Práctica 5: Validación de modelos de regresión

Módulo de modelos lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

1. Para el modelo lineal que ajustaste en la tarea 1 de la práctica 3 (relación entre `mpg` y `horsepower` para el banco de datos `Auto`) valora su ajuste, validando las hipótesis del modelo de regresión una vez ajustado el modelo. Repite dicha validación para el modelo de la tarea 3 en el que se asumía una relación lineal entre `mpg` y `1/horsepower`.

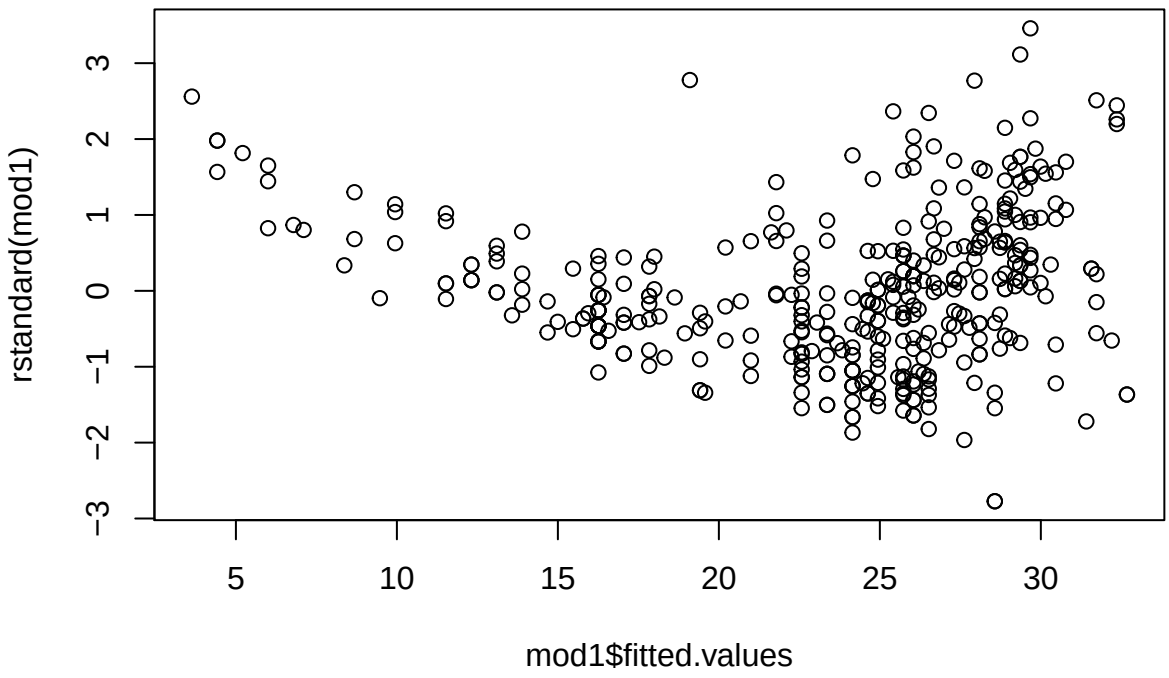
```
data(Auto, package = "ISLR")

# Modelo en función de horsepower
mod1 <- lm(mpg ~ horsepower, data = Auto)
# Valoración de la Normalidad:
ks.test(x = rstudent(mod1), y = "pt", df = dim(Auto)[1] - 3)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(mod1)
## D = 0.060653, p-value = 0.1118
## alternative hypothesis: two-sided
```

```
# Daríamos por buena la hipótesis de normalidad, lo que no significa al fin y al cabo
# que sea cierta ...

# Valoración de la homocedasticidad:
grupos <- cut(mod1$fitted.values, quantile(mod1$fitted.values, (0:4)/4), include.lowest = TRUE)
lawstat::levene.test(rstandard(mod1), grupos)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  rstandard(mod1)
## Test Statistic = 5.9665, p-value = 0.000552
```

```
# La hipótesis de homocedasticidad no se cumple para este modelo de regresión.

# Hipótesis de linealidad
plot(mod1$fitted.values, rstandard(mod1))
```




```
# Evidentemente la hipótesis de linealidad no se cumple. La existencia de tendencia en
# los residuos se podría valorar estadísticamente, y no sólo visualmente, mediante la
# función acf, que veréis con más detalle en series temporales.

# Modelo en función de 1/horsepower
Auto$invhorsepower <- 1/Auto$horsepower
mod2 <- lm(mpg ~ invhorsepower, data = Auto)
# Valoración de la Normalidad:
ks.test(x = rstudent(mod2), y = "pt", df = dim(Auto)[1] - 3)
```

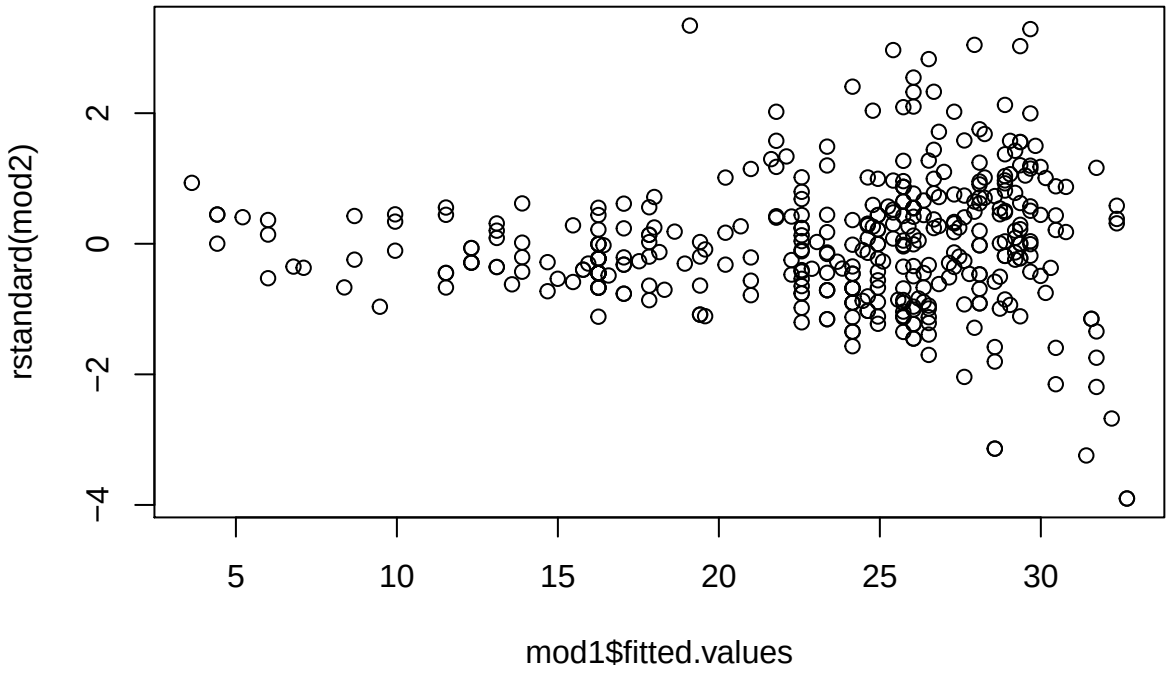
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(mod2)
## D = 0.058367, p-value = 0.1383
## alternative hypothesis: two-sided
```

```
# Daríamos por buena la hipótesis de normalidad, lo que no significa al fin y al cabo
# que sea cierta ...

# Valoración de la homocedasticidad:
grupos <- cut(mod2$fitted.values, quantile(mod2$fitted.values, (0:4)/4), include.lowest = TRUE)
lawstat::levene.test(rstandard(mod2), grupos)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
```

```
##
## data:  rstandard(mod2)
## Test Statistic = 16.215, p-value = 5.992e-10
```

```
# La hipótesis de homocedasticidad no se cumple para este modelo de regresión.

# Hipótesis de linealidad
plot(mod1$fitted.values, rstandard(mod2))
```



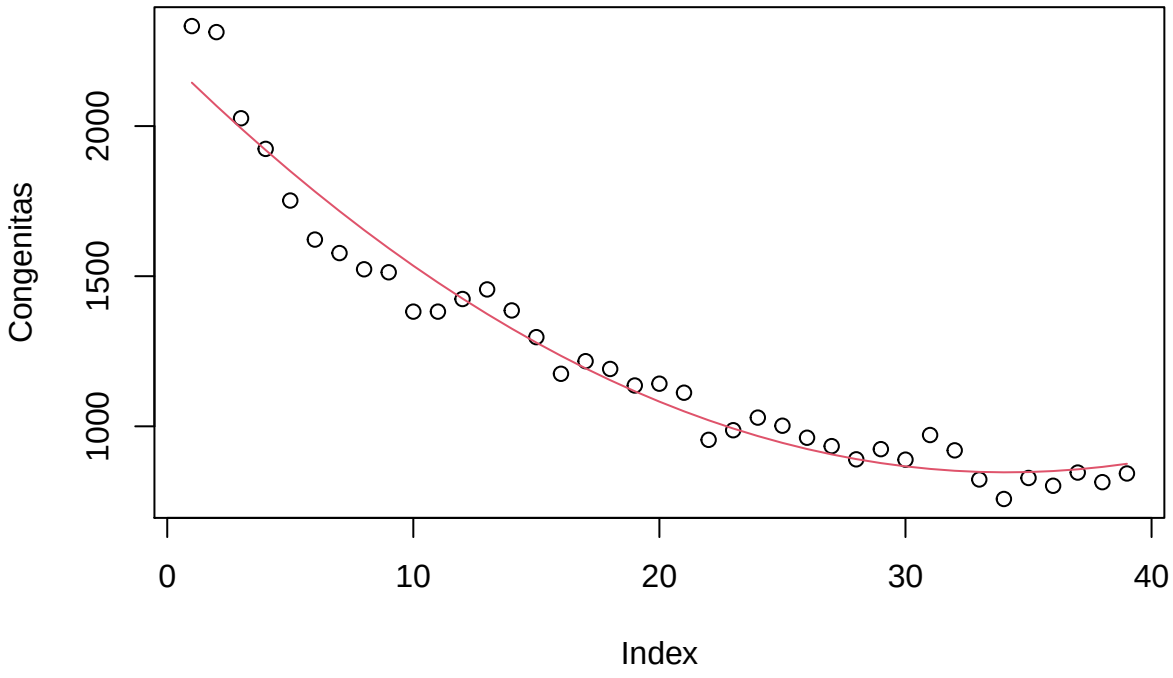

```
# El gráfico representado no muestra evidencia alguna en contra de la hipótesis de
# linealidad, el gráfico no muestra tendencia. Sin embargo la presencia de
# homocedasticidad resulta evidente en este gráfico tal y como ya habíamos demostrado.
# A la vista de los resultados quizás habría sido más conveniente transformar la
# variable respuesta (1/mpg) en lugar de la variable explicativa del modelo.
```

2. El banco de datos `Congenitas.Rdata` contiene las defunciones por enfermedades congénitas ocurridas en España desde el año 1980 hasta 2018, en ese mismo orden. Dado el alto número de defunciones observadas durante el periodo de estudio resulta razonable tratar dicha variable como continua y modelizarla mediante un modelo de regresión lineal Normal.

- Ajusta un modelo de regresión lineal que modeliza las defunciones del banco de datos como función del año del periodo de estudio `x=2:39`. Considera un ajuste cuadrático de esta variable explicativa y representa el ajuste obtenido.

```
load("../Datos/Congenitas.Rdata")
plot(Congenitas)

# Ajuste del modelo cuadrático
```

```r
x <- 1:39
ajuste2 <- lm(Congenitas ~ poly(x, 2))
lines(x = ajuste2$fitted.values, col = 2)
```




```r
summary(ajuste2)
```

```
##
## Call:
## lm(formula = Congenitas ~ poly(x, 2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -160.140  -55.304    3.961   51.815  245.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1232.31      14.13  87.187  < 2e-16 ***
## poly(x, 2)1 -2347.96      88.27 -26.600  < 2e-16 ***
## poly(x, 2)2   837.42      88.27   9.487 2.49e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.27 on 36 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9544
## F-statistic: 398.8 on 2 and 36 DF,  p-value: < 2.2e-16
```

- Evalúa la hipótesis de Normalidad para el modelo que acabas de ajustar.

```r
ks.test(x = rstudent(ajuste2), y = "pt", df = 35)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(ajuste2)
## D = 0.11519, p-value = 0.6369
## alternative hypothesis: two-sided
```

```r
# El modelo cumple la hipótesis de Normalidad
```

- Evalúa la hipótesis de homocedasticidad de los datos alrededor de la curva ajustada ¿a qué crees que se puede deber la heterocedasticidad de los datos?

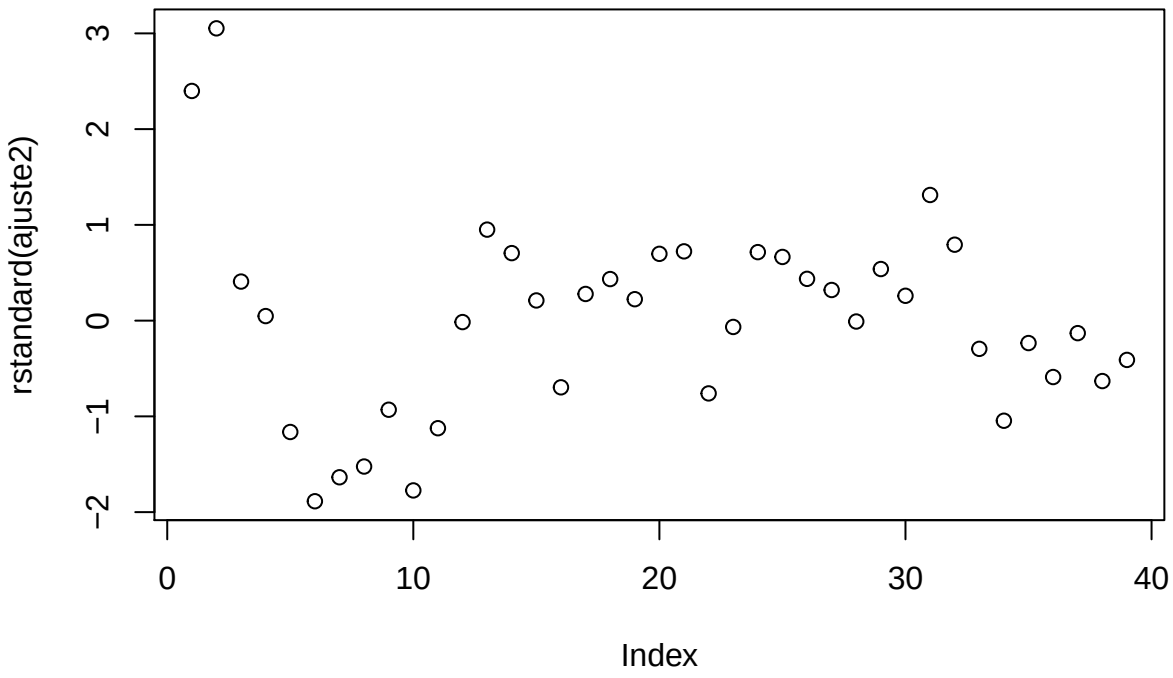
```r
grupos <- cut(ajuste2$fitted.values, quantile(ajuste2$fitted.values, (0:4)/4), include.lowest = TRUE)
lawstat::levene.test(rstandard(ajuste2), grupos)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  rstandard(ajuste2)
## Test Statistic = 3.009, p-value = 0.04316
```

```r
# Los datos no cumplen la hipótesis de homocedasticidad. Esto se puede deber a que, al
# tratarse de conteos, la distribución de Poisson podría ser más adecuada que la
# Normal para modelizar los datos. Pero para datos de tipo Poisson sabemos que su
# media y su varianza coinciden, ese puede ser el motivo por el que no resulta
# admisible la hipótesis de homocedasticidad para este modelo de regresión.
```

- Representa los residuos estandarizados y valora sobre dicha representación la adecuación de la hipótesis de linealidad de la tendencia conforme avanza el periodo de estudio.

```r
plot(rstandard(ajuste2))
```




```
# Parece haber cierta tendencia en los residuos, particularmente evidentemente en la
# parte final e inicial del periodo de estudio.

# La mayor varianza residual en la parte inicial del periodo de estudio (con mayor
# número de casos observados) parece confirmar que la heterocedasticidad se puede
# deber al carácter Poisson de los datos.
```

- Considera un ajuste polinómico de orden superior, hasta el grado que consideres oportuno. Representa el ajuste del nuevo modelo que hayas ajustado.

```
ajuste8 <- lm(Congenitas ~ poly(x, 8))
summary(ajuste8)
```

```
##
## Call:
## lm(formula = Congenitas ~ poly(x, 8))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.480 -23.527  -2.745  28.518  78.698
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1232.308      7.621 161.702  < 2e-16 ***
## poly(x, 8)1 -2347.965     47.592 -49.335  < 2e-16 ***
## poly(x, 8)2   837.425     47.592  17.596  < 2e-16 ***
## poly(x, 8)3  -300.848     47.592  -6.321 5.70e-07 ***
```

```
## poly(x, 8)4   256.714     47.592   5.394 7.66e-06 ***
## poly(x, 8)5  -146.247     47.592  -3.073  0.00448 **
## poly(x, 8)6    74.768     47.592   1.571  0.12667
## poly(x, 8)7   122.944     47.592   2.583  0.01491 *
## poly(x, 8)8  -118.435     47.592  -2.489  0.01860 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.59 on 30 degrees of freedom
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.9867
## F-statistic: 354.7 on 8 and 30 DF,  p-value: < 2.2e-16
```

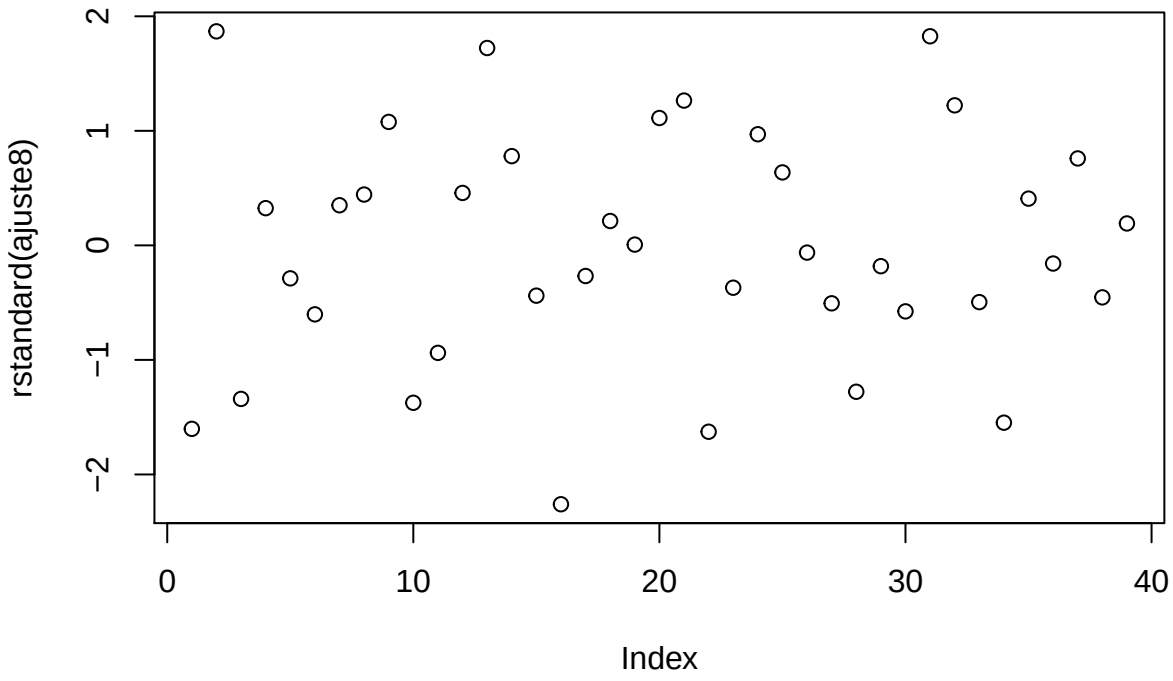
```
# Se puede comprobar facilmente que los coeficientes de polinomios de grado superior
# apenas son significativos, por tanto daríamos por bueno un ajuste polinómico de
# grado 8.
```

- ¿Desaparecen los problemas de homoscedasticidad y tendencia de los residuos para este nuevo modelo?

```
grupos8 <- cut(ajuste8$fitted.values, quantile(ajuste8$fitted.values, (0:4)/4), include.lowest = TRUE)
lawstat::levene.test(rstandard(ajuste8), grupos8)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  rstandard(ajuste8)
## Test Statistic = 0.14369, p-value = 0.933
```

```
# Los problemas de homocedasticidad parecen haber desaparecido, posiblemente a costa
# de sobreajustar las observaciones que anteriormente presentaban más varianza.
plot(rstandard(ajuste8))
```

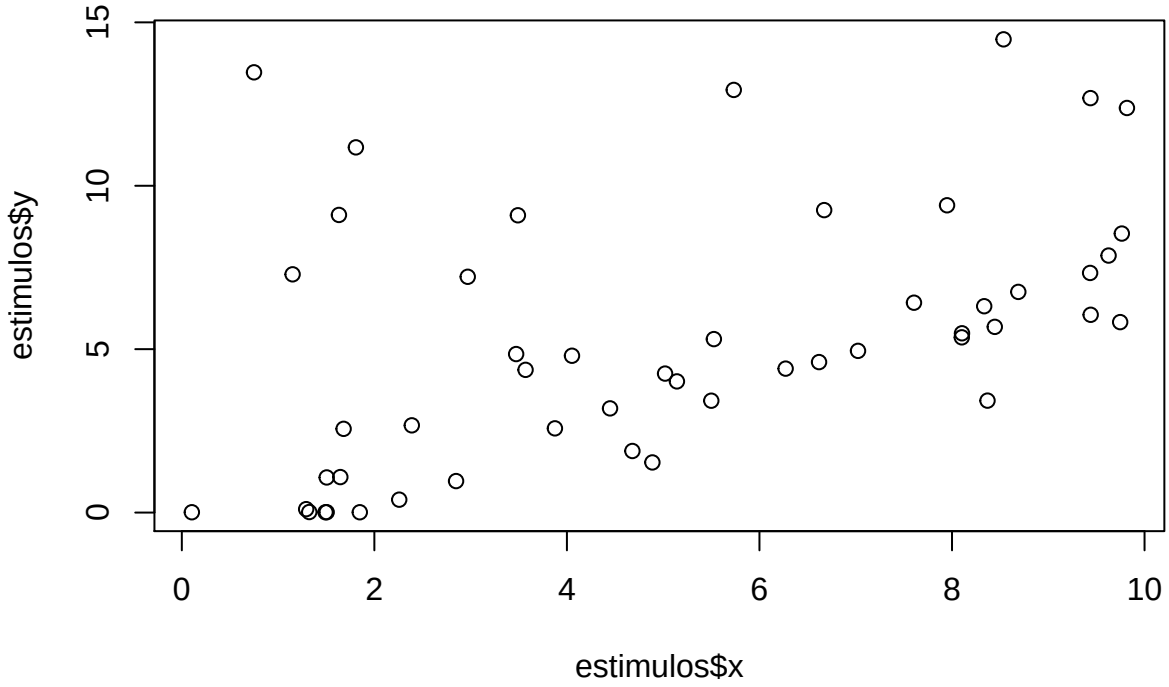



```
# Los problemas de tendencia en el residuo parecen haber desaparecido también.
```

3. El banco de datos `estimulos.Rdata` contiene dos variables, ambas son el tiempo de respuesta del cerébro ante distintos estimulos. La variable `x` es el tiempo de respuesta a un sonido acompañado de un estímulo visual y la variable `y` es el tiempo de respuesta únicamente al estímulo visual, para una serie de 50 individuos.

- Representa gráficamente ambas variables.

```
load("../Datos/estimulos.Rdata")
plot(estimulos$x, estimulos$y)
```



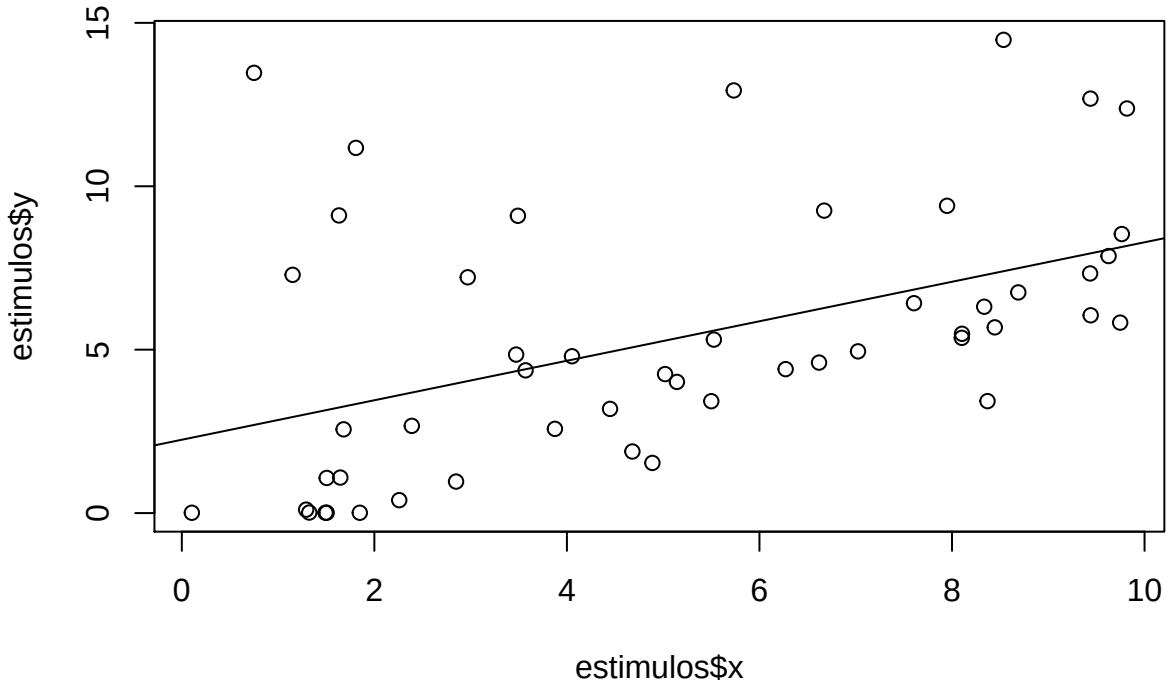
- Ajusta un modelo de regresión lineal para explicar la variable `y` en función de `x`. Comprueba que la relación funcional entre ambas variables es la adecuada y que no resulta necesario incluir un polinomio de mayor grado en función de `x`.

```
ajuste <- lm(y ~ x, data = estimulos)
summary(lm(y ~ poly(x, 5), data = estimulos))
```

```
##
## Call:
## lm(formula = y ~ poly(x, 5), data = estimulos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6329 -2.4672 -0.9713  1.1755  9.8251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.3322     0.5097  10.461 1.63e-13 ***
## poly(x, 5)1  12.9959     3.6041   3.606  0.00079 ***
## poly(x, 5)2   3.1861     3.6041   0.884  0.38150
## poly(x, 5)3  -0.3092     3.6041  -0.086  0.93202
## poly(x, 5)4   0.4110     3.6041   0.114  0.90972
## poly(x, 5)5   1.7633     3.6041   0.489  0.62711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.604 on 44 degrees of freedom
```

```
## Multiple R-squared:  0.2419, Adjusted R-squared:  0.1558
## F-statistic: 2.809 on 5 and 44 DF,  p-value: 0.02756
```

```
# No parece necesitarse un polinomio de mayor grado para mejorar el ajuste
plot(estimulos$x, estimulos$y)
abline(ajuste$coef)
```

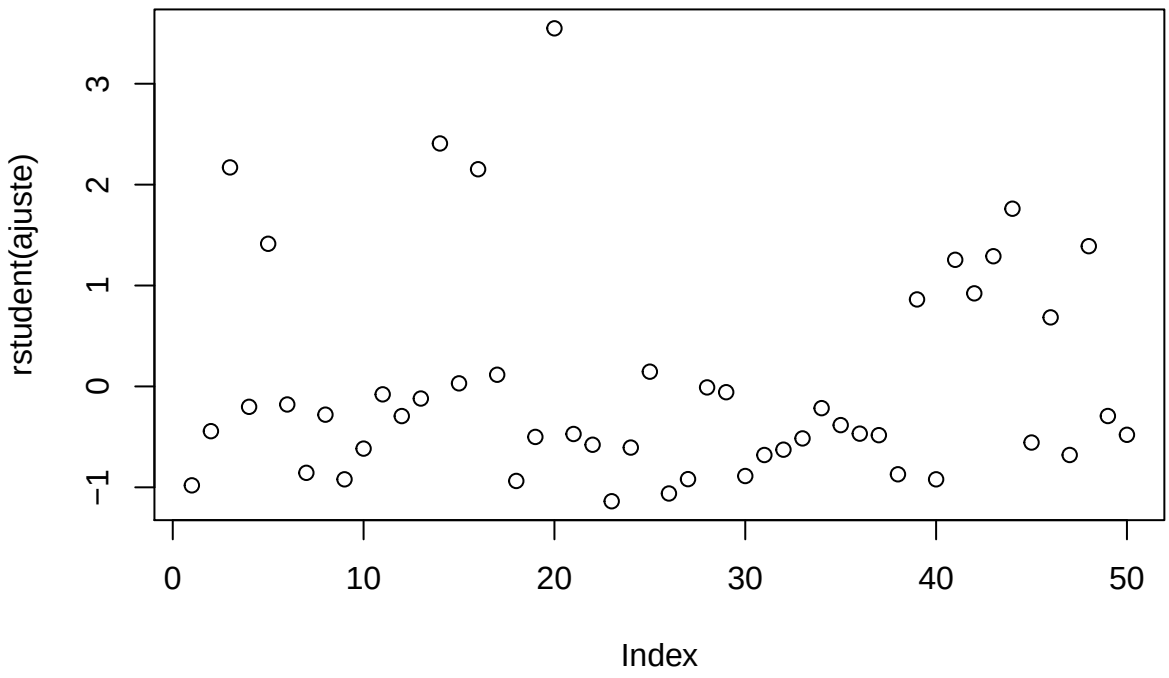


- Evalua la hipótesis de Normalidad para el modelo que acabas de ajustar.

```
ks.test(x = rstudent(ajuste), y = "pt", df = 47)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(ajuste)
## D = 0.20793, p-value = 0.02252
## alternative hypothesis: two-sided
```

```
plot(rstudent(ajuste))
```



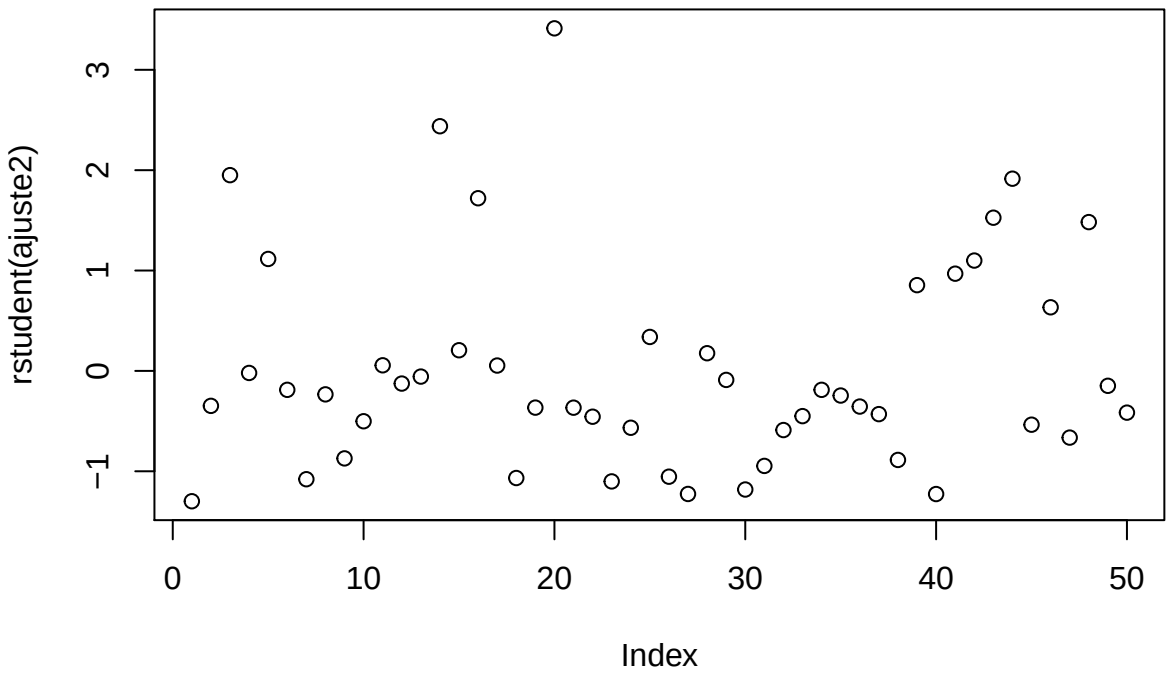

```
# La hipótesis de Normalidad no se cumple para este modelo, posiblemente por la
# asimetría de los residuos alrededor de la recta de regresión. Los valores por encima
# de la recta son más esporádicos, pero más distantes a la recta de regresión.
```

- Determina la transformación de Box-Cox que mejoraría en mayor medida la Normalidad de la variable respuesta `y`. Ajuste de nuevo un nuevo modelo de regresión lineal sobre la nueva variable transformada ¿Observas mejora en cuanto a la hipótesis de Normalidad que habías evaluado anteriormente?

```
# Transformación de Boc-Cox óptima para la variable respuesta
lambda <- EnvStats::boxcox(estimulos$y, lambda = c(-2, 2), optimize = TRUE)$lambda
estimulos$y2 <- (estimulos$y^lambda - 1)/lambda
ajuste2 <- lm(y2 ~ x, data = estimulos)
ks.test(x = rstudent(ajuste2), y = "pt", df = 47)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(ajuste2)
## D = 0.1779, p-value = 0.07429
## alternative hypothesis: two-sided
```

```
plot(rstudent(ajuste2))
```




```
# Tras esta transformación los residuos, además de la variable respuesta, ya se pueden
# considerar Normales, por tanto este modelo satisfaría la hipótesis de Normalidad.
# Los residuos de la variable transformada son algo más simétricos que los de la
# variable original.
```