LAB5- Ana Royuela-Mayo2018

Ana Royuela

2018-05-10

Importing dataset

```
DataReg <- read.table("C:/Users/Ana/Documents/UOC/SOFTWARE ANALISIS DATOS/M
  na.strings="NA", dec=".", strip.white=TRUE)
```

Names of variables

```
names(DataReg)
```

```
## [1] "weight" "age"    "fats"
```

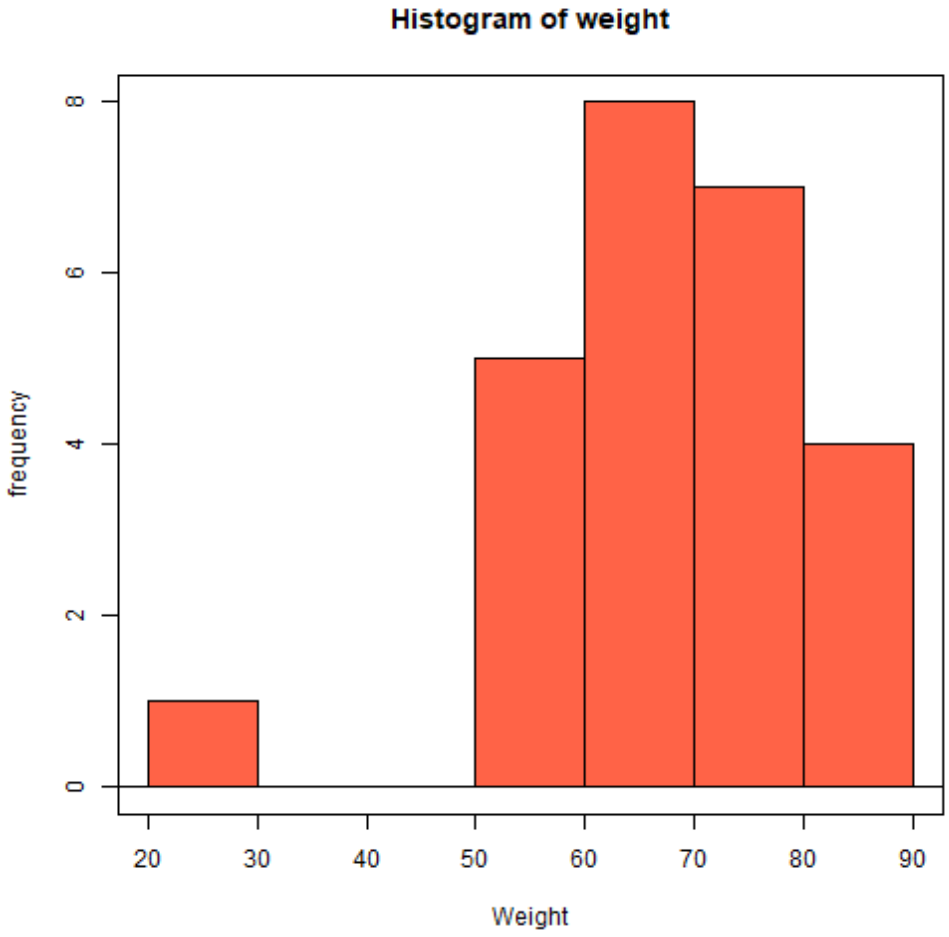
Summary of data

```
summary(DataReg)
```

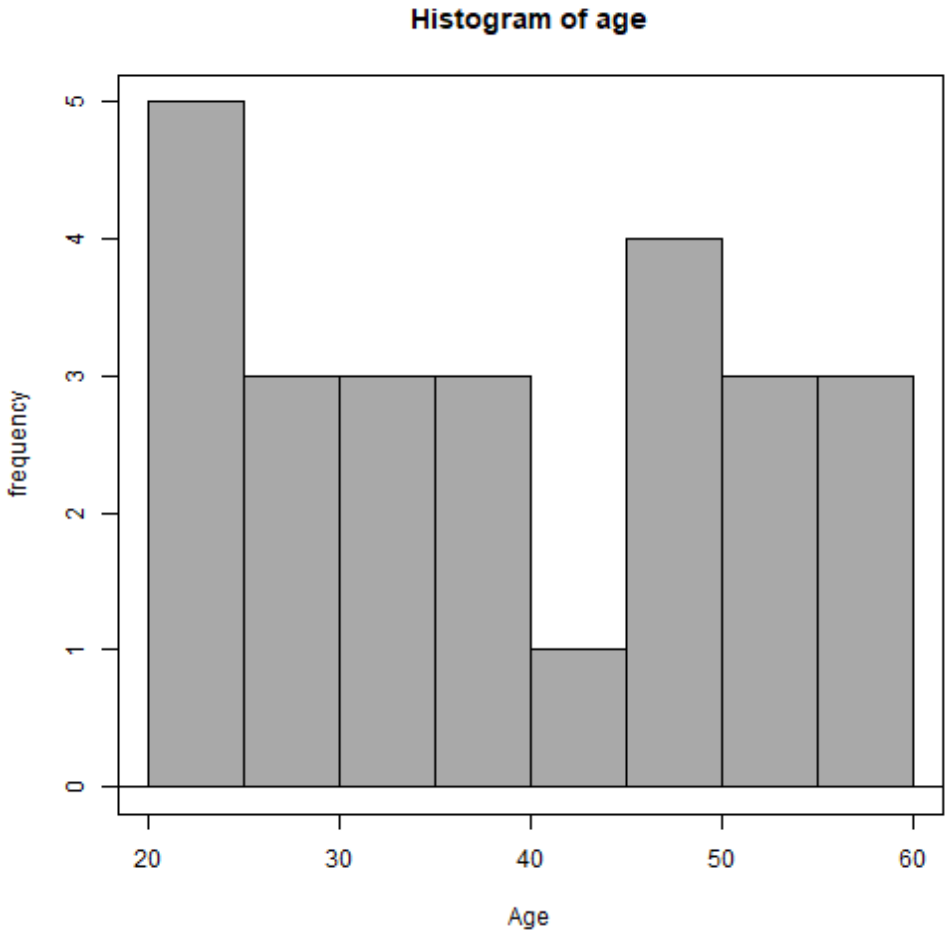
weight	age	fats
Min. :27.00	Min. :20.00	Min. :181.0
1st Qu.:63.00	1st Qu.:30.00	1st Qu.:254.0
Median :69.00	Median :37.00	Median :303.0
Mean :68.68	Mean :39.12	Mean :310.7
3rd Qu.:76.00	3rd Qu.:50.00	3rd Qu.:374.0
Max. :89.00	Max. :60.00	Max. :451.0

Histograms of data

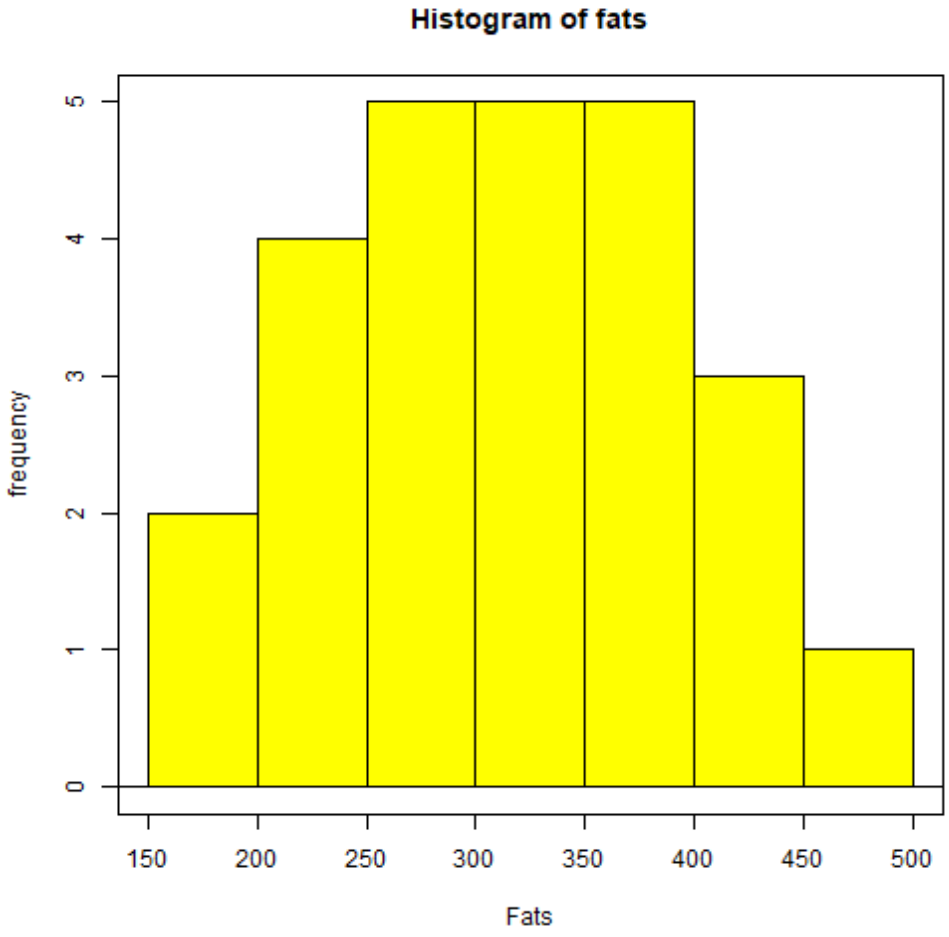
```
with(DataReg, Hist(weight, scale="frequency", breaks="Sturges", col="tomat
```




```
with(DataReg, Hist(age, scale="frequency", breaks="Sturges", col="darkgray'
```

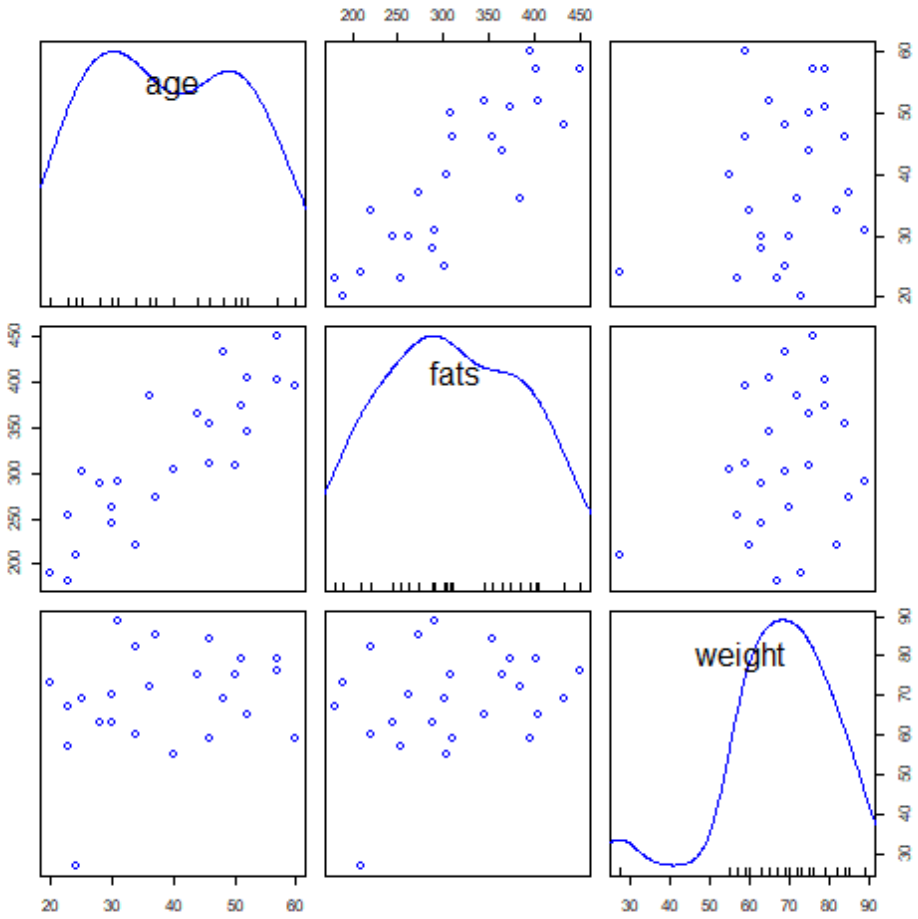



```
with(DataReg, Hist(fats, scale="frequency", breaks="Sturges", col="yellow",
```

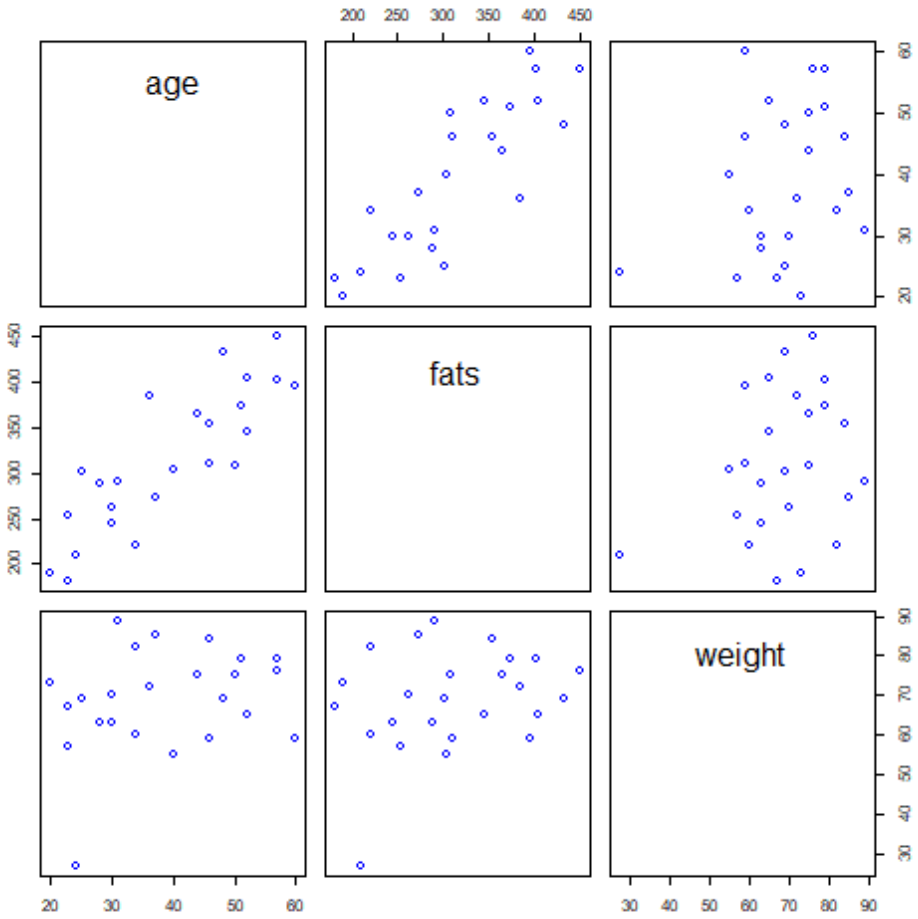


Scatterplot matrix de las tres variables:

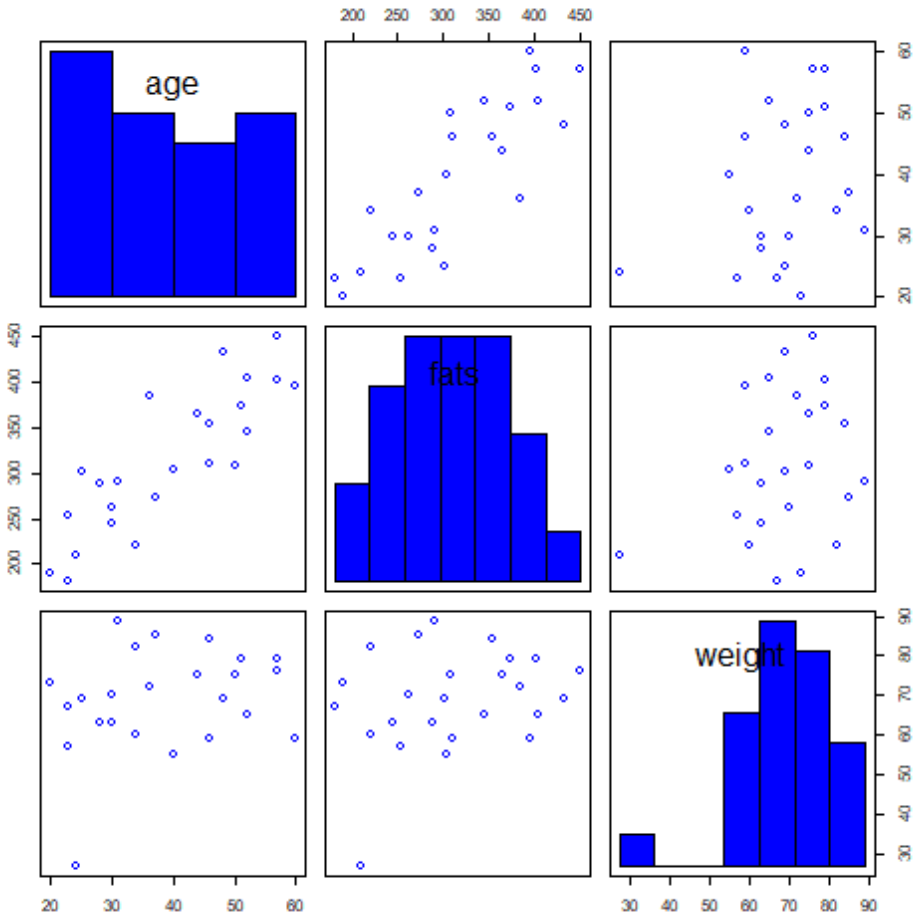
```
scatterplotMatrix(~age+fats+weight, regLine=FALSE, smooth=FALSE, diagonal=
```




```
scatterplotMatrix(~age+fats+weight, regLine=FALSE, smooth=FALSE, diagonal=
```

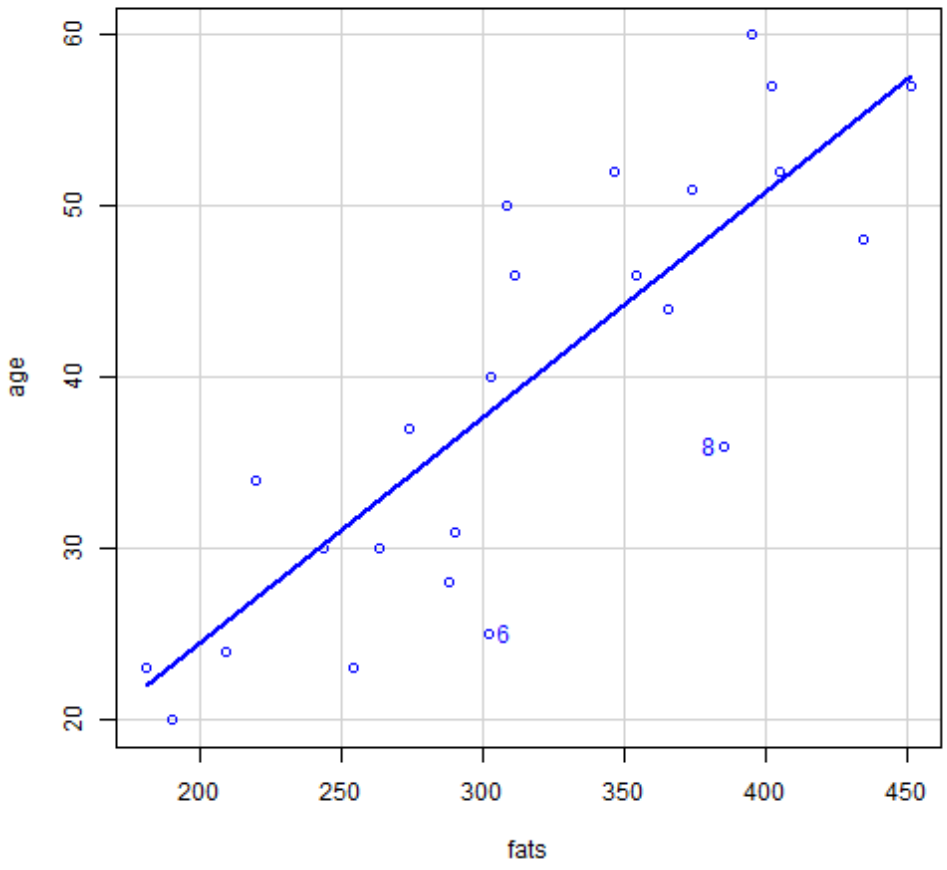



```
scatterplotMatrix(~age+fats+weight, regLine=FALSE, smooth=FALSE, diagonal=
```



Scatterplot between age and fat

```
scatterplot(age~fats, regLine=TRUE, smooth=FALSE, id=list(method='mahal', r
```




```
## [1] 6 8
```

Matriz de correlación de las tres variables

```
cor(DataReg[,c("age","fats","weight")], use="complete")
```

	age	fats	weight
age	1.0000000	0.8373534	0.2400133
fats	0.8373534	1.0000000	0.2652935
weight	0.2400133	0.2652935	1.0000000

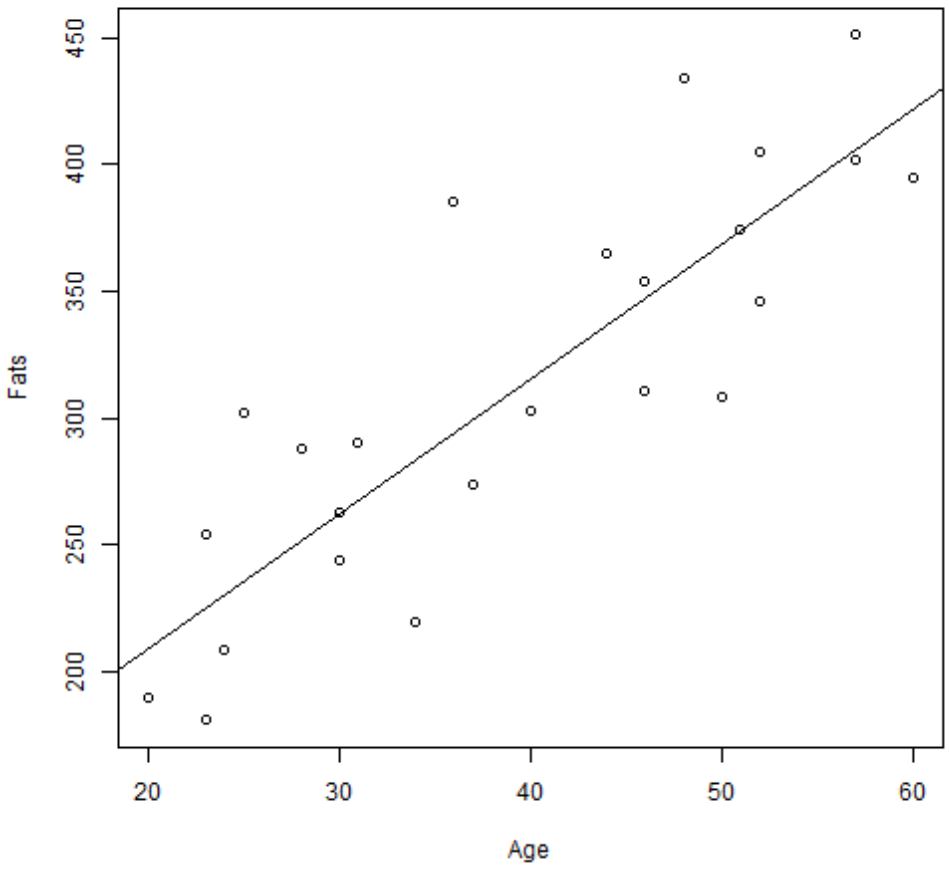
Modelo de regresión lineal entre fats y age

```
RegModel.1 <- lm(fats~age, data=DataReg)
summary(RegModel.1)
```

```
##
## Call:
## lm(formula = fats ~ age, data = DataReg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.478 -26.816  -3.854  28.315  90.881
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 102.5751    29.6376   3.461     0.00212 **
## age           5.3207     0.7243   7.346 0.000000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.46 on 23 degrees of freedom
## Multiple R-squared:  0.7012, Adjusted R-squared:  0.6882
## F-statistic: 53.96 on 1 and 23 DF,  p-value: 0.0000001794
```

Gráfica del modelo

```
plot(DataReg$age, DataReg$fats, xlab="Age", ylab="Fats")
abline(RegModel.1)
```



Generating predictions from a new dataset

```
newages<-data.frame(age=seq(30,50))
predict(RegModel.1, newages)
```

```
##        1        2        3        4        5        6        7        8
## 262.1954 267.5161 272.8368 278.1575 283.4781 288.7988 294.1195 299.4402
##        9       10       11       12       13       14       15       16
## 304.7608 310.0815 315.4022 320.7229 326.0435 331.3642 336.6849 342.0056
##       17       18       19       20       21
## 347.3263 352.6469 357.9676 363.2883 368.6090
```

Intervalos de confianza al 95%

```
Confint(RegModel.1, level=0.95)
```

	Estimate	2.5 %	97.5 %
(Intercept)	102.575142	41.265155	163.885130
age	5.320676	3.822366	6.818986

Intervalos de confianza al 90%

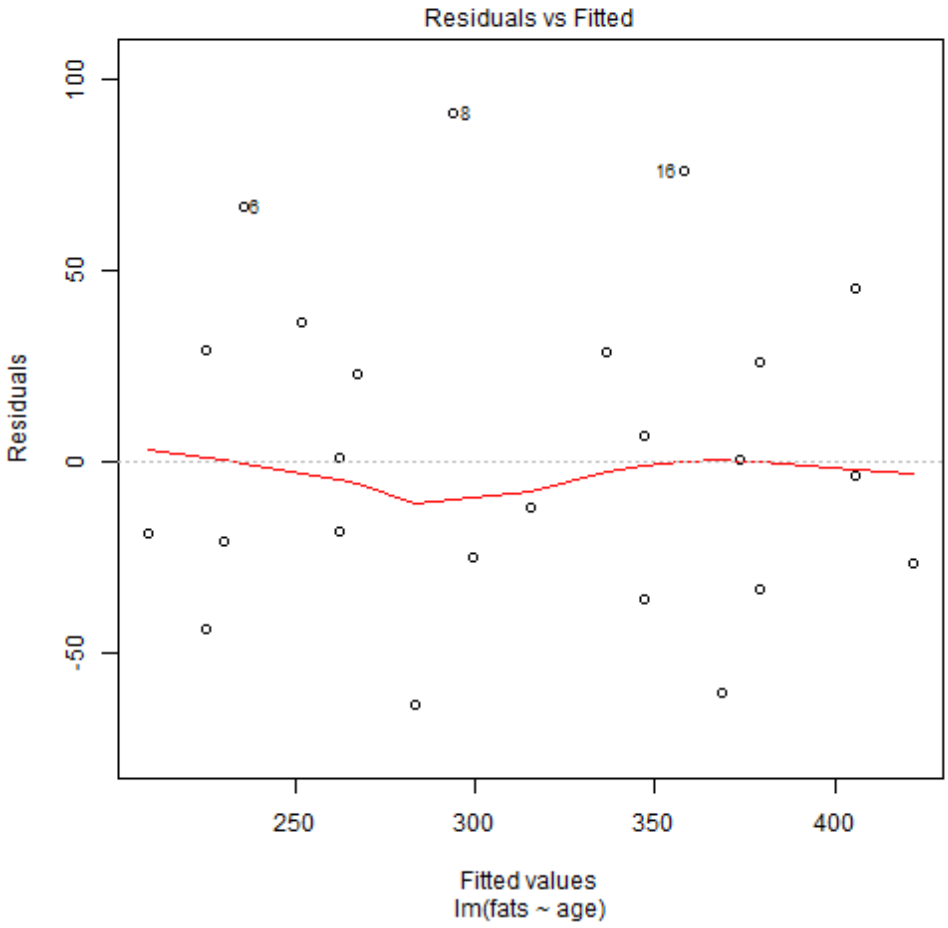
```
Confint(RegModel.1, level=0.90)
```

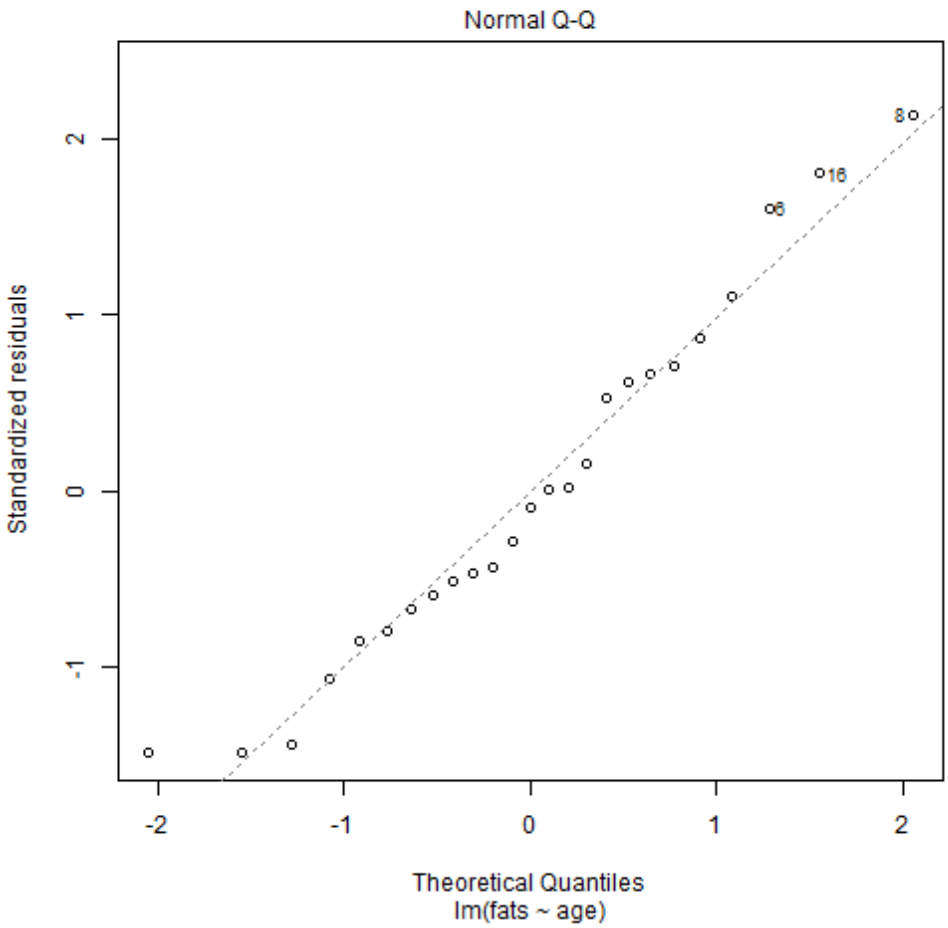
	Estimate	5 %	95 %
(Intercept)	102.575142	51.780153	153.370132
age	5.320676	4.079335	6.562018

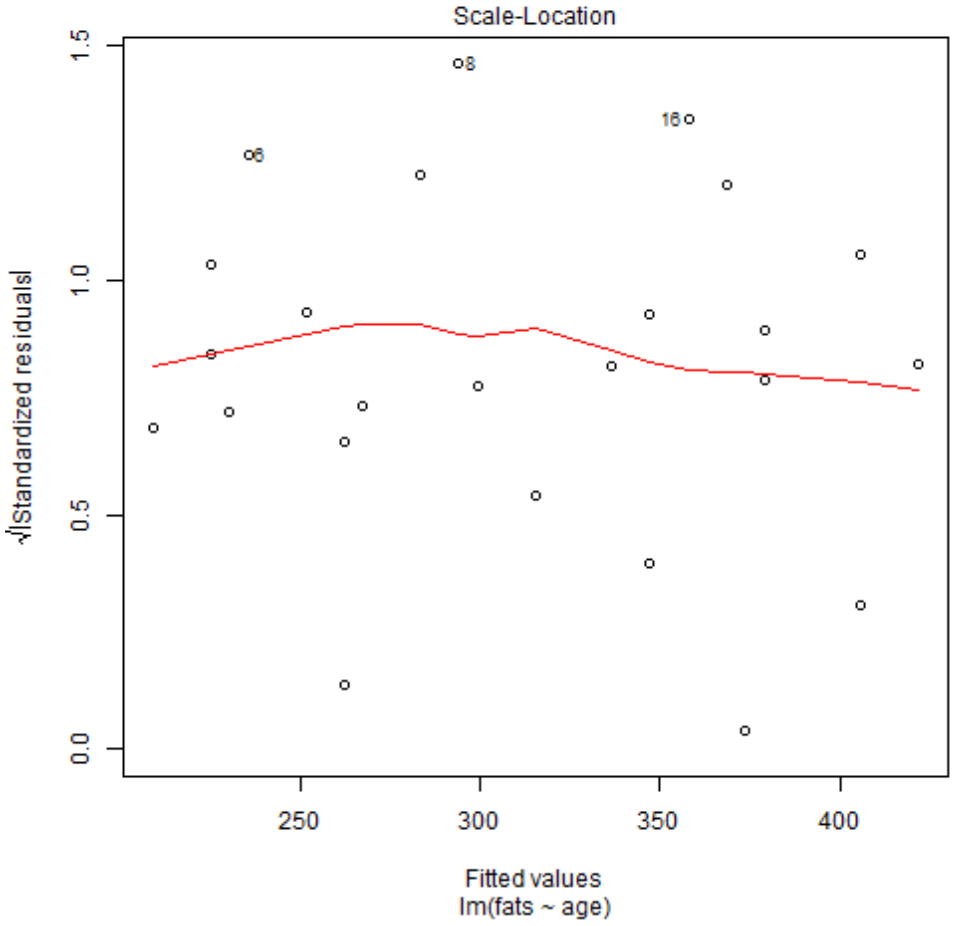
Diagnosis Model

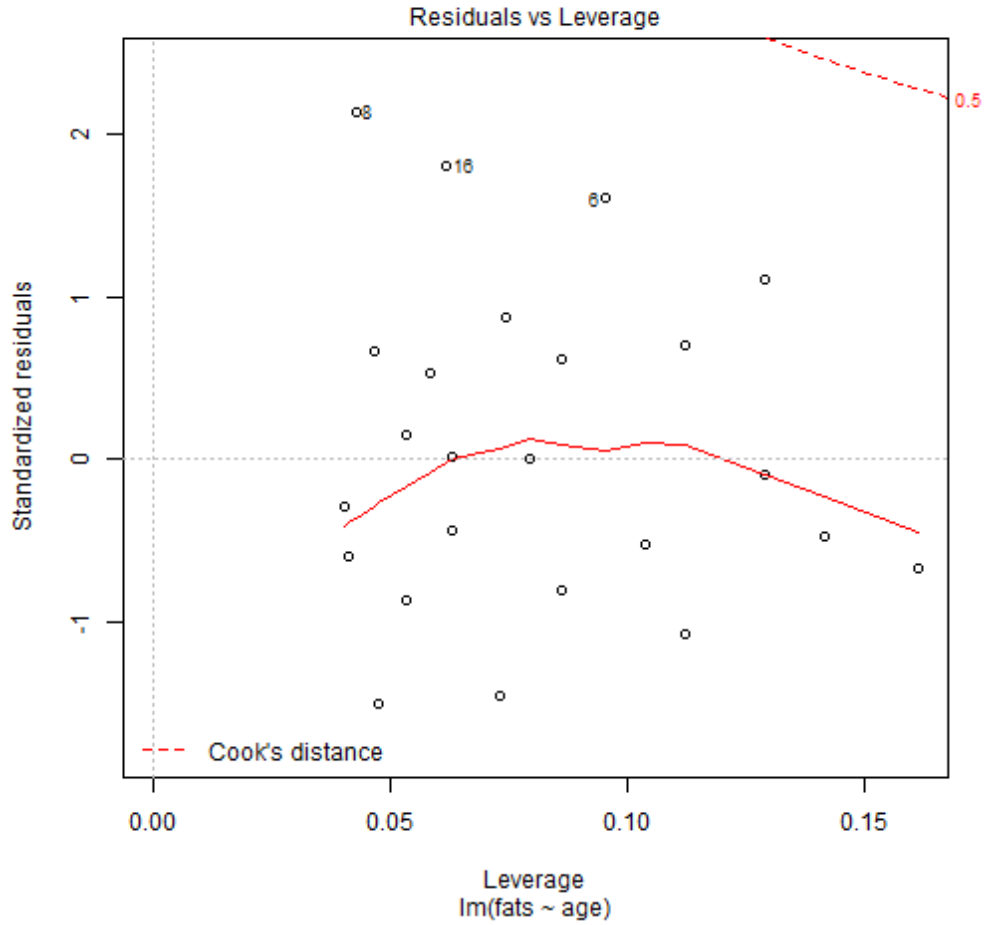
```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
```

```
plot(RegModel.1)
```


Residuals vs Fitted
Residuals
100
50
0
-50
8
16
6
250
300
350
400
Fitted values
lm(fats ~ age)

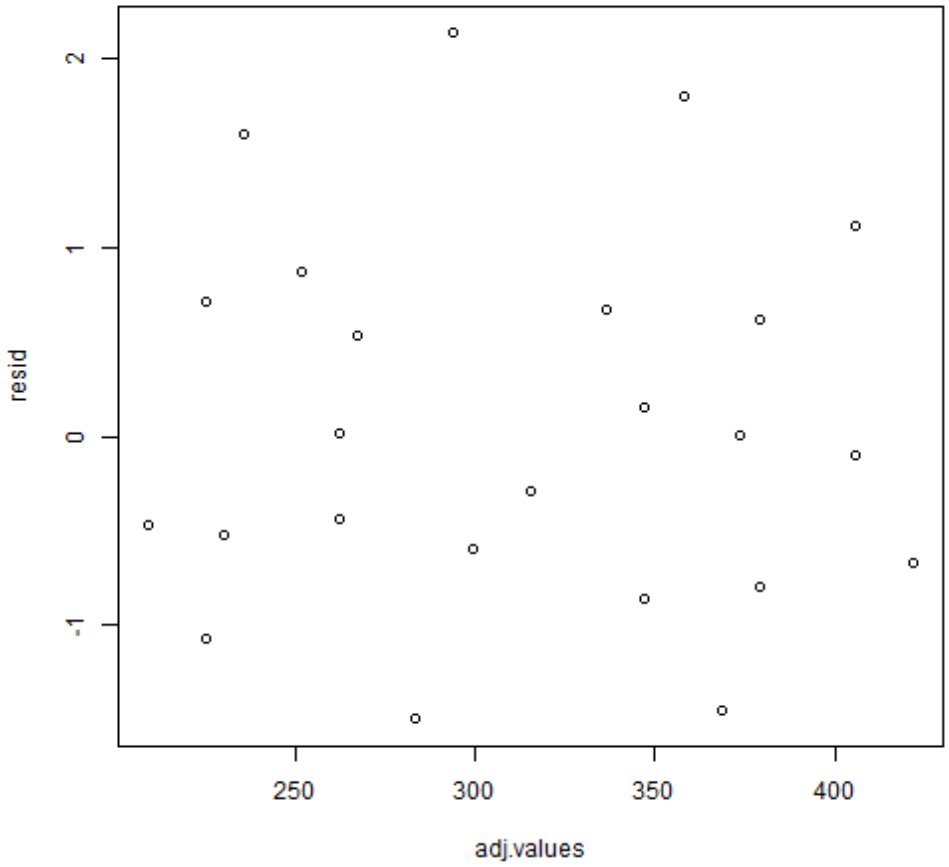

Normal Q-Q
Standardized residuals
2
1
0
-1
8
16
6
-2
-1
0
1
2
Theoretical Quantiles
lm(fats ~ age)


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
8
16
6
250
300
350
400
Fitted values
lm(fats ~ age)


Residuals vs Leverage
Standardized residuals
2
1
0
-1
8
16
6
0.5
Cook's distance
0.00
0.05
0.10
0.15
Leverage
lm(fats ~ age)

```
par(oldpar)
```

```
resid<-rstandard(RegModel.1)
adj.values<-fitted(RegModel.1)
plot(adj.values,resid)
```



Exercise 2

1. With the same data set (previous exercise), create a model that explains the relation between fat and weight.

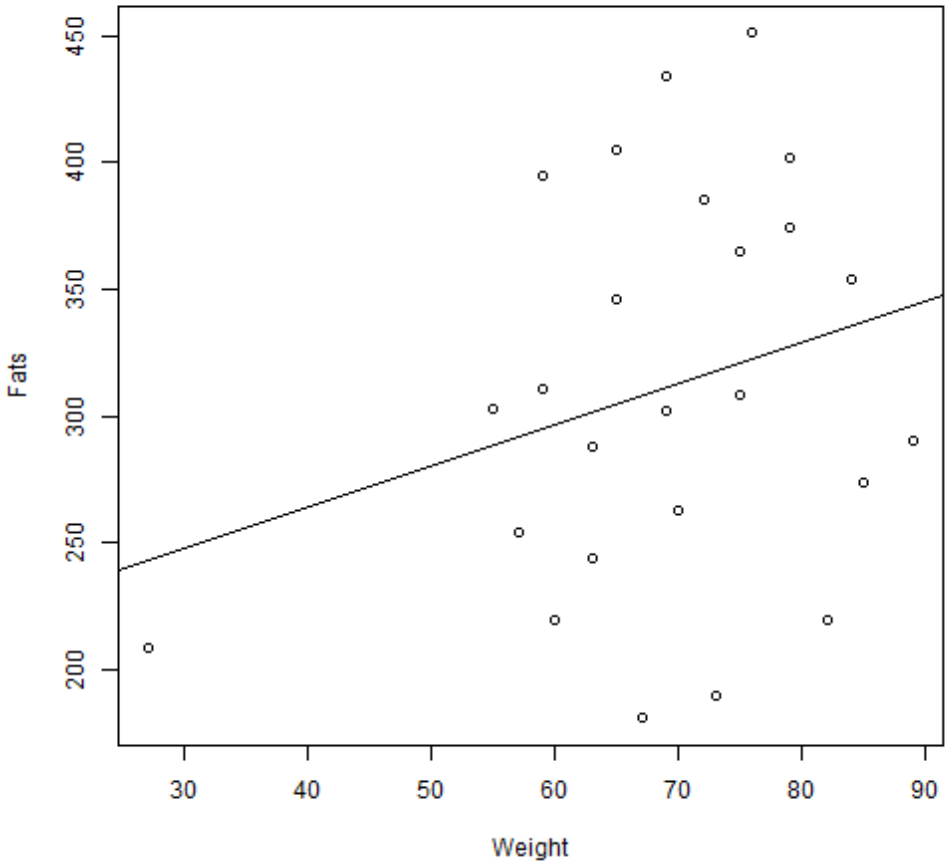
```
RegModel.2 <- lm(fats~weight, data=DataReg)
summary(RegModel.2)
```

```
##
## Call:
## lm(formula = fats ~ weight, data = DataReg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.729  -53.686   -9.239   46.537  128.404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  199.298     85.818   2.322   0.0294 *
## weight         1.622      1.229   1.320   0.2000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.65 on 23 degrees of freedom
## Multiple R-squared:  0.07038,    Adjusted R-squared:  0.02996
## F-statistic: 1.741 on 1 and 23 DF,  p-value: 0.2
```

En este modelo, no se observa asociación entre fats and weight, ya que el coeficiente de weight no es estadísticamente diferente de 0.

2. Calculate and graph the regression line, together with the corresponding point cloud.

```
plot(DataReg$weight, DataReg$fats, xlab="Weight", ylab="Fats")
abline(RegModel.2)
```



3. What is the squared correlation coefficient in this case?

El coeficiente de determinación es igual a 0.07038, mucho menor que en el caso de la edad. La variabilidad de las grasas explicada por el peso es del 7%.

4. What are the estimate parameters of the model?

El intercept es 199.298, es decir, que cuando el peso es igual a 0, el valor de las grasas es 199.298. Por cada unidad de peso, la grasa aumenta en promedio 1.622 unidades.

5. Test the hypothesis that the slope of the line is 0 to 0.05 level. (Note: R?Commander has a specific menu.)

La hipótesis de que la pendiente de la línea es 0, es igual a contrastar si el coeficiente de weight es =0. La p asociada a ese contraste es 0.200, por tanto no se puede rechazar la hipótesis nula de que la pendiente es igual a 0.

6. Calculate a confidence interval for a slope of 90%.

```
Confint(RegModel.2, level=0.90)
```

	Estimate	5 %	95 %
(Intercept)	199.297502	52.2166142	346.378389

	Estimate	5 %	95 %
weight	1.622343	-0.4847468	3.729432

El IC90% de la pendiente abarca entre (-0.4847468; 3.729432). Cruza la línea del 0, por tanto, no es estadísticamente significativo

7. Show the estimated confidence intervals (95%) of the average fat for individuals between 30 and 90 kg.

```
RegModel.3 <- lm(fats~weight, data=DataReg, subset=weight>=30 & weight<=90)
summary(RegModel.3)
```

```
##
## Call:
## lm(formula = fats ~ weight, data = DataReg, subset = weight >=
##     30 & weight <= 90)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -130.85  -52.24   -4.53   55.58  130.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 250.8931   121.1004   2.072   0.0502 .
## weight        0.9098     1.7049   0.534   0.5990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.72 on 22 degrees of freedom
## Multiple R-squared:  0.01278,    Adjusted R-squared:  -0.0321
## F-statistic: 0.2848 on 1 and 22 DF,  p-value: 0.599
```

```
predict(RegModel.3,interval = ("confidence"))
```

	fit	lwr	upr
1	327.3165	269.1001	385.5328
2	317.3087	283.1644	351.4530
3	310.0302	271.9617	348.0988
4	314.5792	281.6465	347.5120
5	320.0381	281.6696	358.4065
6	313.6694	280.3905	346.9484
7	308.2106	266.1380	350.2833
8	316.3989	283.0261	349.7716
9	322.7675	278.0072	367.5277
10	319.1283	282.4536	355.8029
12	331.8655	258.3813	405.3496

	fit	lwr	upr
13	310.0302	271.9617	348.0988
14	302.7518	245.0207	360.4830
15	304.5714	252.4951	356.6477
16	313.6694	280.3905	346.9484
17	305.4812	256.0952	354.8672
18	322.7675	278.0072	367.5277
19	319.1283	282.4536	355.8029
20	325.4969	272.9624	378.0313
21	304.5714	252.4951	356.6477
22	311.8498	276.8021	346.8976
23	328.2263	267.0602	389.3924
24	300.9322	237.2624	364.6021
25	308.2106	266.1380	350.2833

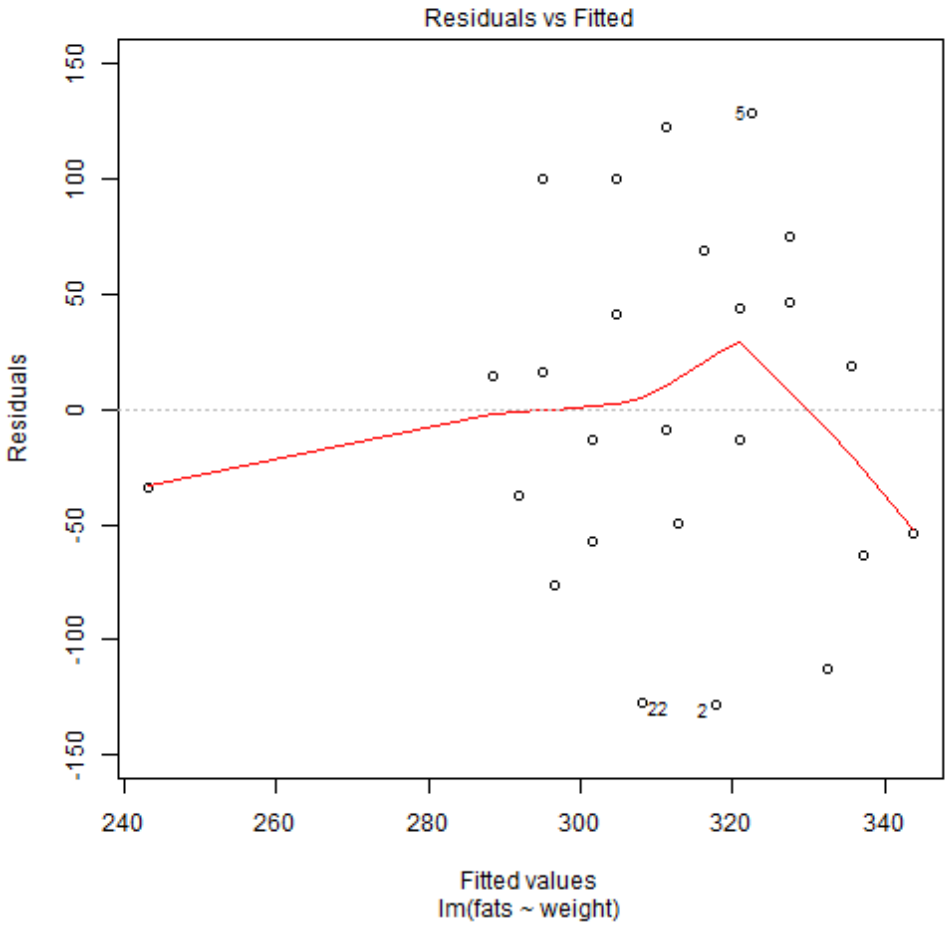
8. Perform a diagnostic model.

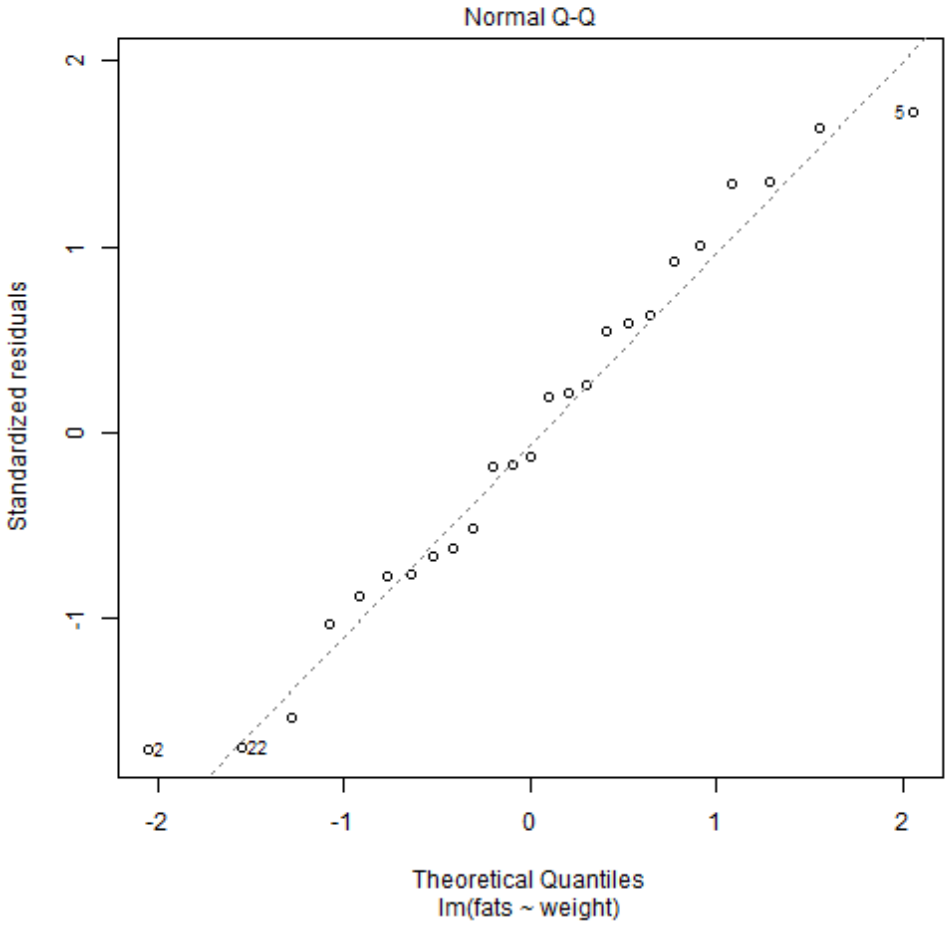
```
RegModel.2 <- lm(fats~weight, data=DataReg)
summary(RegModel.2)
```

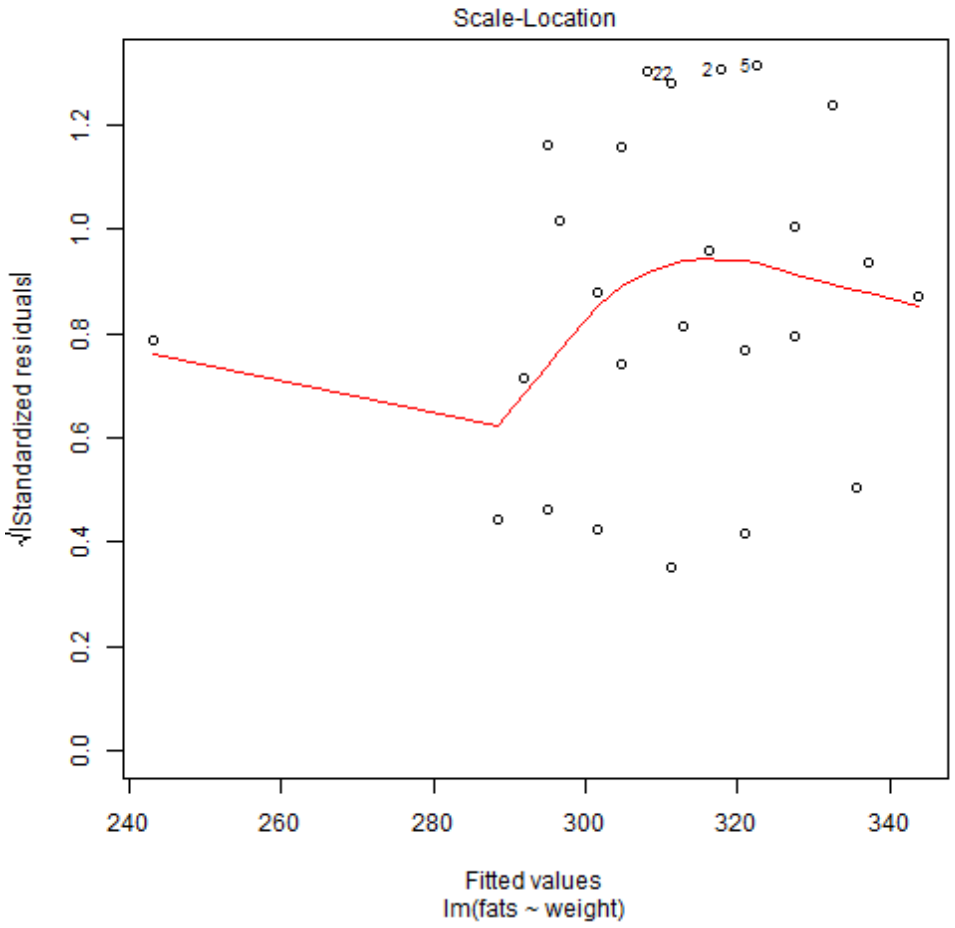
```
##
## Call:
## lm(formula = fats ~ weight, data = DataReg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.729  -53.686   -9.239   46.537  128.404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  199.298     85.818   2.322   0.0294 *
## weight         1.622      1.229   1.320   0.2000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.65 on 23 degrees of freedom
## Multiple R-squared:  0.07038,    Adjusted R-squared:  0.02996
## F-statistic: 1.741 on 1 and 23 DF,  p-value: 0.2
```

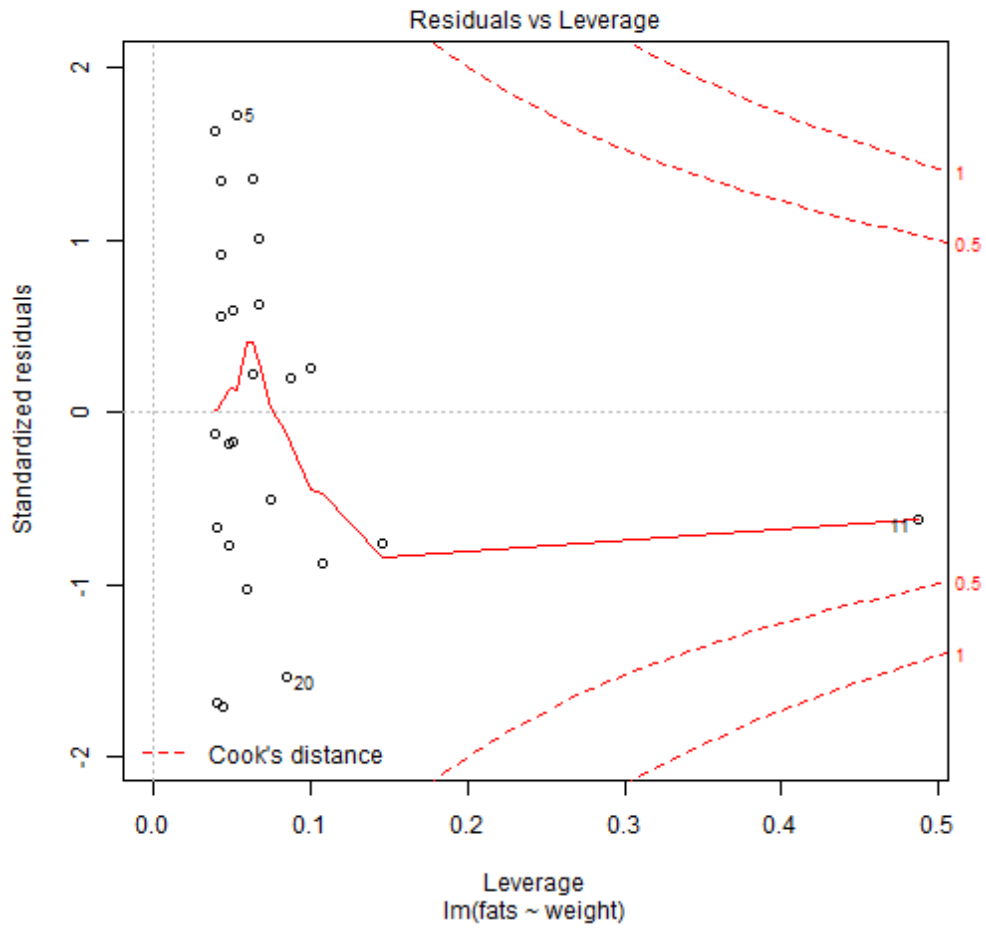
```
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
```

```
plot(RegModel.2)
```


Residuals vs Fitted
Residuals
150
100
50
0
-50
-100
-150
240
260
280
300
320
340
Fitted values
lm(fats ~ weight)
5
22
2


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(fats ~ weight)
5
2
22


Scale-Location
√|Standardized residuals|
Fitted values
lm(fats ~ weight)


Residuals vs Leverage
Standardized residuals
Leverage
lm(fats ~ weight)
Cook's distance

```
par(oldpar)
```

Exercise 3.

1. Use the parent's heights to predict children's heights (prediction).

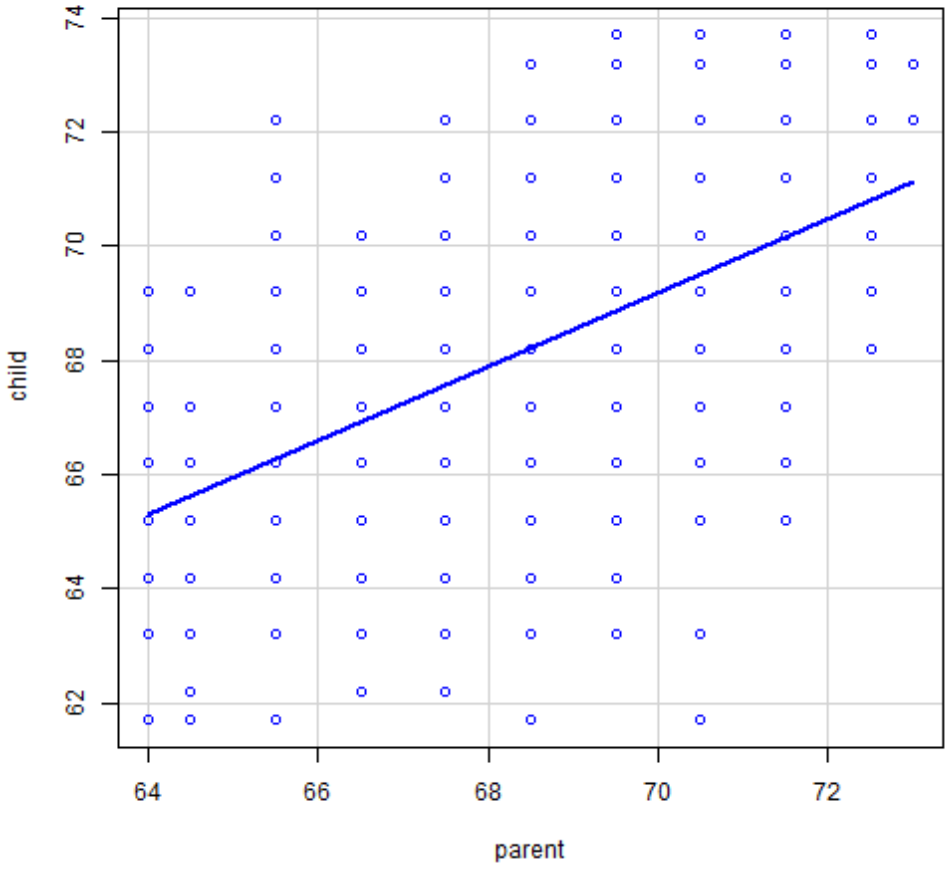
Me he bajado de internet el archivo "Galton.csv" (https://vincentarelbundock.github.io/Rdatasets/datasets.html)

```
library(UsingR)
```

```
Galton <- read.table("C:/Users/Ana/Documents/UOC/SOFTWARE ANALISIS DATOS/M(
```

Diagrama de dispersión para visualizar si ambas variables están relacionadas, parece que sí, parece que hay una correlación positiva:

```
scatterplot(child~parent, regLine=TRUE, smooth=FALSE, boxplots=FALSE, data=
```




```
cor(Galton[,c("child","parent")], use="complete")
```

	child	parent
child	1.0000000	0.4587624
parent	0.4587624	1.0000000

El coeficiente de correlación de Pearson entre ambas variables es de 0.459.

2. Find a relationship between parental and child heights (modelling).

```
Child_height <- lm(child~parent, data=Galton)
summary(Child_height)
```

```
##
## Call:
## lm(formula = child ~ parent, data = Galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

```
Confint(Child_height, level=0.95)
```

	Estimate	2.5 %	97.5 %
(Intercept)	23.9415302	18.4250996	29.4579608
parent	0.6462906	0.5655602	0.7270209

En este caso, la interpretación del intercept no tiene sentido, ya que los padres no pueden medir 0 unidades. Por cada unidad de altura del padre, el hijo medirá 0.646 unidades más (IC 95% 0.565; 0.727). Es decir, hay una asociación estadísticamente significativa entre la altura de los padres y la de los hijos con una magnitud de 0.646 unidades.

3. Investigate the variation in child heights that appears unrelated to parental heights (residual variation), and quantify what impact genotype information has beyond parental height in explaining child height (covariation). An important aspect, especially in questions 2 and 3, is assessing modelling assumptions.

Esa información la tenemos observando la R^2 del modelo, que nos indica el % de variabilidad de la variable dependiente (altura de los hijos) que es explicada por la variable independiente (altura de los padres). En nuestro caso, la R^2 es 0.210, es decir, sólo un 21% de la variabilidad de la altura de los hijos es explicada por la genética (altura de los padres). Por tanto, un 79% de la variabilidad de la altura de los hijos no es explicada por la altura de sus padres.

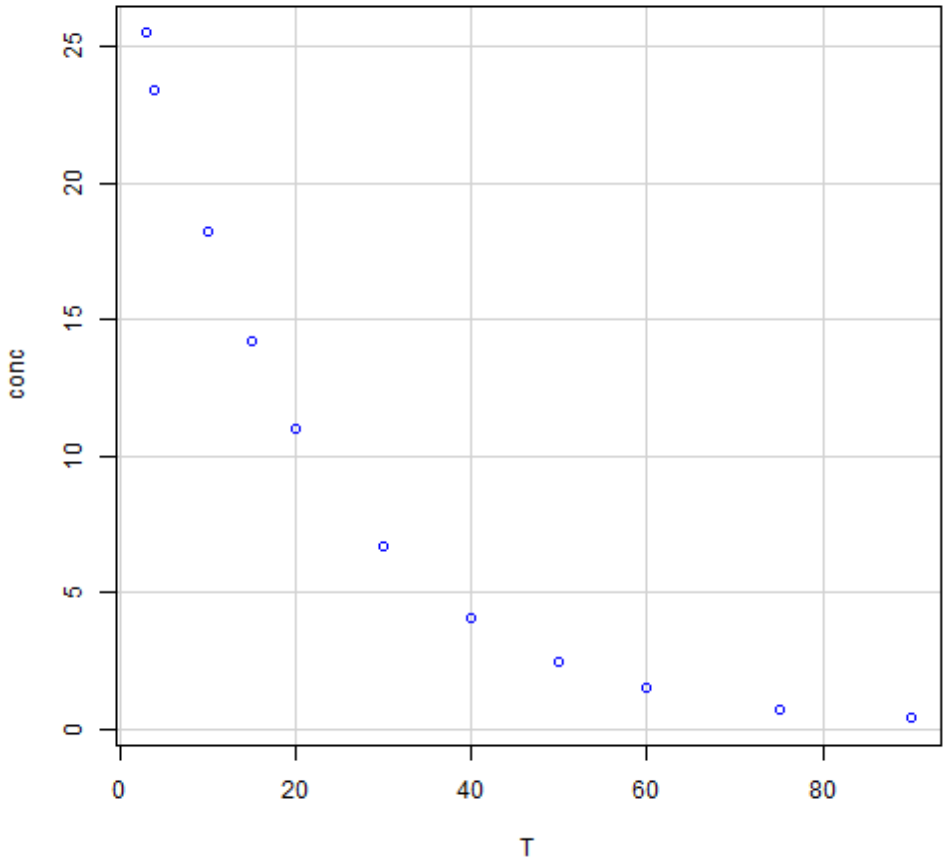
Exercise 4.

Importar los datos EsterData

```
EsterData <- read.table("C:/Users/Ana/Documents/UOC/SOFTWARE ANALISIS DATOS
View(EsterData)
```

Gráfica de dispersión:

```
scatterplot(conc~T, regLine=FALSE, smooth=FALSE, boxplots=FALSE, data=Ester
```

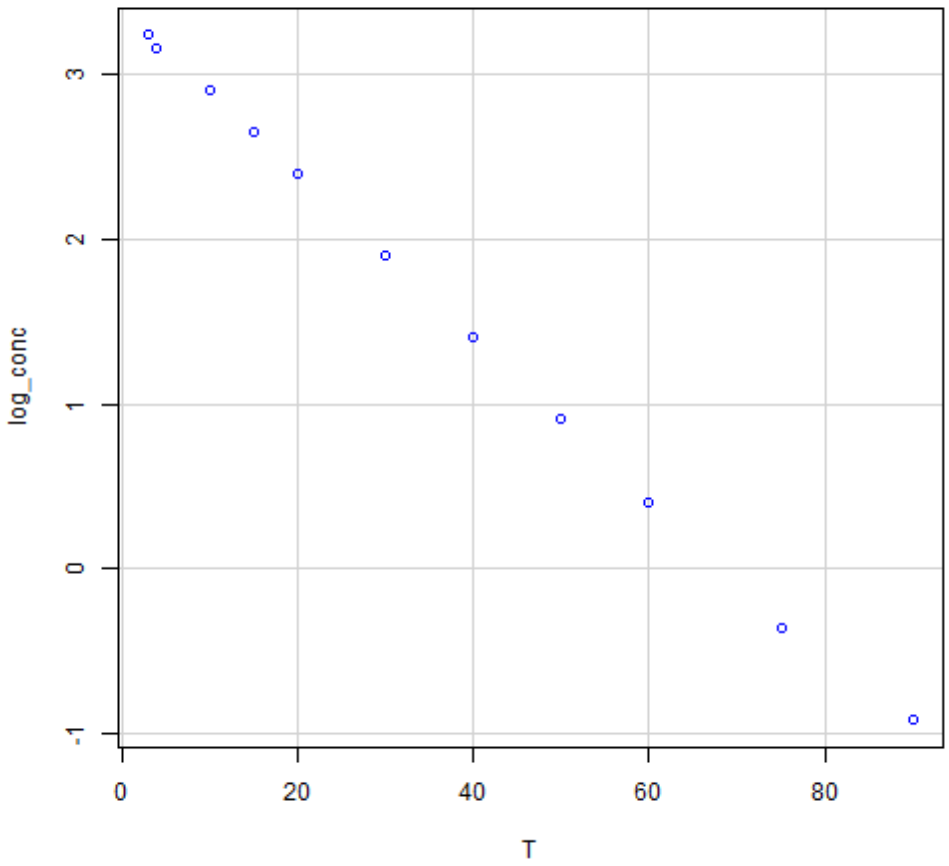


Añadir una nueva columna con el log(conc):

```
EsterData$log_conc <- with(EsterData, log(conc))
```

Gráfica de dispersión con log(conc):

```
scatterplot(log_conc~T, regLine=FALSE, smooth=FALSE, boxplots=FALSE, data=E
```



La asociación es ahora lineal, con una correlación inversa y estadísticamente significativa

Modelo lineal

```
RegModel.2 <- lm(log_conc~T, data=EsterData)
summary(RegModel.2)
```

```
##
## Call:
## lm(formula = log_conc ~ T, data = EsterData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.073561 -0.017338 -0.003809  0.018496  0.096500
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3652700  0.0220995   152.3  < 2e-16 ***
## T           -0.0486451  0.0004825  -100.8 4.72e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04513 on 9 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.999
## F-statistic: 1.016e+04 on 1 and 9 DF,  p-value: 4.715e-15
```