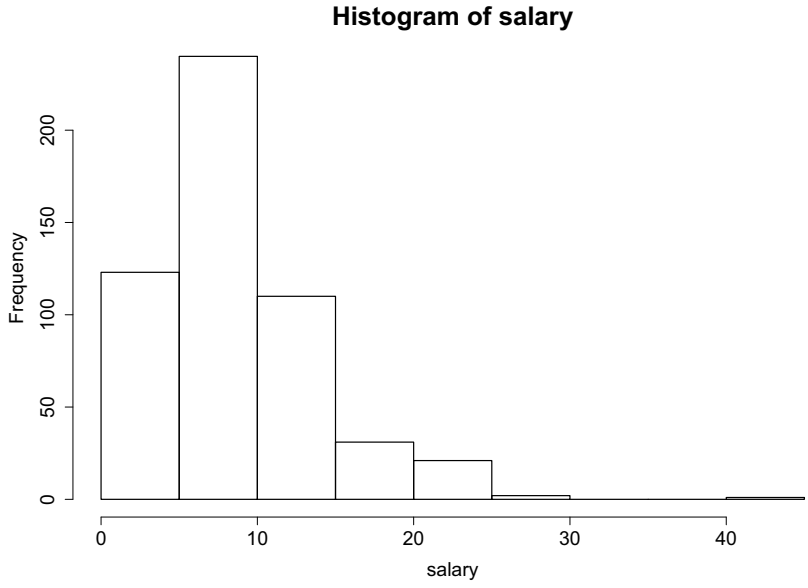
2.5 R and S-plus commands

2.5.1 Data visualisation

The data set contains: education, region, BL, HP, FE, married, experience, union, salary.

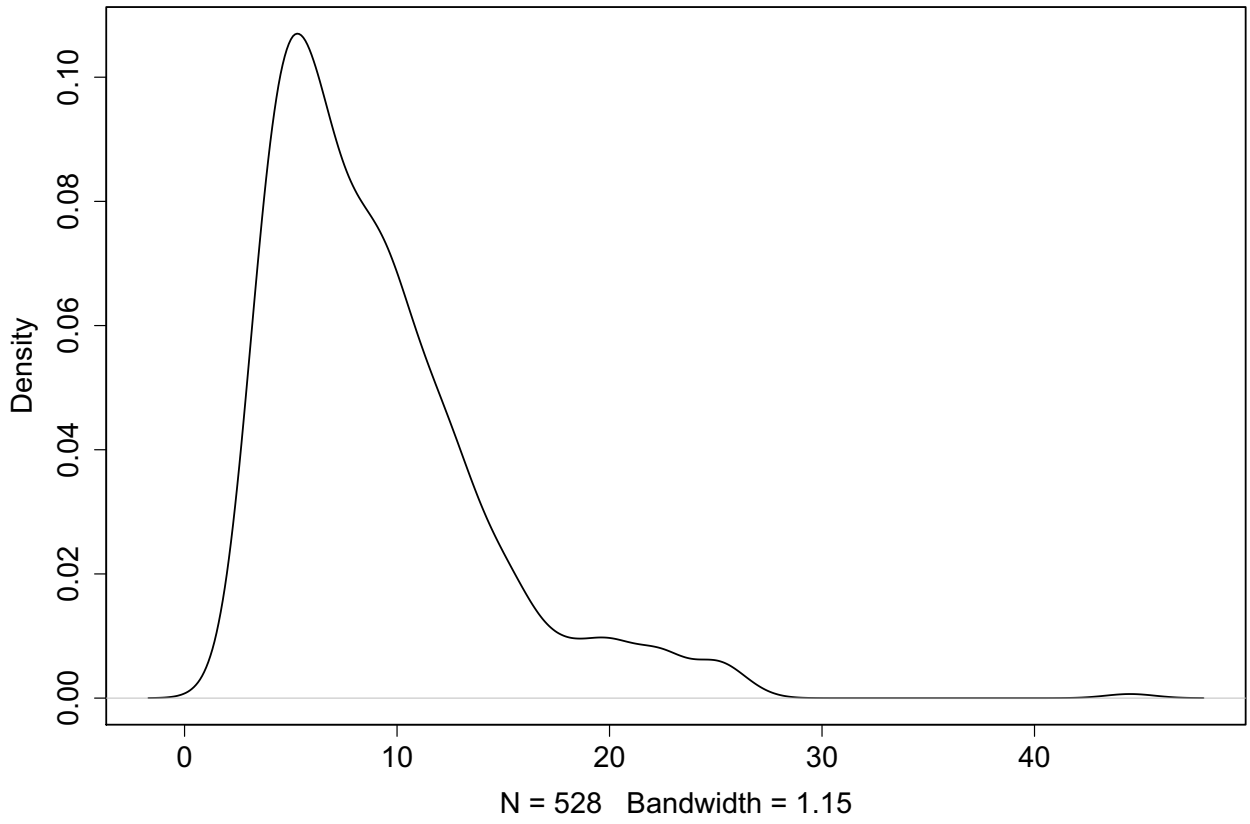
We can visualise the distribution using the histogram Figure 2.5.1, which is produced with the function **hist()** with argument salary.

```
# Read the file "salary_edcuation" which has a header
data<-read.table("./salary_education.dat", h=T)
attach(data)
#Plot the histogram of the variable salary
hist(salary)
```



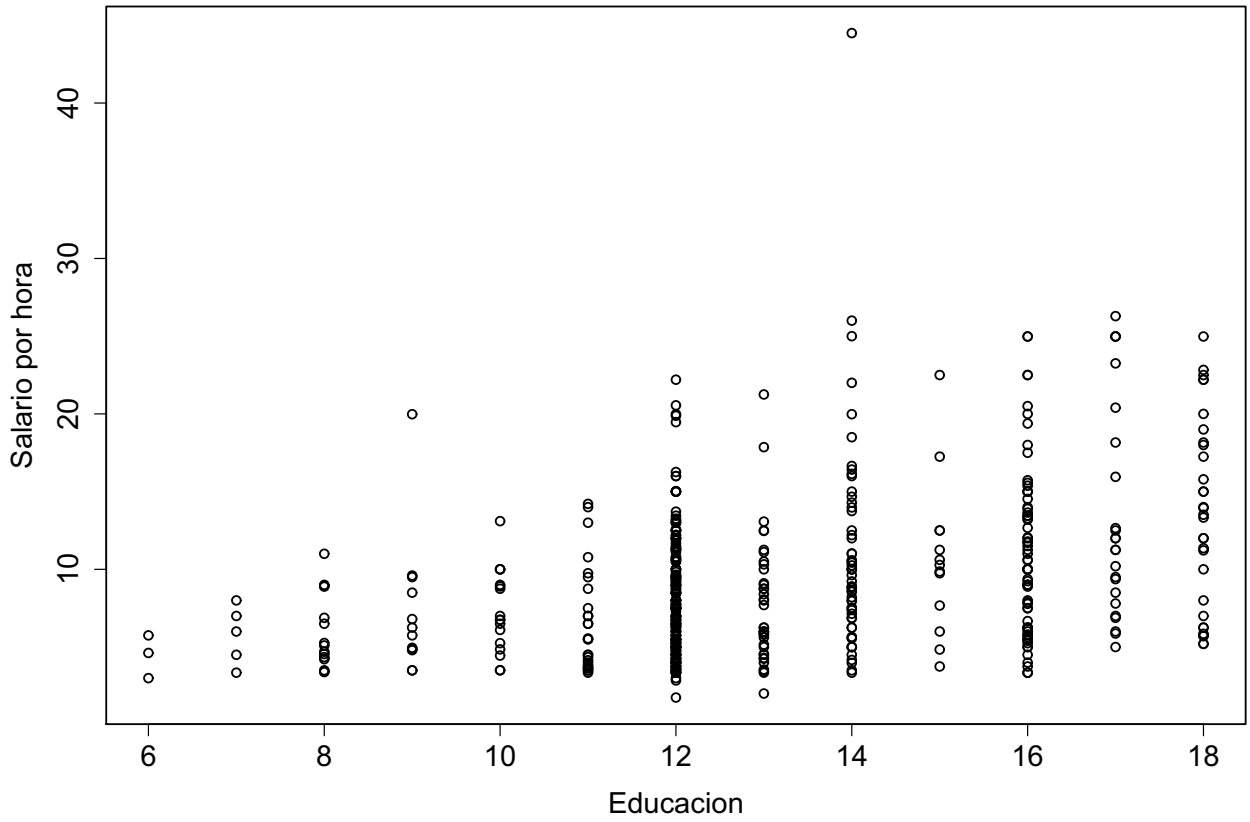
Another way to estimate the density function is to use kernel estimators as in Figure 2.5.1

```
#Plot density function of salary
plot(density(salary), main="")
```



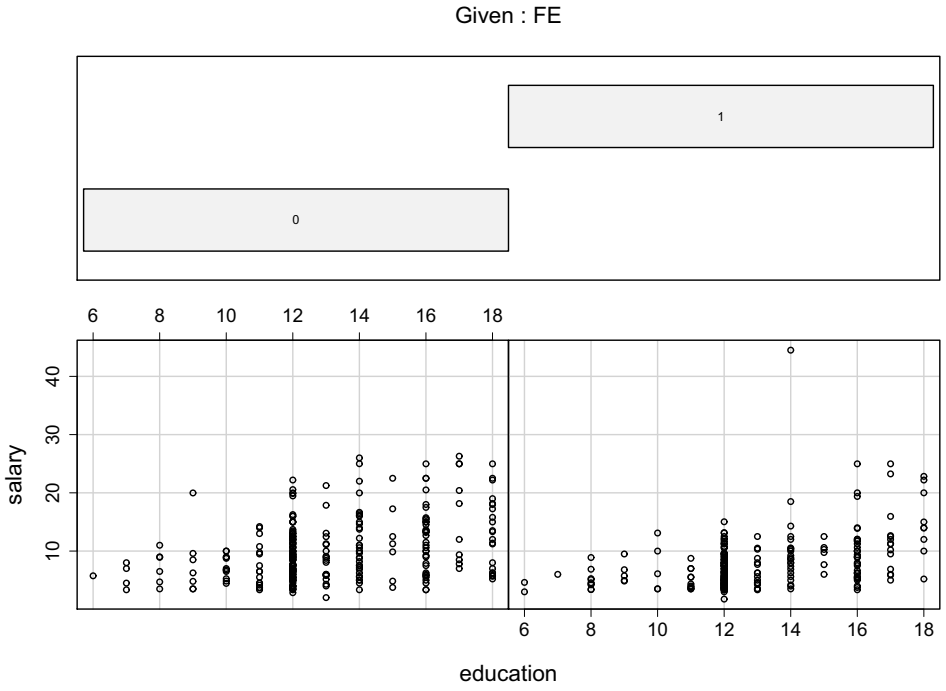
If we have two regressors then the data is bivariate and it is observed with a scatter plot. For example, we can compare the years of education and the salary at the same time as it can be seen in Figure 2.5.1:

```
# Scatter plot of salary versus education
```



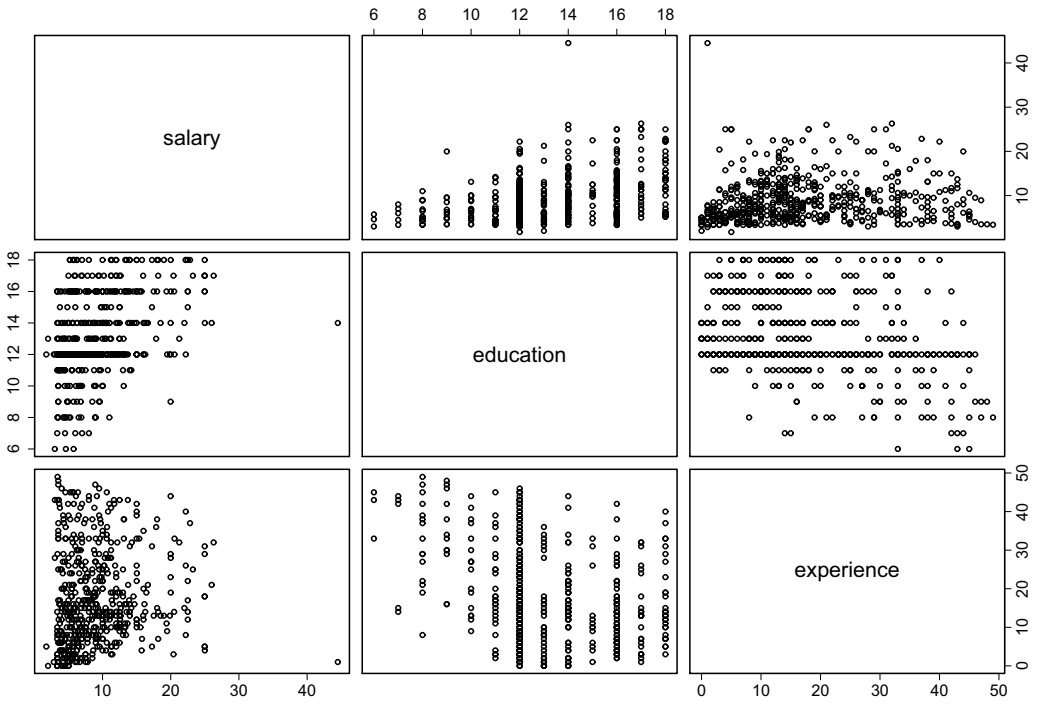
The scatter plot can be used with a third factor variable. For example, the relation between salary and education can be shown for males and females separately, Figure 2.5.1.

```
# Education, salary with sex dependency
FE<-as.factor(FE)
pdf("coplot.pdf")
par(cex.axis=1.5, cex.main=2,lwd=1.5, cex.lab=1.5, mar=c(5,5
coplot(salary~education|FE)
dev.off()
```



We can have a scatter plot for each pair of variables with the command **pairs()**. Figure 2.5.1 shows the relation between three numeric variables of our data set: salary, experience and years of education. We create a new data frame for this.

```
data2<-data.frame(salary, education, experience)
postscript("pairs.ps")
par(cex.axis=1.5, cex.main=2,lwd=1.5,
cex.lab=1.5, mar=c(5,5,3,2))
pairs(data2)
dev.off()
```



Other commands to create plots from the data are:

- plot(). The command plot(x, y, ...) is the general function to plot. If the variables x and y are numeric then they will be used in the axes. If x is a qualitative variable, then it will produce a set of boxplots. Example: plot(x, y , xlab="This is the x axis", ylab="This is the y axis", main="This figure has two axes of evil", type="l", xlim=c(0, 20), lty=3, pch=2)
- lines(), points(), abline(). These commands add dots or lines to existing plots.

2.5.2 Linear model in R and S-Plus

R and S-plus specify the regression model with the following formula:

$$y \sim 1 + x_1 + x_2 \tag{2.6}$$

where

- y is the dependant variable
- x_1, x_2 are the regressors. It can of course have more than two regressors.

The **1** in the formula indicates that the intercept (β_0) is necessary. If the **1** is omitted, then R considers that β_0 is necessary.

Formula (2.6) refers to the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

If the regression line should go through the origin (i.e. $y = 0$ when $x_1 = 0 = x_2$) and $\beta_0 = 0$, the formula en R is

$$y \sim -1 + x_1 + x_2$$

R and S-plus use the function linear model **lm()** to obtain the OLS estimator. The main arguments of lm() are:

lm(formula, data, na.action) where

- formula is the formula of the model (necessary)
- data is name of the data set (optional). If it is not included, the function lm use the variables in memory.
- na.action specifies how to manage the missing values (marked with NA). This argument is also optional. By default, it assumes that there are not missing values.

We are interested in estimate a model that quantifies the relationship between the salary, the experience and the years of education. This model is not going to go through the origin because although the years of education and experience can be zero, the salary can be positive.

```
>salary.lm<-lm(salary~1+experience+education)
> summary(salary.lm)

Call:
lm(formula = salary ~ 1 + experience + education)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5650 -2.8117 -0.5874  2.0026 36.2877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.56732    1.25737  -4.428 1.16e-05 ***
experience   0.10367    0.01734   5.978 4.17e-09 ***
education    0.97685    0.08471  11.532  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 4.584 on 525 degrees of freedom
Multiple R-Squared: 0.209,      Adjusted R-squared: 0.206
F-statistic: 69.36 on 2 and 525 DF,  p-value: < 2.2e-16
```

- The regression model is denoted by *salary.lm* to study the diagnostic results later on.
- The command *summary* can be used in many objects of the language R.
- The regression model (with two decimal digits) is

$$E[salary] = -5.57 + 0.10 experience + 0.98 education$$

- The p-valor 4.17e-09 refers to the t–test $H_0 : \beta_1 = 0$. Therefore, there is a great evidence of linear dependency between salary and experience. The same for the variables salary and education.
- The determination coefficient is $R^2 = 0.209$, i.e. the model explain only a 21% of the variation of salary.
- The F–statistic: 69.36 test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ whose p–value is 2.2e-16.

2.6 Linear model in Stata

```
. regress salary experience education

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  2,   525) =   69.36
       Model |  2914.64782     2  1457.32391            Prob > F      =  0.0000
    Residual |  11030.6042   525  21.0106748           R-squared     =  0.2090
-------------+------------------------------           Adj R-squared =  0.2060
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.5837

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experience |    .103667   .0173408     5.98   0.000     .0696011     .137733
   education |   .9768533   .0847097    11.53   0.000     .8104416    1.143265
       _cons |  -5.567325   1.257373    -4.43   0.000    -8.037426   -3.097223
------------------------------------------------------------------------------
```

We can consider the model through the origin.

```
> salary.lm2<-lm(salary~-1+education+experience)
> summary(salary.lm2)

Call:
lm(formula = salary ~ -1 + education + experience)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4030 -3.3494 -0.9755  1.8099 35.8020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
education    0.61677    0.02413  25.565  < 2e-16 ***
experience  0.06319    0.01499   4.214 2.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 4.664 on 526 degrees of freedom
Multiple R-Squared: 0.7998,     Adjusted R-squared: 0.7991
F-statistic:  1051 on 2 and 526 DF,  p-value: < 2.2e-16
```

2.6.1 Diagnóstico del modelo

Residuals and fitted values are fundamental tools to diagnose whether a linear model is appropriate to our problem. These

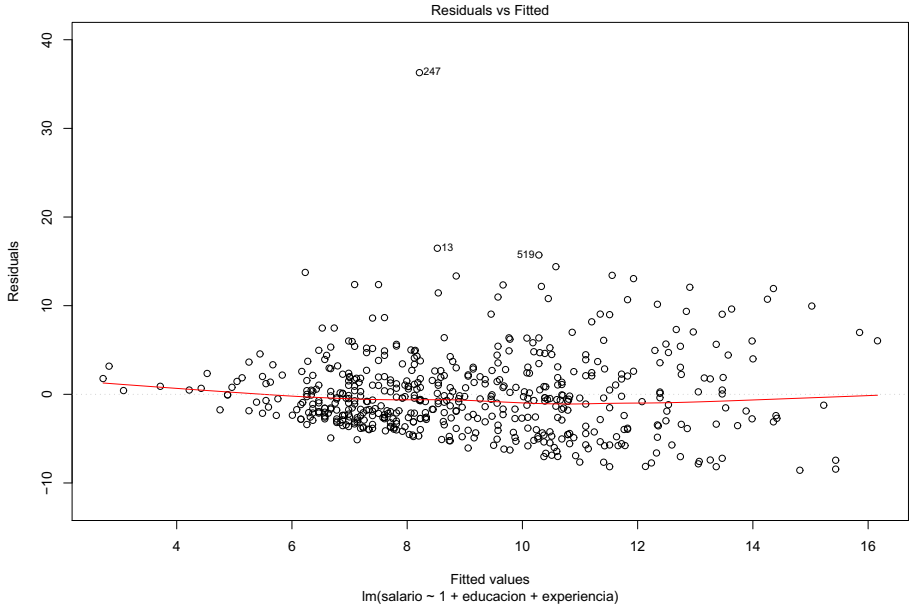
values can be extracted from our model with **resid(salary.lm)** and **fitted(salary.lm)**.

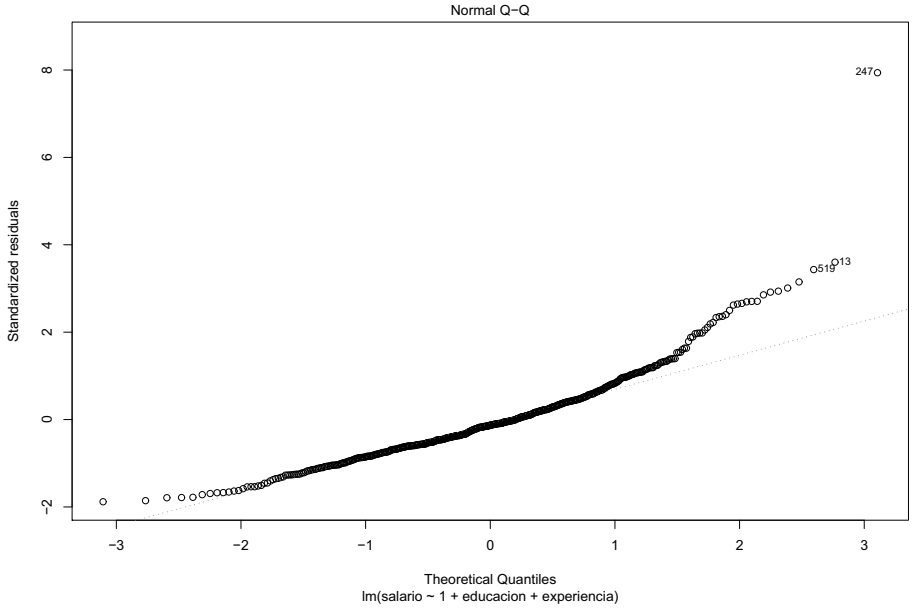
Graphically, the command **plot(salary.lm)** shows the following three figures:

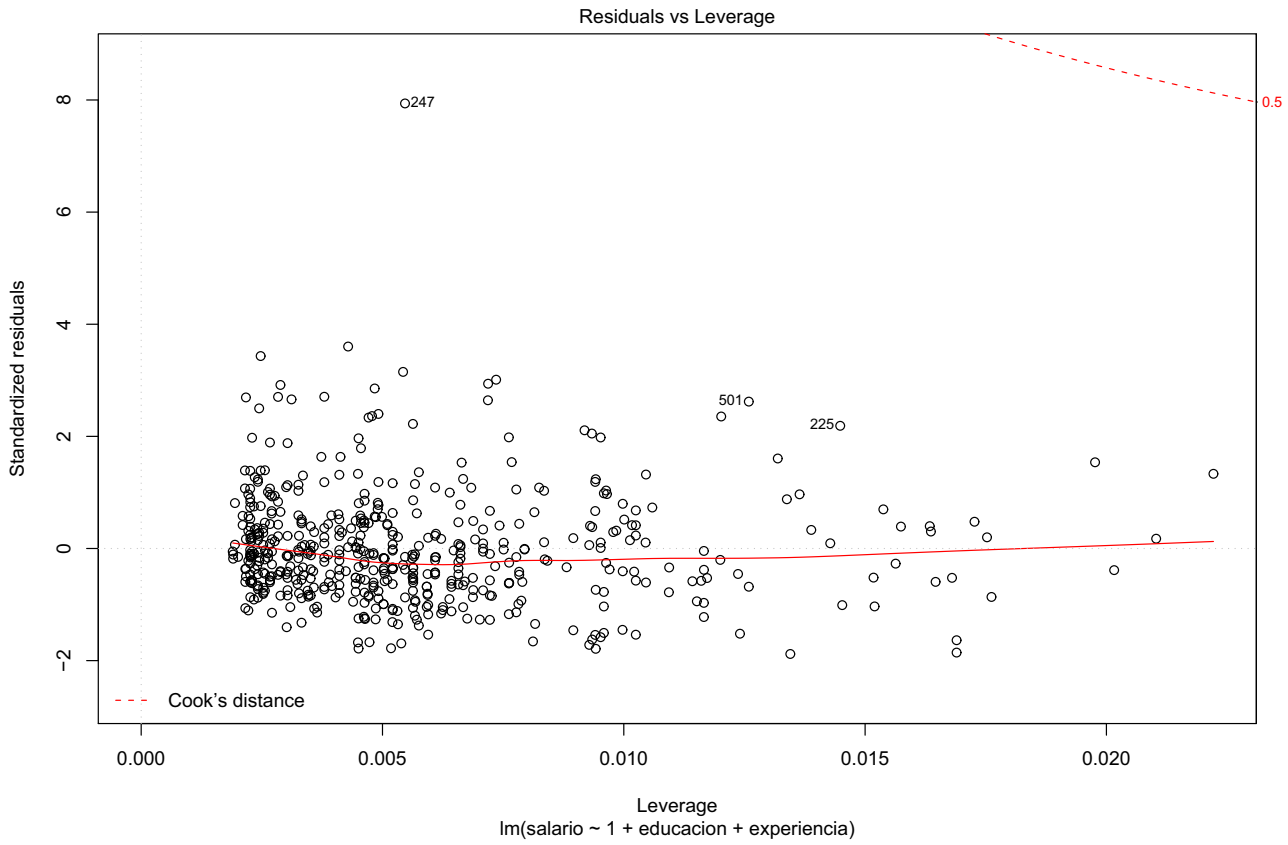
1. **Plot of the residuals vs fitted values**. It is useful to find a) extreme values (in the tails of the residuals); b) heterokedasticity (does the variance around zero changes?); c) no linearity with the endogenous variable (from the curvature of the plot).

2. **Normal QQ plot of residuals**. Are the ϵ_i normally distributed?.

3. **Cook distance plot** Some values might have a great impact on the parameter estimates. The Cook distance measures the influence of each point and measures how the vector of parameters changes if certain point is removed from the data set.


Residuals vs Fitted
Residuals
Fitted values
lm(salario ~ 1 + educacion + experiencia)
247
13
519


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(salario ~ 1 + educacion + experiencia)
247
519
13


Residuals vs Leverage
247
501
225
0.5
Cook's distance
Standardized residuals
8
6
4
2
0
−2
0.000
0.005
0.010
0.015
0.020
Leverage
lm(salario ~ 1 + educacion + experiencia)