9"  "Y10" "Y11" "Y12"
## [13] "Y13" "Y14" "Y15" "Y16" "Y17" "Y18" "Y19" "Y20" "Y21" "Y22" "Y23" "Y24"
## [25] "Y25" "Y26" "Y27" "Y28" "Y29" "Y30" "Y31" "Y32" "Y33" "Y34" "Y35" "Y36"
## [37] "Y37" "Y38" "Y39" "Y40" "Y41" "Y42" "Y43" "Y44" "Y45" "Y46" "Y47" "Y48"
## [49] "Y49" "Y50" "Y51" "Y52" "Y53" "Y54" "Y55" "Y56" "Y57" "Y58" "Y59" "Y60"
## [61] "Y61" "Y62" "Y63" "Y64" "Y65" "Y66" "Y67" "Y68" "Y69" "Y70" "Y71" "Y72"
## [73] "Y73" "Y74" "Y75" "Y76" "X"
```

```
# Comprobamos la estructura de los datos
str(datos)
```

```
## 'data.frame':    120 obs. of  77 variables:
##  $ Y1 : num  2.79 9.77 7.06 6.56 9.6 ...
##  $ Y2 : num  1.48 -5.68 -6.67 -4.74 -7.72 ...
##  $ Y3 : num  -1.45 5.93 6.54 3.69 8.46 ...
##  $ Y4 : num  0.119 -5.897 -4.46 -4.631 -7.643 ...
##  $ Y5 : num  0.898 6.72 6.842 7.888 9.456 ...
##  $ Y6 : num  2.28 12.81 1.94 8.24 13 ...
##  $ Y7 : num  -0.342 -6.821 -1.818 -3.605 -5.134 ...
##  $ Y8 : num  8.34 -6.84 -2.23 -6.64 -1.81 ...
##  $ Y9 : num  2.69 24.83 29.56 20.76 38.62 ...
##  $ Y10: num  0.207 4.931 6.027 7.604 7.695 ...
##  $ Y11: num  3.02 6.39 7.31 8.04 9.96 ...
##  $ Y12: num  1.18 -5.94 -4.57 -4.27 -6.98 ...
##  $ Y13: num  8.41 11.72 13.3 10.95 14.33 ...
##  $ Y14: num  9.86 23.49 22.53 20.92 16.12 ...
##  $ Y15: num  -0.454 -5.978 -7.415 -7.837 -8.05 ...
##  $ Y16: num  0.405 8.123 9.069 8.648 11.521 ...
##  $ Y17: num  -2.01 -4.65 -5.58 -3.64 -7.05 ...
##  $ Y18: num  3.86 7.72 8.77 9.36 14.6 ...
##  $ Y19: num  -0.811 -10.211 -11.897 -6.117 -17.867 ...
##  $ Y20: num  3.25 7.03 6.08 6.4 7.69 ...
##  $ Y21: num  3.69 11.39 9.8 6.79 15.25 ...
##  $ Y22: num  1.41 -8.03 -4.56 -2.01 -2.04 ...
```

```
##  $ Y23: num  -1.579 -5.471 0.766 -7.782 -7.696 ...
##  $ Y24: num  23.1 43.6 40.1 41.9 58.4 ...
##  $ Y25: num  1.06 6.68 5.76 5.97 8.61 ...
##  $ Y26: num  2.41 14.42 13.49 16.01 19.63 ...
##  $ Y27: num  -2.36 -13.77 -10.73 -11.07 -13.08 ...
##  $ Y28: num  8.66 20.19 17.46 16 21.94 ...
##  $ Y29: num  22.7 30.2 24 20.6 30.4 ...
##  $ Y30: num  -2.42 -16.15 -13.25 -12.63 -15.71 ...
##  $ Y31: num  3.56 8.27 8.91 10.24 11.63 ...
##  $ Y32: num  1.9 -4.06 -3.21 -3.98 -6.02 ...
##  $ Y33: num  0.0066 21.0551 15.3934 16.9404 24.9262 ...
##  $ Y34: num  1.74 -14.2 -20.21 -12.05 -25.06 ...
##  $ Y35: num  2.16 7.64 6.48 6.14 8.07 ...
##  $ Y36: num  4.066 11.334 0.882 16.909 20.39 ...
##  $ Y37: num  1.33 -5.49 -6.31 1.63 -1.26 ...
##  $ Y38: num  6.4494 -4.5418 -3.6508 0.0321 -3.6769 ...
##  $ Y39: num  -1.29 62.17 66.13 63.98 109.17 ...
##  $ Y40: num  -0.614 6.626 6.775 5.861 8.451 ...
##  $ Y41: num  5.89 20.07 12.75 22.95 26.73 ...
##  $ Y42: num  1.08 -16.44 -16.9 -21.78 -25.02 ...
##  $ Y43: num  6.55 25.1 20.62 25.91 37.92 ...
##  $ Y44: num  12.3 39.2 31.8 36 47.4 ...
##  $ Y45: num  -3.22 -21.97 -21.76 -18.22 -23.75 ...
##  $ Y46: num  3.9 10.6 9.99 10.4 13.27 ...
##  $ Y47: num  1.46 -1.21 -2.15 -2.13 -4.96 ...
##  $ Y48: num  8.86 17.12 17.67 25.02 27.17 ...
##  $ Y49: num  -1.12 -22.52 -11.54 -21.78 -29.86 ...
##  $ Y50: num  2.38 6.42 7.39 6.58 9.17 ...
##  $ Y51: num  5.03 10.79 1.36 4.51 4.61 ...
##  $ Y52: num  3.316 -0.759 0.315 -0.376 -13.052 ...
##  $ Y53: num  0.69 -5.06 -8.47 -2.9 -8.68 ...
##  $ Y54: num  -5.96 77.74 77.15 90.08 126.67 ...
##  $ Y55: num  2.5 6.03 7.85 5.74 8.41 ...
##  $ Y56: num  3.71 23.25 28.17 25.26 37.03 ...
##  $ Y57: num  -2.48 -22.36 -22.57 -20.97 -31.39 ...
##  $ Y58: num  4.59 38.3 28.95 26.6 34.42 ...
##  $ Y59: num  44.4 57.6 34.5 41.1 68.6 ...
##  $ Y60: num  -9.53 -22.58 -23.85 -21.03 -28.32 ...
##  $ Y61: num  6.57 12.53 10.44 11.22 13.04 ...
##  $ Y62: num  3.269 0.206 -0.388 -2.959 -2.69 ...
##  $ Y63: num  -1.71 20.29 24.27 33.42 47.31 ...
##  $ Y64: num  -1.62 -19.49 -23.65 -25.67 -33.93 ...
##  $ Y65: num  1.88 6.23 6.86 6.58 8.81 ...
##  $ Y66: num  5.59 13.45 7.18 5.37 15.77 ...
##  $ Y67: num  3.093 -6.001 -0.293 1.032 -8.416 ...
##  $ Y68: num  4.14 -4.88 -3.82 -6.89 -5.88 ...
##  $ Y69: num  -19.2 83.8 124 83.7 186.4 ...
##  $ Y70: num  1.08 4.95 6.07 5.28 9.68 ...
##  $ Y71: num  8.63 26.93 33.61 30.2 44.78 ...
##  $ Y72: num  -8.52 -33.4 -27.29 -32.7 -35.87 ...
##  $ Y73: num  3.92 34.69 29.21 40.23 42.38 ...
##  $ Y74: num  21.5 43.5 45.3 64.9 70.1 ...
##  $ Y75: num  1.64 -30.5 -32.09 -33.13 -40.3 ...
##  $ Y76: num  5.88 12.77 11.56 11.1 12.84 ...
```

```
##  $ X  : num  1.14 6.22 6.09 6.23 8.61 ...
```

```
# Y realizamos un pequeño análisis estadístico
summary(datos)
```

```
##        Y1                Y2                Y3                Y4
##  Min.   :-0.2477   Min.   :-9.4897   Min.   :-2.1023   Min.   :-9.1588
##  1st Qu.: 3.0484   1st Qu.:-5.7411   1st Qu.: 0.6724   1st Qu.:-5.9943
##  Median : 5.0805   Median :-2.7276   Median : 2.9093   Median :-2.7178
##  Mean   : 5.3895   Mean   :-3.2972   Mean   : 3.2825   Mean   :-3.2029
##  3rd Qu.: 7.5942   3rd Qu.:-0.9511   3rd Qu.: 5.7836   3rd Qu.:-0.5092
##  Max.   :11.8944   Max.   : 2.4110   Max.   :10.4587   Max.   : 2.4227
##        Y5                Y6                Y7                Y8
##  Min.   :-0.7795   Min.   :-1.866   Min.   :-14.3113   Min.   :-11.0489
##  1st Qu.: 1.6522   1st Qu.: 2.270   1st Qu.: -5.2127   1st Qu.: -4.1702
##  Median : 3.8990   Median : 4.546   Median : -2.3262   Median : -0.8625
##  Mean   : 4.3346   Mean   : 5.418   Mean   : -3.2343   Mean   : -1.2905
##  3rd Qu.: 6.5736   3rd Qu.: 7.871   3rd Qu.: -0.1593   3rd Qu.:  1.4933
##  Max.   :10.9457   Max.   :21.604   Max.   :  6.7544   Max.   :  8.3449
##        Y9              Y10               Y11               Y12
##  Min.   :-1.829   Min.   :-0.8901   Min.   :-0.2592   Min.   :-10.4924
##  1st Qu.: 6.989   1st Qu.: 1.7983   1st Qu.: 2.8566   1st Qu.: -5.4696
##  Median :13.039   Median : 4.2346   Median : 5.1034   Median : -2.2569
##  Mean   :16.112   Mean   : 4.3589   Mean   : 5.2499   Mean   : -3.1751
##  3rd Qu.:23.215   3rd Qu.: 6.4514   3rd Qu.: 7.8987   3rd Qu.: -0.7196
##  Max.   :43.531   Max.   :10.7299   Max.   :10.8283   Max.   :  2.5404
##       Y13              Y14               Y15               Y16
##  Min.   : 4.138   Min.   : 6.574   Min.   :-10.381   Min.   : 0.4048
##  1st Qu.: 6.980   1st Qu.:10.931   1st Qu.: -6.495   1st Qu.: 3.8066
##  Median : 8.796   Median :13.473   Median : -3.727   Median : 5.6679
##  Mean   : 9.279   Mean   :14.015   Mean   : -4.190   Mean   : 6.3914
##  3rd Qu.:11.422   3rd Qu.:16.799   3rd Qu.: -1.598   3rd Qu.: 9.1575
##  Max.   :16.136   Max.   :23.486   Max.   :  1.400   Max.   :12.9407
##       Y17               Y18               Y19               Y20
##  Min.   :-8.7036   Min.   :-4.973   Min.   :-19.249   Min.   :-1.422
##  1st Qu.:-4.8080   1st Qu.: 1.843   1st Qu.:-11.477   1st Qu.: 2.304
##  Median :-2.2672   Median : 6.061   Median : -5.399   Median : 3.688
##  Mean   :-2.3692   Mean   : 6.351   Mean   : -6.473   Mean   : 4.314
##  3rd Qu.: 0.3271   3rd Qu.:11.011   3rd Qu.: -1.515   3rd Qu.: 6.573
##  Max.   : 4.3058   Max.   :20.155   Max.   :  4.032   Max.   :10.490
##       Y21               Y22               Y23               Y24
##  Min.   : 0.7946   Min.   :-18.4719   Min.   :-9.925   Min.   :-9.932
##  1st Qu.: 3.2239   1st Qu.: -4.5732   1st Qu.:-5.144   1st Qu.: 9.515
##  Median : 5.0702   Median : -1.4525   Median :-1.068   Median :21.659
##  Mean   : 6.3172   Mean   : -2.5908   Mean   :-1.430   Mean   :28.843
##  3rd Qu.: 8.2181   3rd Qu.:  0.6842   3rd Qu.: 2.182   3rd Qu.:46.626
##  Max.   :18.1741   Max.   :  5.6872   Max.   : 7.590   Max.   :92.268
##       Y25               Y26               Y27               Y28
##  Min.   :-2.223   Min.   :-2.213   Min.   :-21.125   Min.   : 3.714
##  1st Qu.: 1.915   1st Qu.: 4.587   1st Qu.:-12.480   1st Qu.: 8.816
##  Median : 3.854   Median : 7.980   Median : -7.104   Median :12.078
##  Mean   : 4.288   Mean   : 9.530   Mean   : -7.484   Mean   :13.260
##  3rd Qu.: 6.531   3rd Qu.:14.420   3rd Qu.: -2.084   3rd Qu.:17.261
##  Max.   :10.695   Max.   :20.826   Max.   :  3.126   Max.   :24.713
```

```
##       Y29              Y30               Y31              Y32
##  Min.   : 7.941   Min.   :-21.329   Min.   : 1.422   Min.   :-7.0392
##  1st Qu.:17.240   1st Qu.:-13.804   1st Qu.: 4.783   1st Qu.:-3.7479
##  Median :22.875   Median : -7.595   Median : 6.683   Median :-0.5868
##  Mean   :23.859   Mean   : -8.606   Mean   : 7.209   Mean   :-1.2995
##  3rd Qu.:29.656   3rd Qu.: -3.494   3rd Qu.: 9.472   3rd Qu.: 1.0191
##  Max.   :50.584   Max.   :  1.947   Max.   :13.591   Max.   : 4.2334
##       Y33              Y34               Y35              Y36
##  Min.   :-7.728   Min.   :-28.198   Min.   :-2.408   Min.   : 0.2674
##  1st Qu.: 1.371   1st Qu.:-15.426   1st Qu.: 2.034   1st Qu.: 4.0594
##  Median : 7.525   Median : -8.321   Median : 3.588   Median : 6.3597
##  Mean   : 9.261   Mean   : -9.564   Mean   : 4.239   Mean   : 7.0433
##  3rd Qu.:17.133   3rd Qu.: -2.240   3rd Qu.: 6.788   3rd Qu.: 9.2219
##  Max.   :28.809   Max.   :  6.655   Max.   :10.742   Max.   :20.3899
##       Y37               Y38               Y39               Y40
##  Min.   :-15.1027   Min.   :-11.146   Min.   : -3.547   Min.   :-1.478
##  1st Qu.: -2.5464   1st Qu.: -4.271   1st Qu.: 12.994   1st Qu.: 1.958
##  Median :  0.1824   Median : -1.122   Median : 33.292   Median : 4.036
##  Mean   : -0.8192   Mean   : -1.620   Mean   : 43.782   Mean   : 4.396
##  3rd Qu.:  1.7337   3rd Qu.:  1.098   3rd Qu.: 68.176   3rd Qu.: 6.805
##  Max.   :  4.3262   Max.   :  6.449   Max.   :136.722   Max.   :10.735
##       Y41              Y42               Y43               Y44
##  Min.   :-7.186   Min.   :-31.783   Min.   :-0.7223   Min.   : 9.775
##  1st Qu.: 7.086   1st Qu.:-17.317   1st Qu.: 9.6409   1st Qu.:23.755
##  Median :12.739   Median :-10.253   Median :16.1155   Median :31.898
##  Mean   :13.523   Mean   :-11.593   Mean   :17.3573   Mean   :32.008
##  3rd Qu.:21.354   3rd Qu.: -4.772   3rd Qu.:25.1589   3rd Qu.:39.326
##  Max.   :31.027   Max.   :  7.163   Max.   :37.9203   Max.   :63.723
##       Y45              Y46               Y47               Y48
##  Min.   :-30.379   Min.   : 3.455   Min.   :-7.0839   Min.   :-5.669
##  1st Qu.:-20.006   1st Qu.: 5.991   1st Qu.:-2.4687   1st Qu.: 3.989
##  Median :-11.557   Median : 7.841   Median : 0.1661   Median :12.319
##  Mean   :-12.766   Mean   : 8.315   Mean   :-0.1815   Mean   :13.861
##  3rd Qu.: -5.239   3rd Qu.:10.537   3rd Qu.: 2.3278   3rd Qu.:23.917
##  Max.   :  3.775   Max.   :14.261   Max.   : 5.9654   Max.   :38.238
##       Y49              Y50               Y51               Y52
##  Min.   :-36.180   Min.   :-0.8581   Min.   : 1.365   Min.   :-14.871
##  1st Qu.:-21.109   1st Qu.: 2.0911   1st Qu.: 4.943   1st Qu.: -2.533
##  Median :-11.047   Median : 3.8673   Median : 6.649   Median :  0.872
##  Mean   :-12.865   Mean   : 4.3877   Mean   : 8.586   Mean   : -0.389
##  3rd Qu.: -2.041   3rd Qu.: 6.7008   3rd Qu.:12.362   3rd Qu.:  2.842
##  Max.   :  6.676   Max.   :10.5254   Max.   :25.002   Max.   :  5.006
##       Y53              Y54               Y55               Y56
##  Min.   :-10.383   Min.   :-14.72   Min.   :-0.9298   Min.   :-3.889
##  1st Qu.: -4.030   1st Qu.: 16.51   1st Qu.: 1.6218   1st Qu.: 8.417
##  Median : -1.738   Median : 39.72   Median : 3.5498   Median :15.384
##  Mean   : -1.451   Mean   : 55.65   Mean   : 4.1061   Mean   :18.391
##  3rd Qu.:  1.187   3rd Qu.: 86.76   3rd Qu.: 6.8756   3rd Qu.:28.227
##  Max.   :  7.928   Max.   :195.65   Max.   :10.1064   Max.   :45.738
##       Y57              Y58               Y59              Y60
##  Min.   :-42.630   Min.   : 0.1507   Min.   :11.16   Min.   :-41.59
##  1st Qu.:-24.336   1st Qu.:11.5121   1st Qu.:27.56   1st Qu.:-26.61
##  Median :-14.001   Median :20.0358   Median :41.77   Median :-14.08
##  Mean   :-15.735   Mean   :22.2223   Mean   :43.05   Mean   :-17.26
```

```
##  3rd Qu.: -5.564   3rd Qu.:33.8884   3rd Qu.:54.77   3rd Qu.: -6.92
##  Max.   :  8.046   Max.   :55.7775   Max.   :97.85   Max.   :  4.94
##       Y61              Y62               Y63               Y64
##  Min.   : 3.246   Min.   :-4.6786   Min.   :-16.123   Min.   :-46.083
##  1st Qu.: 6.560   1st Qu.:-1.8111   1st Qu.:  4.627   1st Qu.:-28.714
##  Median : 8.561   Median : 1.0312   Median : 13.647   Median :-13.595
##  Mean   : 9.141   Mean   : 0.8388   Mean   : 16.574   Mean   :-16.909
##  3rd Qu.:11.588   3rd Qu.: 3.4876   3rd Qu.: 27.989   3rd Qu.: -5.593
##  Max.   :15.708   Max.   : 6.1329   Max.   : 53.417   Max.   :  9.246
##       Y65              Y66               Y67               Y68
##  Min.   :-1.674   Min.   : 1.439   Min.   :-12.7320   Min.   :-12.656
##  1st Qu.: 1.561   1st Qu.: 6.403   1st Qu.: -1.7827   1st Qu.: -4.001
##  Median : 3.826   Median : 8.417   Median :  1.8860   Median : -1.635
##  Mean   : 4.314   Mean   : 8.997   Mean   :  0.8352   Mean   : -1.548
##  3rd Qu.: 6.787   3rd Qu.:10.678   3rd Qu.:  3.8491   3rd Qu.:  1.265
##  Max.   :10.992   Max.   :18.891   Max.   :  6.3926   Max.   :  6.570
##       Y69              Y70               Y71              Y72
##  Min.   :-19.18   Min.   :-1.064   Min.   :-6.198   Min.   :-54.773
##  1st Qu.: 19.91   1st Qu.: 2.043   1st Qu.: 9.681   1st Qu.:-32.094
##  Median : 48.11   Median : 3.807   Median :18.659   Median :-15.862
##  Mean   : 68.49   Mean   : 4.330   Mean   :21.680   Mean   :-19.948
##  3rd Qu.:101.87   3rd Qu.: 6.860   3rd Qu.:33.659   3rd Qu.: -7.352
##  Max.   :220.42   Max.   :11.401   Max.   :56.576   Max.   :  3.507
##       Y73              Y74               Y75               Y76
##  Min.   :-4.935   Min.   : 14.30   Min.   :-56.984   Min.   : 4.809
##  1st Qu.:12.821   1st Qu.: 35.79   1st Qu.:-33.665   1st Qu.: 7.861
##  Median :23.417   Median : 45.85   Median :-19.400   Median : 9.875
##  Mean   :24.692   Mean   : 49.42   Mean   :-21.114   Mean   :10.239
##  3rd Qu.:36.956   3rd Qu.: 64.77   3rd Qu.: -7.322   3rd Qu.:12.595
##  Max.   :58.703   Max.   :109.29   Max.   :  5.546   Max.   :16.729
##         X
##  Min.   :0.095
##  1st Qu.:1.795
##  Median :3.232
##  Mean   :4.268
##  3rd Qu.:6.667
##  Max.   :9.921
```

```r
# Seleccionamos las dos variables de interés
X <- datos[,"X"]
Y <- datos[,"Y47"]

# Guardamos el número de datos
n <- length(Y)
```

1) Relación entre variable explicativa y variable respuesta

En primer lugar, calculamos la covarianza entre las variables. Debemos tener en cuenta que R la calcula como una 'cuasi'covarianza, es decir, dividiendo entre $n-1$ en lugar de entre n. Para corregirlo, multiplicamos por $n-1$ y dividimos entre n, aunque también mostraremos el valor original.

y el coeficiente de correlación de los datos, con el objetivo de ver si existe relación lineal entre las variables.

```r
covar = cov(X,Y)*(n-1)/n; covar          # Covarianza
```

```
## [1] -8.275695
```

```r
cov(X,Y)                                 # Cuasicovarianza
```

```
## [1] -8.345238
```

```r
cor(X, Y)                                # Coeficiente de correlación
```

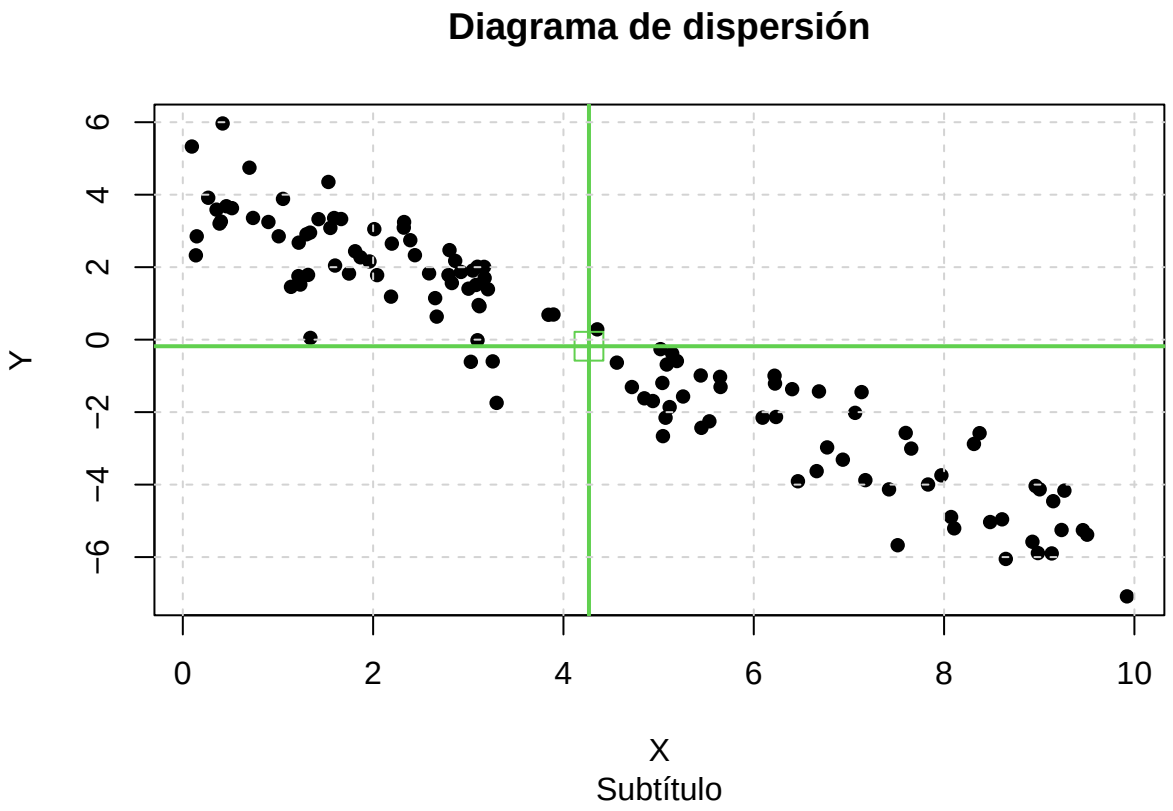
```
## [1] -0.9506962
```

Como vemos, tanto la covarianza como la correlación son negativas. En el caso de la covarianza, esta no nos indica una medida fiable de la relación entre los datos, ya que depende de la escala de los datos. Sin embargo, la correlación, con un valor de -0.95 nos da a entender una relación de proporcionalidad inversa entre X e Y, que podremos corroborar posteriormente al ver el diagrama de dispersión.

A continuación hallamos el vector de medias o centro de gravedad aplicando `mean` en ambas variables:

```r
mX <- mean(X)
mY <- mean(Y)
```

Y con la siguiente función generamos el diagrama de dispersión de los datos:



Fácilmente observamos que la nube de puntos toma una forma descendente, lo cuál encaja con el hecho de

que la correlación entre X e Y sea negativa. También vemos que los datos están, de forma aproximada, uniformemente alineados en torno a una forma rectilínea. Todo esto motiva el establecimiento de un modelo lineal para la relación entre ambas variables. Recordemos que los modelos lineales son de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Ajustamos entonces este modelo a nuestros datos mediante la función lm:

```
modelo = lm(Y~X); modelo
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.184       -1.023
```

y obtenemos un intercepto $\beta_0 = 4.184$ y una pendiente de $\beta_1 = -1.023$, lo cuál concuerda con lo observado anteriormente en la nube de puntos.

En los siguientes ejercicios, analizaremos más en profundidad este modelo. Además, lo validaremos frente a otros modelos como los polinómicos o los no paramétricos.

Ejercicio 2

Estimación puntual a mano

Para la estimación puntual de los parámetros intercepto β_0, pendiente $beta_1$ y varianza del error σ^2 podemos aplicar directamente las fórmulas obtenidas en la parte teórica de la asignatura:

```
var.X <- var(X)*(n-1)/n
beta0.gorro = mY - covar*mX/var.X; beta0.gorro
```

```
## [1] 4.184343
```

```
beta1.gorro = covar/var.X; beta1.gorro
```

```
## [1] -1.022889
```

```
var.error = sum((Y - beta0.gorro - beta1.gorro*X)^2)/(n-2); var.error
```

```
## [1] 0.916048
```

```
sd.error = sqrt(var.error); sd.error
```

```
## [1] 0.957104
```

Estimación puntual automática

De manera alternativa, podemos obtenrlas a partir del propio modelo creado anteriomente por $\mathbb{R}$:

```r
modelo    # Información del modelo
```

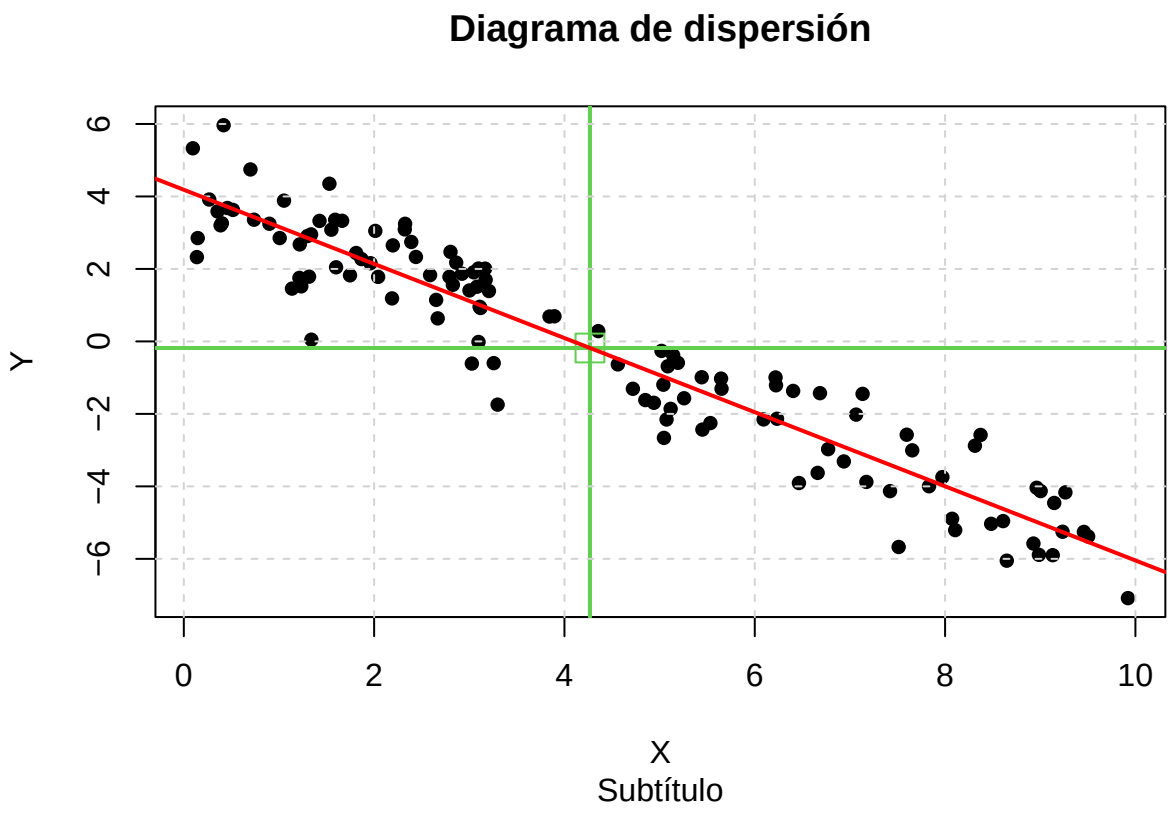
```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.184       -1.023
```

```r
modelo$coefficients         # beta0 gorro y beta1 gorro
```

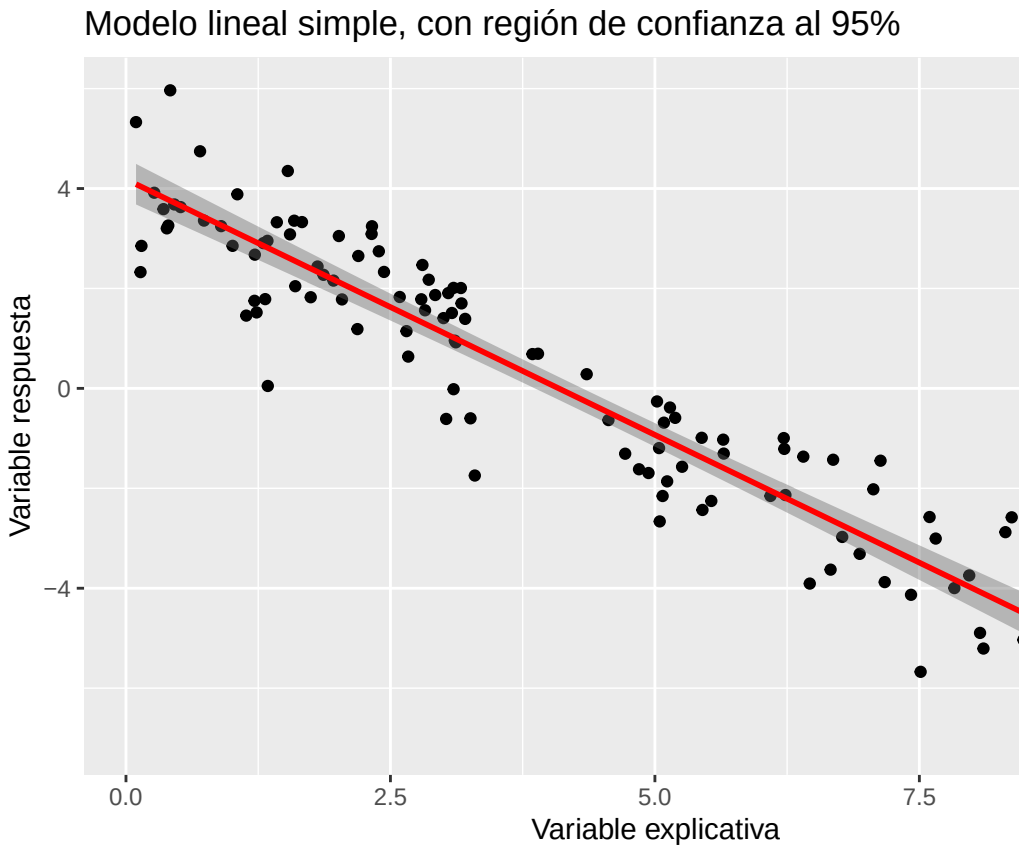
```
## (Intercept)           X
##    4.184343   -1.022889
```

```r
# En modelo$residuals están los residuos
sum(modelo$residuals^2)/(n-2)
```

```
## [1] 0.916048
```



Incluimos también una gráfica adicional usando la librería *ggplot2* e incluyendo la región o intervalo de confi-



anza para los datos al nivel del 99%:

Ejercicio 3

Calcula los intervalos de confianza para los parámetros del modelo de nivel 99Interpreta los resultados obtenidos. En primer lugar, establecemos el nivel de significación α.

```
alfa <- 1 - 0.99
```

A continuación, hallamos los intervalos para

```
beta0.cuantil <- qt(1-alfa/2, df=n-2); beta0.cuantil
```

```
## [1] 2.618137
```

```
beta0.extremoinferior <- beta0.gorro - beta0.cuantil * sqrt(var.error * (1/n + mX^2/(n*var.X)))
beta0.extremosuperior <- beta0.gorro + beta0.cuantil * sqrt(var.error * (1/n + mX^2/(n*var.X)))
beta0.IC <- c(beta0.extremoinferior, beta0.extremosuperior); beta0.IC
```

```
## [1] 3.771852 4.596833
```

```
beta1.cuantil <- beta0.cuantil
beta1.extremoinferior <- beta1.gorro - beta1.cuantil*sqrt(var.error/(var.X * n))
beta1.extremosuperior <- beta1.gorro + beta1.cuantil*sqrt(var.error/(var.X * n))
beta1.IC <- c(beta1.extremoinferior, beta1.extremosuperior); beta1.IC
```

```
## [1] -1.1033105 -0.9424673
```

```
var.error.cuantilinferior <- qchisq(alfa/2, df=n-2)
var.error.cuantilsuperior <- qchisq(1-alfa/2, df=n-2)
var.error.extremoinferior <- (n-2)*var.error^2/var.error.cuantilsuperior
var.error.extremosuperior <- (n-2)*var.error^2/var.error.cuantilinferior
var.error.IC <- c(var.error.extremoinferior, var.error.extremosuperior); var.error.IC
```

```
## [1] 0.6138273 1.2048240
```

Pero también podemos utilizar las funciones de R para hacerlo de forma automática: - IC para beta0 y beta1 asumiendo que la varianza es desconocida

```
confint(modelo, level=0.99)
```

```
##                 0.5 %     99.5 %
## (Intercept)  3.771852  4.5968334
## X           -1.103311 -0.9424673
```

No hay una automatización del cálculo de la varianza del error

Ejercicio 4

Realiza los contrates de significación asociados al intercepto y a la pendiente del modelo de regresión considerado. Interpreta los resultados obtenidos. En base a los resultados obtenidos, ¿tendría sentido considerar otro modelo más sencillo? A continuación, realizaremos los contrastes de significación sobre el modelo con el objetivo de determinar si el modelo se podría simplificar a uno con menos variables o no. En primer lugar, realizaremos el contraste de forma manual a partir de los estadísticos de contraste basado en el pivote de la estimaciones puntuales previas:

```
# Contraste de significacion para beta0
beta0.t <- abs(beta0.gorro) / (sqrt(var.error * (1/n + mX^2/(n*var.X)))); beta0.t
```

```
## [1] 26.55861
```

```
beta1.t <- abs(beta1.gorro) / (sd.error / sqrt(n*var.X)); beta1.t
```

```
## [1] 33.30029
```

```
# Rechazamos la hipótesis nula de que el modelo tiene origen de 0
beta0.t > beta0.cuantil
```

```
## [1] TRUE
```

```
# Rechazamos la hipótesis nula de que el modelo no tiene pendiente
beta1.t > beta1.cuantil
```

```
## [1] TRUE
```

```r
# El p-valor es 0 --> La hipótesis nula es falsa para cualquier nivel de signif. --> El modelo tiene un
beta0.pvalor = dt(beta0.t, df=n-2); beta0.pvalor
```

```
## [1] 2.508369e-51
```

```r
# El p-valor es 0 --> La hipótesis nula es falsa para cualquier nivel de signif. --> El modelo tiene un
beta1.pvalor = dt(beta1.t, df=n-2); beta1.pvalor
```

```
## [1] 1.237751e-61
```

De esto deducimos que existen pruebas estadísticamente significativas de que $\beta_0 \neq 0$, lo cuál nos indica que el intercepto es distinto de 0. Por otro lado, también existen pruebas de $\beta_1 \neq 0$, de dónde deducimos que realmente la variable explicativa influye en la variable respuesta.

Alternativamente, podemos obtener los valores de estos dos contrastes de significación y su p-valor a partir de los datos presentes en el modelo de R. Para esto, usaremos la función summary:

```r
summary(modelo)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76465 -0.72493  0.00685  0.71260  2.20924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.18434    0.15755   26.56   <2e-16 ***
## X           -1.02289    0.03072  -33.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9571 on 118 degrees of freedom
## Multiple R-squared:  0.9038, Adjusted R-squared:  0.903
## F-statistic:  1109 on 1 and 118 DF,  p-value: < 2.2e-16
```

En concreto, los valores relevantes son el t-value y el Pr(>|t|) de las filas (Intercept) y X que se corresponden al valor observado del estadístico observado y su p-valor en el contraste sobre el intercepto β_0 y la pendiente β_1. Obtenemos los mismos datos que en el cálculo manual.

En base a los resultados obtenidos anteriormente, decidimos no simplificar más nuestro modelo y continuar realizando regresión lineal.

Ejercicio 5

Si consideramos que la variable X toma 3 nuevos valores: 2, 4 y 6 unidades, proporciona intervalos de predicción e intervalos de confianza para la media condicionada de la variable Y. Interpreta los resultados obtenidos.

En este apartado, consideramos 3 nuevos valores para la variable explicativa $X = 2, 4, 6$. Para obtener intervalos de confianza para la media de Y condicionada a estos valores y de predicción, es necesario comprobar

primero que estos datos están dentro del rango de observación de X. Esto es debido a que no sabemos como se comporta el modelo fuera del rango observado, y nuestro objetivo es predecir y no extrapolar.

```
nuevosValores <- c(2, 4, 6)
# El rango está contenido
min(X) < min(nuevosValores) && max(X) > max(nuevosValores)
```

```
## [1] TRUE
```

```
# Construimos un data.frame  con los nuevos datos ya que predict necesita este formato para sus predicc
nuevosDatos = data.frame("X" = nuevosValores)
```

Habiendo realizado esta comprobación, ya podemos obtener los intervalos utilizando la función predict sobre el modelo de R. Obtendremos ambos intervalos para los niveles de significación 0.95 y 0.99. ((TODO: REVISAR, ESTOS ESTÁN BIEN PERO DEBERÍAN SER CONSISTENTES CON OTRAS PARTES DONDE COJAMOS ALFAS ARBITRARIOS))

En primer lugar, pasando el argumento `interval = "confidence"` obtenemos los asociados a la media condicionada.

```
predict(modelo, newdata = nuevosDatos, interval = "confidence", level=0.95)
```

```
##           fit       lwr        upr
## 1  2.13856482  1.917271  2.3598582
## 2  0.09278705 -0.080999  0.2665731
## 3 -1.95299072 -2.155557 -1.7504246
```

```
predict(modelo, newdata = nuevosDatos, interval = "confidence", level=0.99)
```

```
##           fit        lwr        upr
## 1  2.13856482  1.8459908  2.4311389
## 2  0.09278705 -0.1369772  0.3225513
## 3 -1.95299072 -2.2208054 -1.6851761
```

Y para obtener los intervalos de predicción, los cuáles serán más amplios que los anteriores, pasamos el argumento `interval = "prediction"`

```
predict(modelo, newdata = nuevosDatos, interval = "prediction", level=0.95)
```

```
##           fit        lwr         upr
## 1  2.13856482  0.2303633  4.04676635
## 2  0.09278705 -1.8104901  1.99606420
## 3 -1.95299072 -3.8591112 -0.04687022
```

```
predict(modelo, newdata = nuevosDatos, interval = "prediction", level=0.99)
```

```
##           fit        lwr       upr
## 1  2.13856482 -0.3842867 4.6614163
## 2  0.09278705 -2.4235539 2.6091280
## 3 -1.95299072 -4.4730909 0.5671095
```

Ejercicio 6

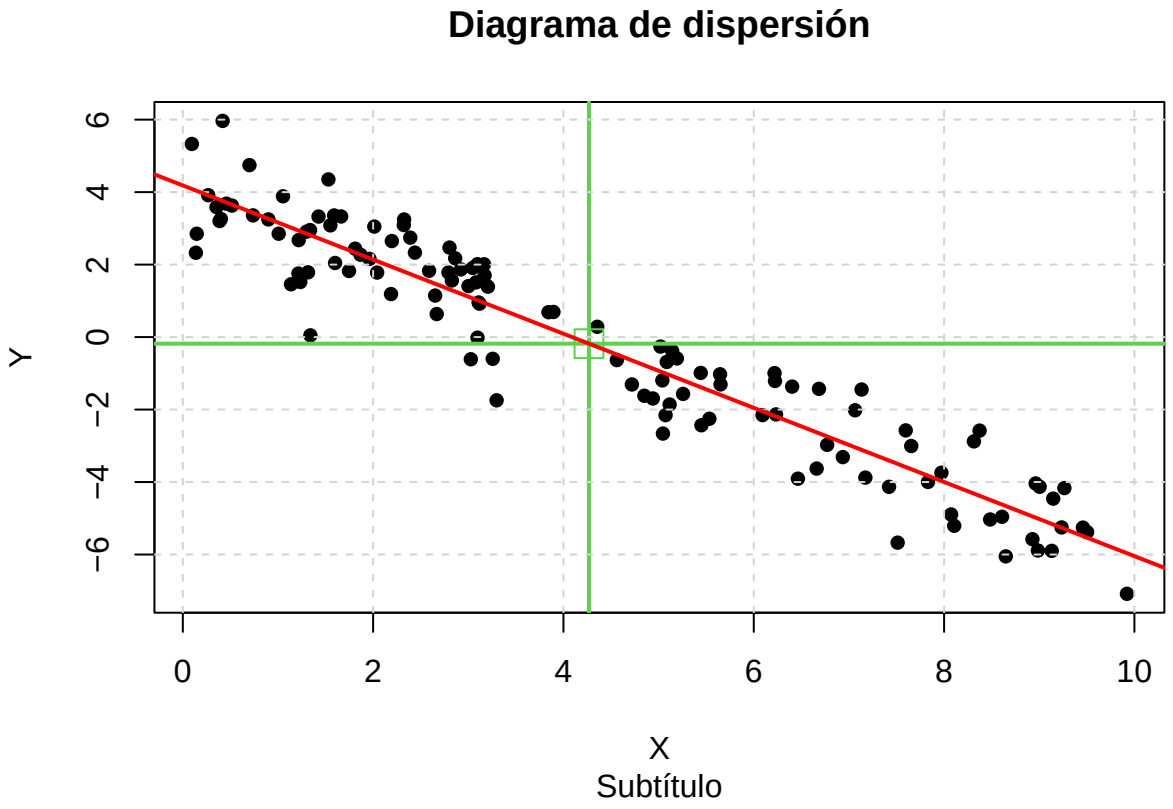
TODO

Ejercicio 7

Las técnicas de inferencia empleadas hasta el momento son ciertas bajo el supuesto de que las 4 hipótesis del modelo de regresión lineal simple (linealidad, homocedasticidad, normalidad e independencia) se verifican. De lo contrario, no todas las interpretaciones obtenidas seguirían siendo válidas. Por ejemplo, si no se cumplieran las hipótesis de homocedasticidad, normalidad e independencia, los intervalos de confianza que hemos obtenido no serían válidos.

Linealidad

En primer lugar, podemos tratar de aventurar si se los datos siguen una tendencia lineal. Emplearemos una aproximación exploratoria, a través de una interpretación gráfica. Para ello, revisitemos la representación previamente definida.



Vemos que los puntos parecen distribuirse en torno a la recta de forma lineal. Si bien hay datos un tanto atípicos, especialmente en los extremos, esto no es lo suficientemente significativo como para rechazar la hipótesis. Tampoco se ve un patrón evidente en los datos (es esto lo que debemos tratar de detectar, y no solo corroborar que haya el mismo número de puntos por encima/debajo de la recta, que no es suficiente como para indicar linealidad).

Nótese que aunque se puede apreciar una menor concentración de puntos para valores de X comprendidos alrededor del valor 4, esto no es indicativo de una falta de linealidad. Dado que trabajamos bajo diseño fijo,

se tiene que achacar a decisiones sobre las condiciones de medición o al propio diseño del experimento. Esta observación se puede comprobar a través del siguiente cuadro:

```
# Representamos el número de valores de X en cada intervalo de longitud 0.5, comenzando desde el mayor
# o igual que el dato mínimo, y finalizando en el menor entero mayor o igual que el dato máximo.
table(cut(X, breaks=seq(from=floor(min(X)), to=ceiling(max(X)), by=0.5)))
```

```
##
##  (0,0.5]  (0.5,1]  (1,1.5]  (1.5,2]  (2,2.5]  (2.5,3]  (3,3.5]  (3.5,4]
##        9        4       11        9        8        8       13        2
##  (4,4.5]  (4.5,5]  (5,5.5]  (5.5,6]  (6,6.5]  (6.5,7]  (7,7.5]  (7.5,8]
##        1        4       11        3        6        4        4        5
##  (8,8.5]  (8.5,9]  (9,9.5] (9.5,10]
##        5        5        6        2
```

Con el objetivo de realizar una prueba más precisa, planteamos el siguiente contraste de hipótesis. Como hipótesis nula tenemos que la variable respuesta siga el modelo lineal simple que hemos estado considerando, y como hipótesis nula, que siga un modelo parabólico, donde hay dependencia de la variable explicativa al cuadrado:

$$\begin{cases} H_0 : Y = \beta_0 + \beta_1 X + \epsilon \\ H_a : Y = \beta_0 + \beta_1 X + \beta_2 * X^2 \epsilon \end{cases}$$

Ejecutamos la prueba:

```
# Empleamos power = 2 porque estamos considerando una alternativa cuadrática
resettest(modelo, power = 2)
```

```
##
##  RESET test
##
## data:  modelo
## RESET = 0.09269, df1 = 1, df2 = 117, p-value = 0.7613
```

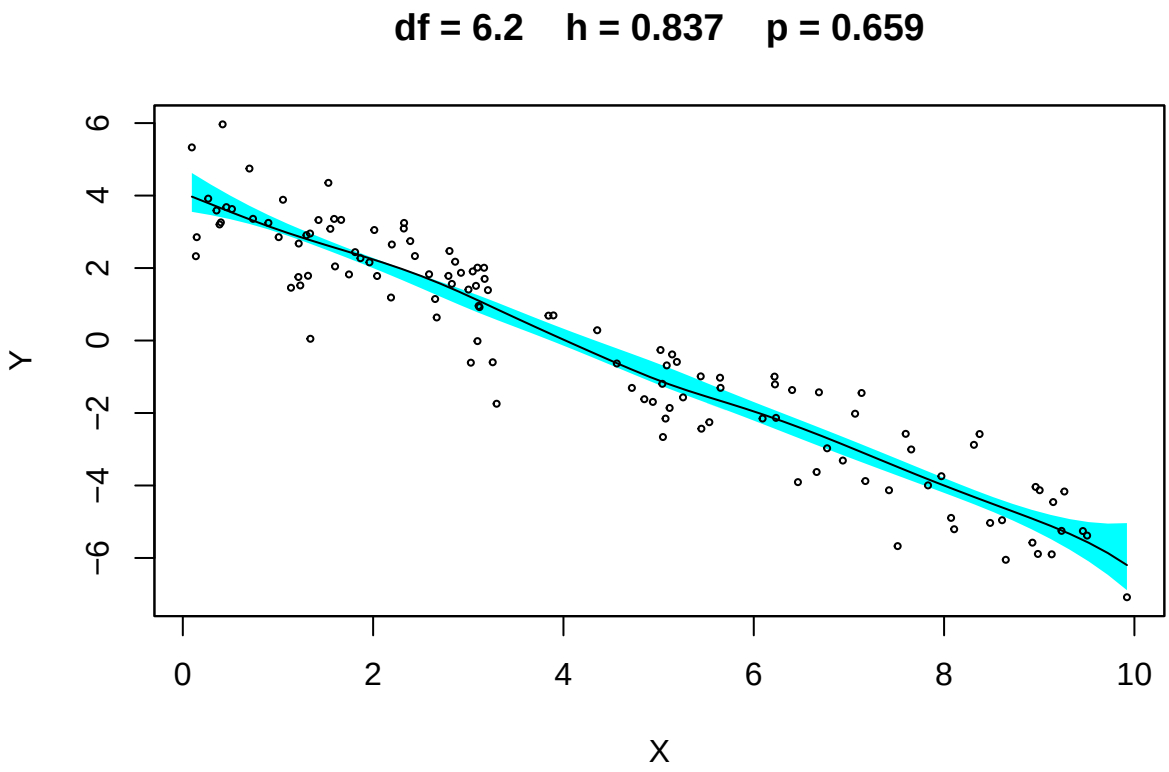
Vemos que el p-valor es de 0.7613. INTERPRETAR.

No obstante, este contraste solo nos ha aportado información sobre la equiparación con un modelo cuadrático. Si buscáramos una confirmación perfecta, teórica, deberíamos seguir contrastando con todos los valores de power. Dado que esto es impracticable experimentalmente, podemos plantearnos en su lugar un contraste más general, con una alternativa no parámetrica:

$$\begin{cases} H_0 : Y = \beta_0 + \beta_1 X + \epsilon \\ H_a : Y = m(X) + \epsilon \end{cases}$$

Haciendo uso del paquete *sm*, realizamos la prueba de hipótesis:

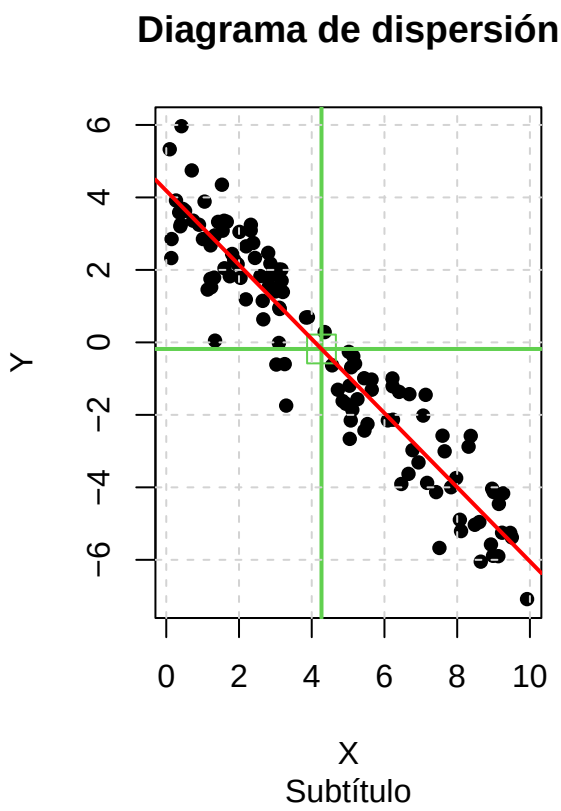
```
# Importamos rpanel para abrir un panel interactivo para la representación
# Los valores que sabemos interpretar son los que aparecen con las opciones por defecto
# Indicamos test=T para que se nos muestre un p-valor.
sm.regression(X, Y, model="linear", panel=T, test=T)
```

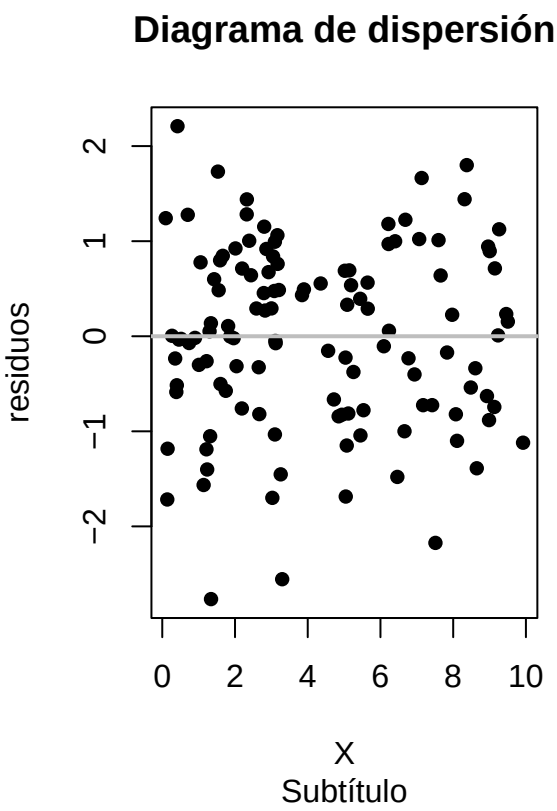


La interpretación de la figura resultante es la siguiente. Con una línea negra nos aparece marcada una estimación no paramétrica de la regresión (sin asumir linealidad), y en azul, una región de confianza para el modelo lineal simple. Vemos que la línea negra se encuentra siempre dentro de la región azul. Por tanto, podemos asumir que la hipótesis nula es cierta, esto es, que los datos verifican la hipótesis de linealidad. FALTA ANALIZAR EL P-VALOR

Homocedasticidad

Contrapongamos ahora los residuos del modelo a la variable explicativa. Se muestra también el diagrama de





dispersión original:

Queremos comprobar ahora si la varianza del error, σ^2, es la misma independientemente del valor que tome la variable explicativa. Vemos que la distribución de los residuos en el diagrama no sigue un patrón evidente, y que su desviación con respecto a la recta $x = 0$ parece ser la misma sin importar el intervalo de X considerado.

Tampoco sobre el diagrama de dispersión de la variable respuesta observamos una tendencia significativa acera de las desviaciones con la recta de regresión. En conjunción con lo anterior, podríamos aventurar, a primera vista, que los datos muestrales son verdaderamente homocedásticos.

Sí destacamos que la interpretación para la región central, en aproximadamente $(4, 4.5)$, puede no ser muy precisa, por falta de datos. Sin embargo, esto no basta para desmentir la hipótesis de homocedasticidad.

Para tener una confirmación precisa, nos planteamos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \text{modelo homocedástico} \\ H_a : \text{modelo heterocedástico} \end{cases}$$

Ejecutamos un test de Harrison-McCabe con R, haciendo uso del previamente cargado paquete *lmtest*:

```
hmctest(Y~X)
```

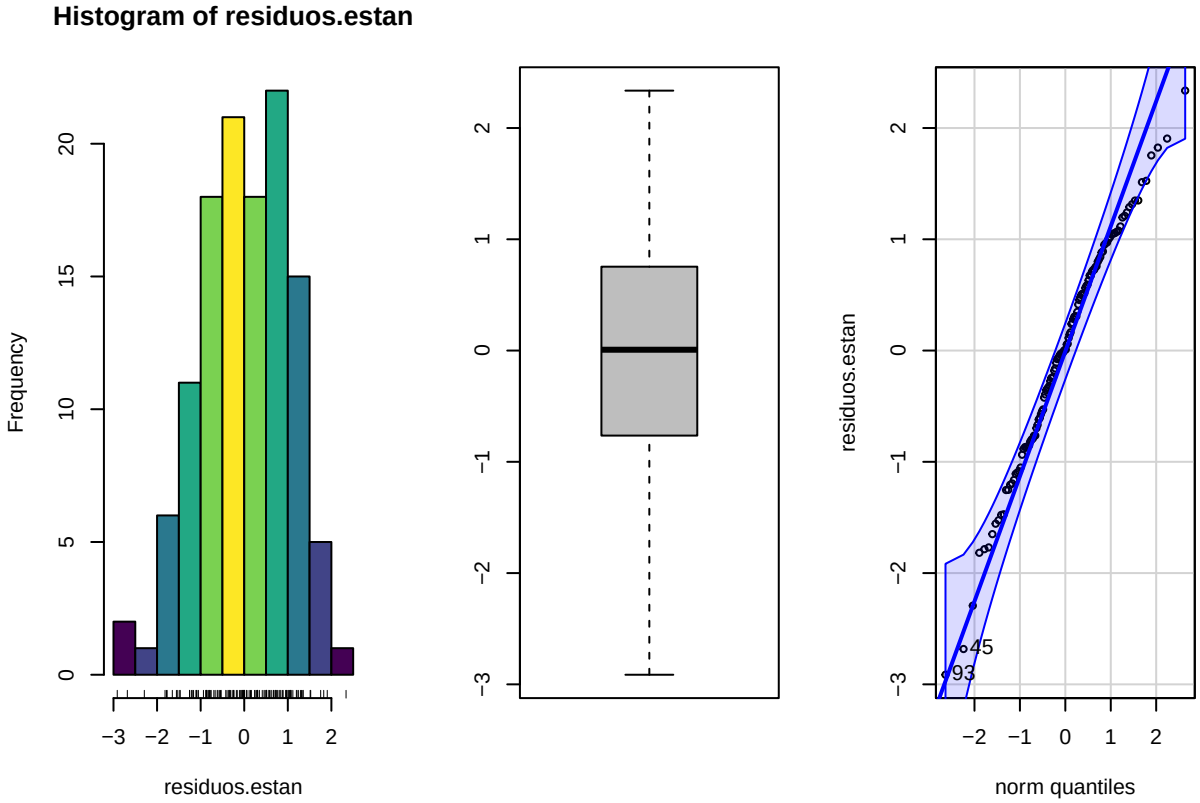
```
##
##  Harrison-McCabe test
##
## data:  Y ~ X
## HMC = 0.55113, p-value = 0.783
```

El p-valor es de 0.763.

Normalidad

Para corroborar que el error tiene distribución normal, haremos varias representaciones gráficas que nos permitan intuir si la hipótesis se ajusta a los datos. Trabajaremos con los residuos estandarizados, pues no tienen la misma varianza y la correlación entre cada 2 de ellos puede ser distinta (provienen de distribuciones diferentes).

Presentamos 3 gráficos: un histograma, un boxplot y un qqplot (para el cual necesitamos la librería *car*), aunque centraremos nuestra atención en el último de ellos, el más relevante en lo que concierne al estudio de la normalidad.




```
## [1] 93 45
```

En el histograma podemos apreciar una cierta asimetría hacia la derecha (valores más altos). En el boxplot o diagrama de caja vemos que la media está centrada en el centro de la caja, un buen indicador. No obstante, la cola izquierda es de una longitud ligeramente mayor, lo cual es indicativo de la asimetría mencionada, al estar los datos más concentrados alrededor de valores más altos.

El QQPlot o diagrama cuantil-cuantil nos presenta una comparativa entre los cuantil muestrales de los residuos estandarizados y los cuantiles teóricos de una normal estándar. Si los residuos estandarizados presentaran una distribución normal de media 0 y varianza 1, se situarían alrededor de la recta diagonal resaltada. En nuestro caso, vemos que en la zona central el ajuste es bueno, pero hay una cierta desviación en las colas. Esto es especialmente notorio en la superior, donde los cuantiles muestrales son algo inferiores a los cuantiles teóricos de una normal, que es lógico y coherente con la asimetría indicada anteriormente.

Ahora bien, una representación visual es solamente un apoyo al estudio, y no podemos inferir de ella una conclusión estadísticamente definitoria. Para ello , emplearemos directamente un test de bondad de ajuste sobre los errores estandarizados con respecto a una distribución normal. Aunque hay varias opciones adecuadas, como el test de Kolmogorov-Smirnov y el test de Lilliefoids, el más ampliamente usado con este propósito es el test de Shapiro-Wilk, especialmente diseñado para contrastes de normalidad:

$$\begin{cases} H_0 : \epsilon \text{ sigue una distribución normal} \\ H_a : \epsilon \text{ no sigue una distribución normal} \end{cases}$$

Ejecutemos pues el contraste de especificación mencionado:

```
shapiro.test(residuos.estan)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos.estan
## W = 0.98768, p-value = 0.3518
```

También podemos comprobar los resultados de otros tests:

```
# TODO FIXME
# el argumento y esta ausente
# ks.test(residuos.estan)
lillie.test(residuos.estan)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuos.estan
## D = 0.057751, p-value = 0.4207
```

Una observación adicional: en este caso, tenemos que el tamaño de la muestra, n, es mayor que 30, de modo que se pueden despreciar las impurezas debidas a utilizar los residuos en el estudio de la normalidad, en lugar de los errores (que no están sujetos a la aplicación del ajuste de mínimos cuadrados).

```
n
```

```
## [1] 120
```

Independencia

De entre las 4 hipótesis con las que trabaja el modelo, la independencia de los errores es la más difícil de corroborar. No tenemos información acerca del proceso de recogida de muestras, por lo que no podemos garantizarla en base a que los datos hayan sido medidos sobre objetos o individuos de forma independiente.

Debido a la complejidad inherente a este apartado, nos limitaremos a comprobar la independencia temporal. Para ello, asumiremos que nuestros datos han sido medidos a lo largo del tiempo.

Nos preguntamos entonces si existe algún tipo de relación entre las observaciones, esto es:

$$\begin{cases} H_0 : \epsilon \text{ son incorrelacionados} \\ H_a : \epsilon \text{ son correlacionados de orden k} \end{cases}$$

En el contraste planteado, $k \in \mathbb{N}$, $k > 1$, es el retardo, esto es, la separación entre los instantes de tiempo que influyen sobre el instante actual. Así, fijado un k y dados unos errores

BONDAD DE AJUSTE ES EL COEFICIENTE DE DETERMINACIÓN (ES UNA MEDIDA DE CUÁNTO DE BUENO ES EL MODELO). EL R^2 AJUSTADO ES OTRA MEDIDA DE BONDAD DE AJUSTE. TAMBIÉN HAY CONTRASTES. ESTÁ EN EL SUMMARY.

ACEPTAR LA H0 EN CONTRASTE LINEALIDAD POLINÓMICO SOLO SIGNIFICA QUE MI MODELO ES MEJOR QUE UN POLINÓMICO DE ORDEN 2, 3...

EL SM TE LO CONTRASTA CON ALTERNATIVA NO PARAMÉTRICA -> ES MEJOR MI MODELO QUE CUALQUIER OTRA COSA? CON HACER ESTE ES SUFICIENTE

a veces las formas raras en homocedasticidad (qqplot) se pueden deber a falta de linealidad. Es más concluyente el contraste

HAY QUE FIJAR EL ALFA DESDE EL PRINCIPIO. METER YA DESDE LA INTRODUCCIÓN. EN ESTE CASO TENDREMOS QUE FIJAR ALFA = 1% (NIVEL 99%) POR EL APARTADO 3