Exam_cheat_sheet

Haardt Vittorio

2023-04-30

Contents

Exploratory analyses

Univariate summaries

First lines

`head()`

```
head(school)
```

```
##   students calworks expenditure    income    score
## 1      195   0.5102    6384.911 22.690001 6.089709
## 2      240  15.4167    5099.381  9.824000 3.685556
## 3     1550  55.0323    5501.955 14.695077 5.346915
## 4      243  36.4754    7101.831  8.978000 3.457385
## 5     1335  33.1086    5235.988  9.080333 4.655332
## 6      137  12.3188    5580.147 13.623322 4.510078
```

Each row reports information for a different school district. All the variables are quantitative (numeric); more specifically, the first one (*students*) is discrete, while all the others are continuous. The variables exhibit very different variability among themselves. For instance, the percentage of students eligible for income assistance (variable *calworks*) varies widely across observations, ranging from 0.5% for the first unit to 55% for the third. On the contrary, *expenditure* shows similar values across the 6 displayed observations. In particular, we observe that the first observation, among these 6, has the highest value of the variable *score* and presents the highest value of the variable *income*, while a very low value (the lowest among these 6 observations) for the *calworks*.

Descriptive statistics

`skim_without_charts()`

```
library(skimr)
```

```
skim_without_charts(school)
```

Table 1: Data summary

Name	school
Number of rows	388
Number of columns	5
Column type frequency:	
numeric	5

Table 1: Data summary

Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
students	0	1	2611.06	3801.45	81.00	373.00	909.50	3142.25	25151.00
calworks	0	1	13.43	10.79	0.00	5.26	10.75	19.14	71.71
expenditure	0	1	5267.55	602.56	3926.07	4894.33	5186.03	5523.18	7711.51
income	0	1	14.08	6.26	0.46	10.03	13.33	17.51	36.17
score	0	1	5.89	1.15	3.32	5.04	5.78	6.60	10.38

Skipping directly to the part of descriptive analysis, we summarize hereafter the main results:

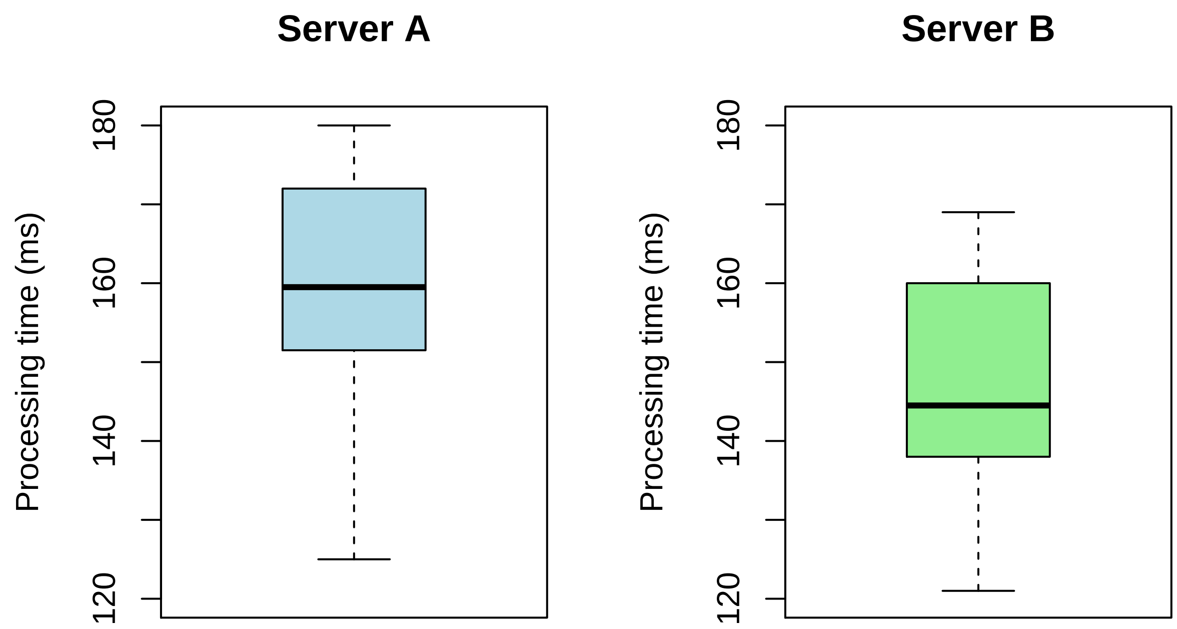
- The range of the *score* variable is quite narrow, being the minimum and maximum values equal to 3.32 and 10.38, respectively. From the median, we can observe that half of the considered school districts obtained a score lower than or equal to 5.78. The mean assumes a very similar value (approximately 5.89), which can lead to a preliminary assumption of a reasonably symmetrical distribution. From the analysis of quartiles, we obtain that one-quarter of the examined districts obtained a score lower than 3.32, and another quarter obtained a score higher than 6.60. The observation of quartiles also confirms the hypothesis of a good symmetry of the distribution. The standard deviation is reported as a measure of variability: the score variable has a rather low variability, with a value of 1.15; we can interpret this result by stating that the score obtained by school districts has an average variability around the mean of approximately ±1.15.
- The *students* variable shows the highest values of standard deviation (approximately equal to 3800) and range (25070). The minimum and maximum values are indeed far apart: the smallest school is attended by only 81 students, while the largest by more than 25000. On average, a school district accommodates 2600 students, but half of the schools are attended by less than 910 students.
- The variable *calworks* measures the percentage of students eligible to receive economic assistance. The minimum value is equal to 0, indicating that in at least one school district no student is eligible. Conversely, the school with the highest percentage has as many as 72% of students receiving assistance. The distribution is significantly skewed, with a long tail to the right, indicating that there are few schools with a very high percentage and most districts with lower values. On average, the percentage of eligible students is around 13%, and three-quarters of the schools do not exceed 20%.

Variability indices

Boxplot

```
boxplot()
```

```
par(mfrow=c(1,2))
# Box plot Server A
boxplot(data1$Server_A, col="lightblue",
        main="Server A",
        ylab="Processing time (ms)", ylim= c(120,180))
# Box plot Server B
boxplot(data1$Server_B, col="lightgreen", main="Server B",
        ylab="Processing time (ms)",
        ylim= c(120,180))
```



The box plot make clear how the Server A has an intervall slightly higher than Server B. The interval of server A is larger going form 125 to 180, while for server B the interval goes form 120 to 170. In sever A the meadian is around 160 while for sever B it is around 145. The box contain the central 50% of the data form le lower quartile to the upper quartile, and the lenght of the central box is the interquartile range $(Q_3 - Q_1)$ that in this case is similar for the two variable's distributions. In this case thare are not outliers, identified as observation falling 1.5 interquantile range.

Extra

Are they from an observational or a randomized experiment?

- In uno studio osservazionale, i soggetti sono assegnati ai gruppi di studio sulla base di fattori naturali o circostanze, senza un controllo diretto da parte dei ricercatori. Ad esempio, uno studio osservazionale sulla relazione tra l'esposizione al fumo di sigaretta e la comparsa di malattie respiratorie potrebbe includere soggetti che scelgono di fumare o meno. In uno studio osservazionale, i ricercatori possono solo osservare e registrare le associazioni tra le variabili di studio, ma non possono manipolare direttamente le condizioni di studio.
- In un esperimento randomizzato, invece, i soggetti sono assegnati in modo casuale a diversi gruppi di studio, in modo da ridurre al minimo gli effetti delle variabili di confondimento e garantire una distribuzione equilibrata delle caratteristiche dei partecipanti nei gruppi di studio. Ad esempio, in uno studio randomizzato sulla stessa relazione tra esposizione al fumo di sigaretta e malattie respiratorie, i soggetti verrebbero assegnati a caso a un gruppo sperimentale che riceve l'esposizione al fumo di sigaretta e un gruppo di controllo che non riceve l'esposizione al fumo di sigaretta.

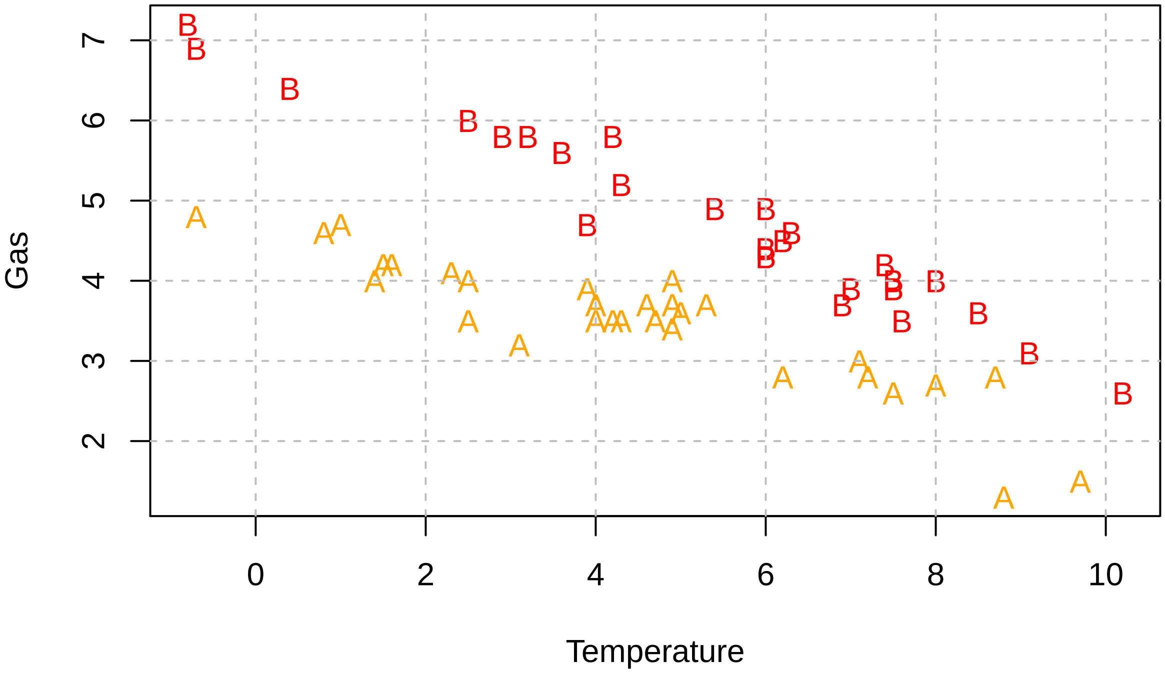
```r
load('/Users/Vitto_1/Desktop/Data (for the exercises)-20230430/dtclass.Rdata')
mu <- function(x){round(mean(x),3)}
math <- by(dtclass$math, dtclass$class, mu)
read <- by(dtclass$read, dtclass$class, mu)
data.frame(math_avg = c(math), read_avg= c(read))
```

```
##            math_avg read_avg
## classtwo      4.900   24.900
## classthree    4.483   21.414
## classone      3.739   16.174
## classzero     4.438   21.017
```

```
by(dtclass$read, dtclass$class, mean)
```

```
## dtclass$class: classtwo
## [1] 24.9
## ------------------------------------------------------------
## dtclass$class: classthree
## [1] 21.41379
## ------------------------------------------------------------
## dtclass$class: classone
## [1] 16.17391
## ------------------------------------------------------------
## dtclass$class: classzero
## [1] 21.01685
```

```
library(MASS)
data(whiteside)

col <- c("red", "orange")
plot(whiteside$Temp, whiteside$Gas,
     pch = as.character (whiteside$Insul),
     col = col[whiteside$Insul],
     ylab= "Gas", xlab = "Temperature")
grid(lwd = 1, lty = 2, col = "gray")
```



Generating values from random variables

Realisations from the univariate Gaussian distribution

Realizations from the Gaussian distribution

`rnorm()`

$X \sim N(\mu = 7, \sigma^2 = 2)$

```
set.seed(130)
x <- rnorm(1000, mean = 7, sd = sqrt(2))
```

We describe the generated values using univariate summaries.

`summary()`

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.399   6.015   6.961   7.011   7.934  12.065
```

First of all, we observe that the values of the mean and median are very similar to each other and both close to 7, which is the theoretical value assigned to the mean. The data genera- tion process seems to be quite accurate. Furthermore, the fact that the mean and median are essentially coincident suggests the symmetry of the distribution, consistent with known theoretical results. The minimum value of the generated data is 2.3, and the maximum is 12; these values are also approximately equidistant from the center (median) of the distribution.

We also calculate some dispersion indices, such as the variance and interquartile range.

`var()`

```
var(x)
```

```
## [1] 2.137289
```

The variance takes a value of 2, which is also very close to the theoretical value of the assumed random variable.

`IQR()`

```
IQR(x)
```

```
## [1] 1.918718
```

The interquartile range (calculated as the third quartile minus the first quartile) is 1.9. The simulated data are therefore quite concentrated around the mean.

Finally, we can also compute the values of $\mu \pm 3\sigma$; indeed it is known that, for the Gaussian distribution, more than 99% of the realizations are in the interval $[\mu - 3\sigma, \mu + 3\sigma]$. We can therefore evaluate if the number of generated data outside this interval is very limited.

```
x_Max <- mean(x)+3*sd(x); x_Max
```

```
## [1] 11.39685
```

```
x_Min <- mean(x)-3*sd(x); x_Min
```

```
## [1] 2.625163
```

```
length(x[x < x_Min])
```

```
## [1] 2
```

```
length(x[x > x_Max])
```

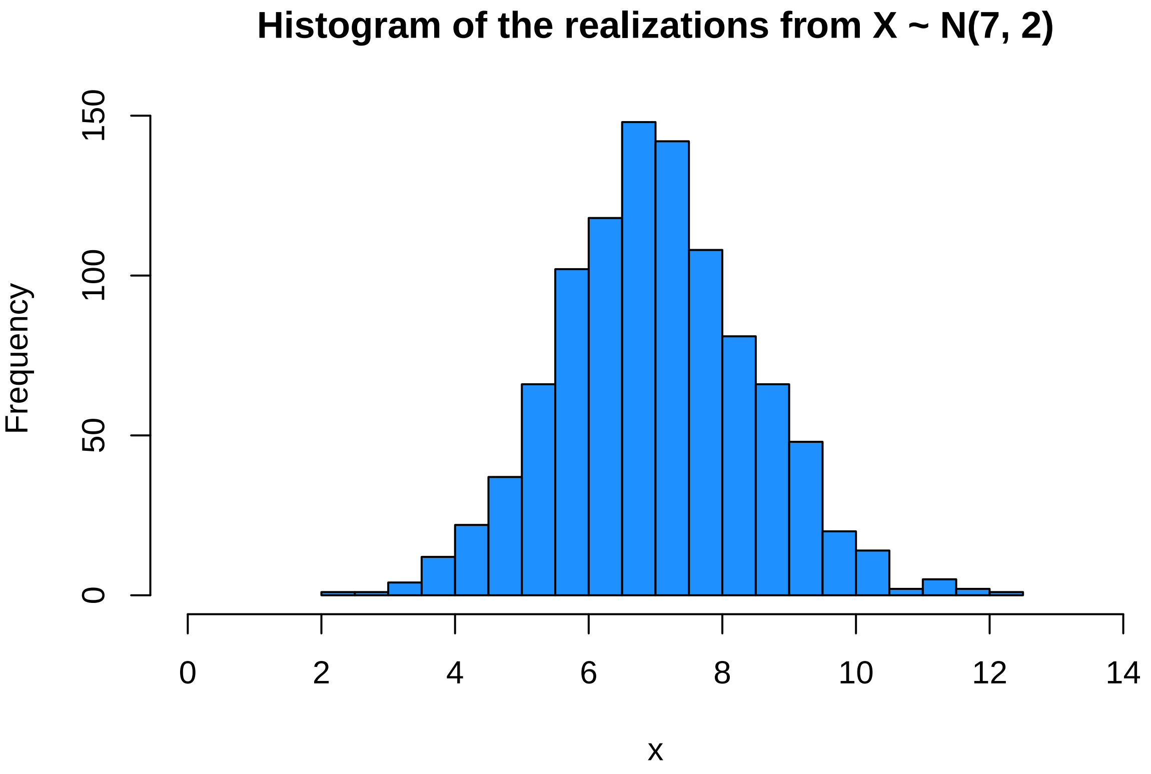
```
## [1] 3
```

We detect the presence of only 5 observations that fall outside the calculated interval, a very low number, which is consistent with the information known from a theoretical point of view; in particular, 2 of them are in the left tail and 3 in the right tail.

Finally, we draw a histogram.

```
hist()
```

```
hist(x, main = "Histogram of the realizations from X ~ N(7, 2)",
     breaks = 16,
     col = "dodgerblue",
     #prob = TRUE,
     xlim = c(0,14))
```



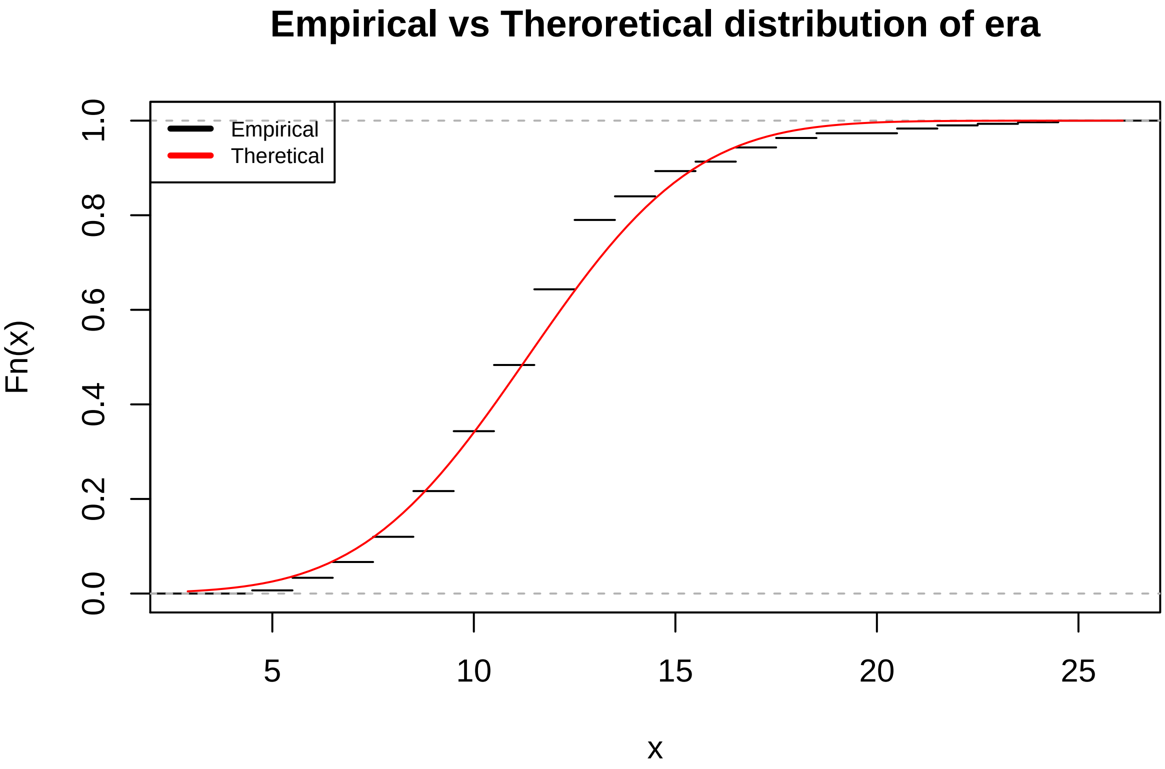
The histogram shows a bell-shaped curve, which is expected for a random variable with a univariate Gaussian distribution. The center of the distribution appears to be around 7, which is the mean of the random variable we used as reference. The spread of the distribution appears to be roughly consistent with the standard deviation we specified in the distribution. Overall, the histogram suggests that the realized values are consistent with those of a random variable having a Gaussian distribution we generated from.

Emprirical cumulative distribution function

```
plot(ecdf())
```

```
plot(ecdf(shell$eta), do.points = FALSE, main = "Empirical vs Theroretical distribution of era")
curve(pnorm(x, mean =mean(shell$eta), sd = sd(shell$eta)), add=TRUE, col="red")
```

```
legend("topleft",
       c("Empirical","Theretical"), lty=c(1,1),
       lwd=c(3,3),
       col=c("black", "red" ),
       cex= 0.7)
```



As we can see the empirical distribution of eta is composed by many segments, this because the observation in this variable are grouped under the same age and the variable not assume all the possible values in R, this cause the empirical distribution to be more sparse on the segments. Despite that it seams that the empirica distribution follow the theoretical one pretty well, only in 0.8 we observe a separation, but it is not vary influent. In this case, there are no significant deviations observed between the empirical and theoretical functions, which would confirm that the random variable in the reference population is normally distributed. However, the graphical analysis of the empirical distribution function is generally not very sensitive to subtle deviations from the reference distribution (normal in this case), and therefore a further examination of the data with other tools (QQ-plot, examined in the following, or hypothesis test) is appropriate.

Probability and quantiles of a Normal distribution

`pnorm()`

$P(Y \leq 130)$

```
pnorm(130, mean=140, sd=6)
```

```
## [1] 0.04779035
```

`qnorm()`

$P(Y \leq q) = 0.975$

```r
qnorm(0.975, mean=140, sd=6)
```

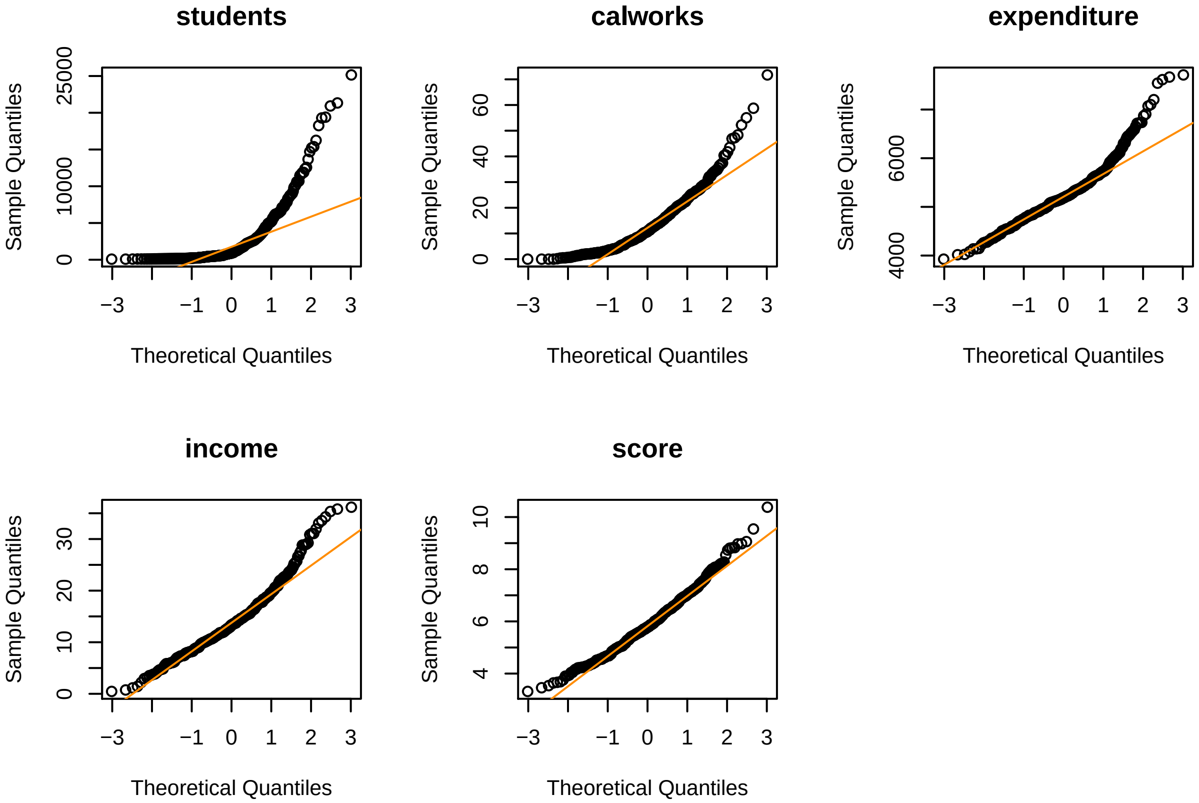
```
## [1] 151.7598
```

97.5% of the values of the distribution are les then or equal to 152.

Q-Q plot

```r
qqnorm()
```

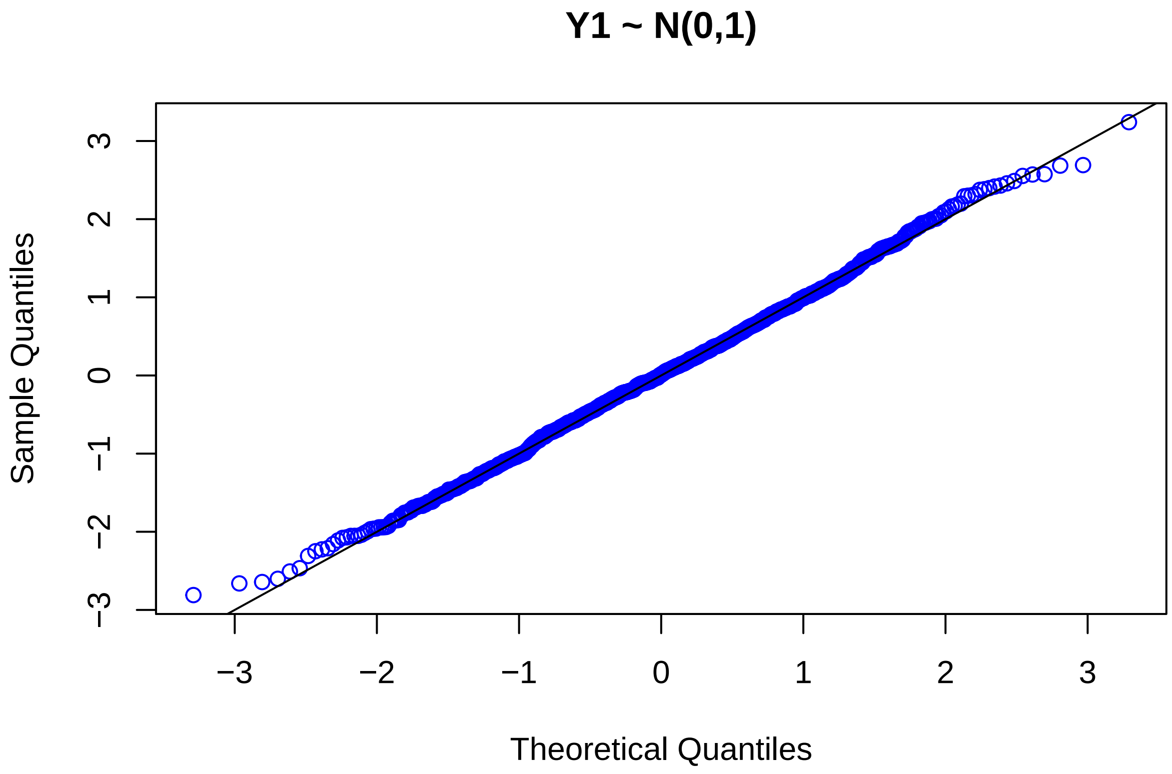
```r
par(mfrow = c(2, 3))
for (i in 1:ncol(school)) {
  qqnorm(school[, i],
         main = colnames(school)[i])
  qqline(school[, i], col = "darkorange")
}
```



We represent the quantile quantile (QQ) plot for each variable. If the data perfectly follows the reference distribution (Gaussian), the points on the QQ-plot will perfectly follow a straight line. Suppose the data deviates from the reference distribution; in this case the points on the QQ-plot will depart from the straight line more or less noticeably, depending on the severity of the deviation. Regarding the *students* and *calworks* variables, their respective QQ-plots confirm the non-normality of the data: the distribution of points on the plot deviates significantly from the reference line (the bisector of the first and third quadrant); in particular, they form a very pronounced curve, completely above the reference line. This indicates the presence of a particularly heavy right tail and, conversely, a very light left tail. This also implies the asymmetry of the two distributions. The other variables exhibit a similar behavior, albeit much less pronounced. In particular, the *expenditure* and *income* variables have a heavier right tail than normal, while the left tail follows the normal assumption. Finally, the *score* variable deviates minimally from the reference line, so the normality assumption for the reference random variable in the population can be considered satisfied.

Single one:

```
set.seed(123)
Y1 <-  rnorm(1000)
qqnorm(Y1, col = "blue", main = "Y1 ~ N(0,1)")
abline(0,1)
```



Realisations from the Student-t distribution

Realizations from the Student distribution

rt()

```
set.seed(4326)
t1 <- rt(1000, df = 1)
set.seed(4326)
t2 <- rt(1000, df = 10)
```

We briefly inspect the generated data.

summary()

```
summary(t1)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -821.621   -0.999    0.043    4.050    0.957 3422.254
```

```
summary(t2)
```

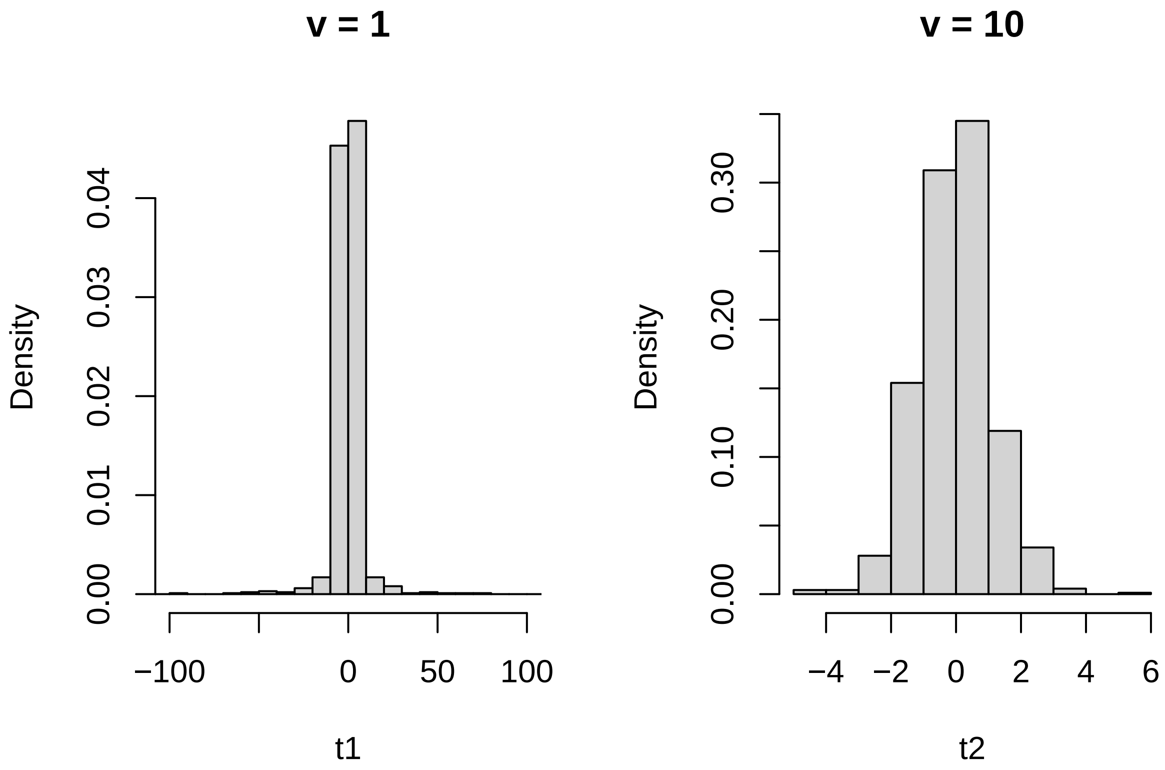
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -4.470076 -0.711690  0.002668 -0.023289  0.658093  5.061821
```

The first set of data (generated with $\nu = 1$) presents a very wide range of variation, with maximum and minimum values of -820 and 3420, respectively. It is worth noting that the values of the first and third quartiles are considerably smaller, indicating that most of the data are concentrated around the median, while there is a small minority of highly dispersed data far from the median value (it should be noted that it is not possible to talk about outliers, as they are points generated according to the given random variable). Considering instead the second set of data ($\nu = 10$), the maximum and minimum values are much smaller, resulting in a less wide range of variation. In both cases, the median is close to zero, as expected from known theoretical results. As for the mean, it should be noted that it is only defined for $\nu > 1$ (and in this case equal to 0), so the value of 4.05 obtained for the first set of data should not be considered.

We now represent, in the same graphical window, the histograms of the generated data. Regarding the first set of data ($\nu = 1$), due to the presence of extremely dispersed values compared to the median, we choose to display only a central interval of the range in order to obtain a better visualization of the data (note that there are only 3 values smaller than -100 and only 3 values bigger than 100).

hist()

```r
par(mfrow = c(1, 2))
hist(t1, main = "v = 1", breaks = 500, freq = FALSE, xlim = c(-100, 100))
hist(t2, main = "v = 10", freq = FALSE)
```

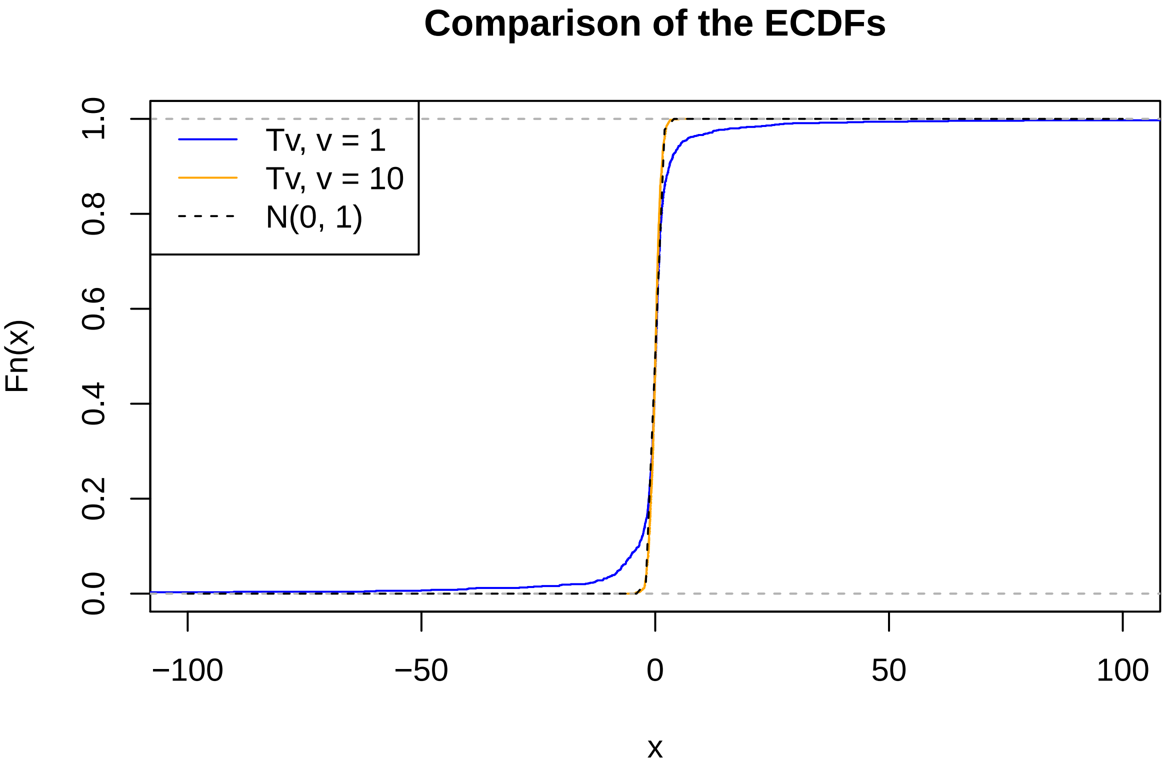


The comparison between the histograms of the data generated with the two Student-t distri- butions confirms the results obtained before. In particular, the first distribution with $\nu = 1$ has a much wider range of variation than the second one with $\nu = 10$, as can be seen from the horizontal axis of the two histograms. These features are consistent with the properties of the Student-t distribution, which becomes more similar to a Gaussian distribution as the number of degrees of freedom increases. Finally, it is worth noting that the frequency of occurrence of central values is much smaller in the distribution depicted on the left, as shown by values on the y-axis.

Emprirical cumulative distribution function

`plot(ecdf())`

```
plot(ecdf(t1), do.points = FALSE, col = "blue", xlim = c(-100, 100),
     main = "Comparison of the ECDFs")
lines(ecdf(t2), do.points = FALSE, col = "orange")
curve(pnorm(x), lty = 2, add = TRUE)
legend("topleft", c("Tv, v = 1", "Tv, v = 10", "N(0, 1)"),
       lty = c(1, 1, 2), col = c("blue", "orange", "black"))
```



Also in this case, we restrict the x-axis range, in order to show the results more clearly. We also add the curve of the cumulative distribution function of the standard Gaussian distribution. We can see that the spread of the Student-t distribution with 1 degree of freedom is much wider than the spread of the Student-t distribution with 10 degrees of freedom. This is because the Student-t distribution with 1 degree of freedom has heavier tails and thus more extreme values, while the T distribution with 10 degrees of freedom has lighter tails and is more concentrated around the mean. In addition, we can notice that the ECDF for the Student-t distribution with 10 degrees of freedom is closer to the standard Gaussian distribution (the dashed line) than the ECDF for the Student-t distribution with 1 degree of freedom. This is because as the degrees of freedom increase, the T distribution becomes more and more similar to the standard normal distribution.

Probability and quantiles of a Student distribution

`pt()`

$P(-1 < Y < 1)$

```
pt(1, df = 1) - pt(-1, df = 1)
```

```
## [1] 0.5
```

Note that this value is smaller that the corresponding one obtained using the standard Gaussian distribution.

Bivariate Normal distribution

Generate values from the bivariate Normal distribution

`rmvnorm()`

$(X, Y) \sim N(\mu_x = 0, \mu_y = 0, \sigma_x = 10, \sigma_y = 10, \sigma_{xy} = -25)$

```
require(mvtnorm)
```

```
## Loading required package: mvtnorm
```

```
set.seed(14263)
mu <- c(0, 0)
sigma1 <- matrix(c(100, -25, -25, 100), ncol = 2)
X <- rmvnorm(n = 3000, mean = mu, sigma = sigma1)
summary(X)
```

```
##        V1                  V2
##  Min.   :-37.2481   Min.   :-39.6197
##  1st Qu.: -6.9605   1st Qu.: -7.2802
##  Median : -0.2517   Median : -0.3906
##  Mean   : -0.1601   Mean   : -0.4059
##  3rd Qu.:  6.3459   3rd Qu.:  6.3571
##  Max.   : 34.2921   Max.   : 34.2013
```

The object returned in the output by the *rmvnorm()* function is a matrix with two columns and 3000 rows: the two columns represent the two components (X and Y) of the bivariate variable, while each row corresponds to a different observation. We briefly analyze the generated data by calculating its main descriptive statistics. Firstly, we recall that the components of a bivariate Gaussian distributed random variable have univariate Gaussian distributions; in this case, both components have the same distribution: $X, Y \sim N(0, 10^2)$. From the descriptive statistics we observe that, for both univariate components, the mean and median assume very similar values to each other and are approximately equal to zero (theoretical value assigned to the mean). The study of quartiles also highlights the symmetry of the data around the central values of the mean and median. In summary, the values generated for both components appear to be consistent with those of a random variable having the required Gaussian distribution.

`sd()`

```
sd(X[, 1])
```

```
## [1] 9.978415
```

```
sd(X[, 2])
```

```
## [1] 10.11361
```

`sd()`

```
cov(X[, 1], X[, 2])
```

```
## [1] -25.36925
```

The standard deviations of the two components are both very close to the assigned theoretical value of 10; also in this case, the value of the index calculated on the simulated data is consistent with the theoretical one. Finally, the value of the covariance (equal to -2.1) is also approximately equal to the theoretical one.

The two components are therefore (weakly) linearly and negatively associated with each other. Note that since the two components are not uncorrelated, they are definitely not independent either.

cov()

```
cov(X)
```

```
##           [,1]      [,2]
## [1,]  99.56876 -25.36925
## [2,] -25.36925 102.28508
```

Empircial variance-covariance matrix, veriy similar to theoretical values.

cor()

```
cor(X)
```

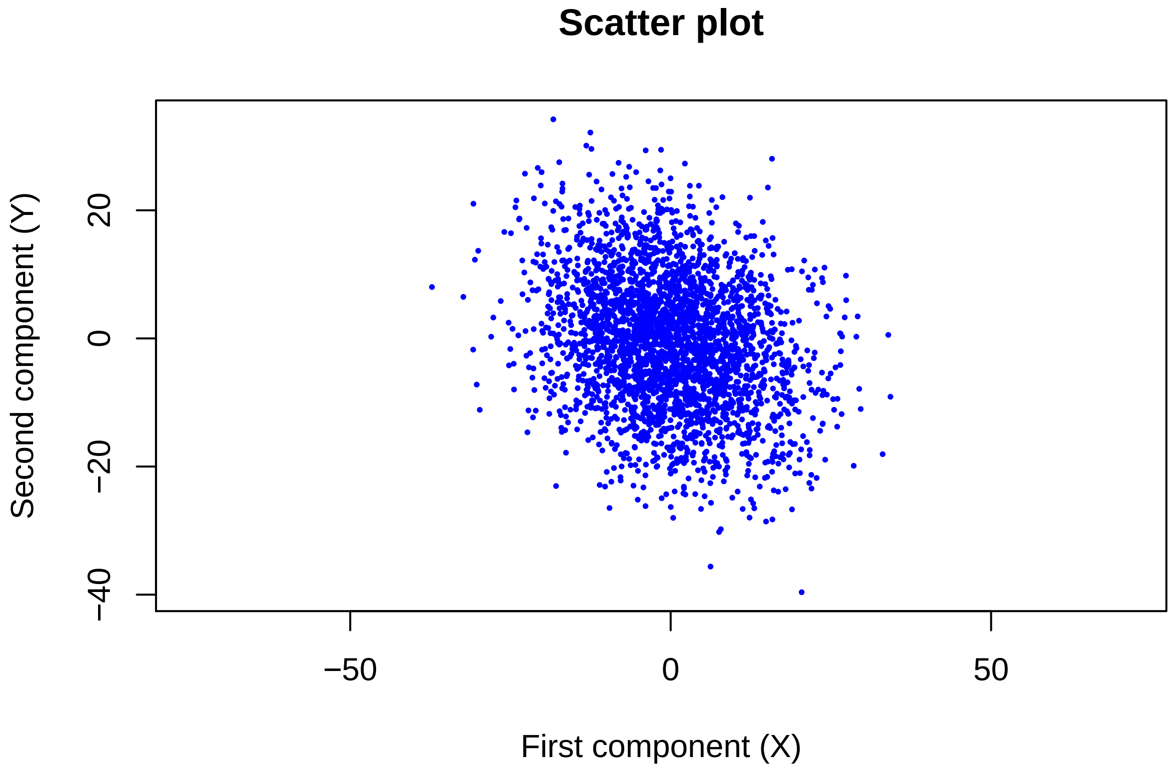
```
##            [,1]       [,2]
## [1,]  1.0000000 -0.2513853
## [2,] -0.2513853  1.0000000
```

Emprical correlation matrix is also very similar to the theoretical one. The theoretical correlation is optained form $\frac{\sigma_{xy}}{\sqrt{\sigma_x \cdot \sigma_y}}$ that in this case is -0.25.

Scatterplot

plot()

```
plot(X[, 1], X[, 2], col = "blue", main = "Scatter plot",
     xlab = "First component (X)", ylab = "Second component (Y)", pch=16, cex=0.4,asp = 1)
```



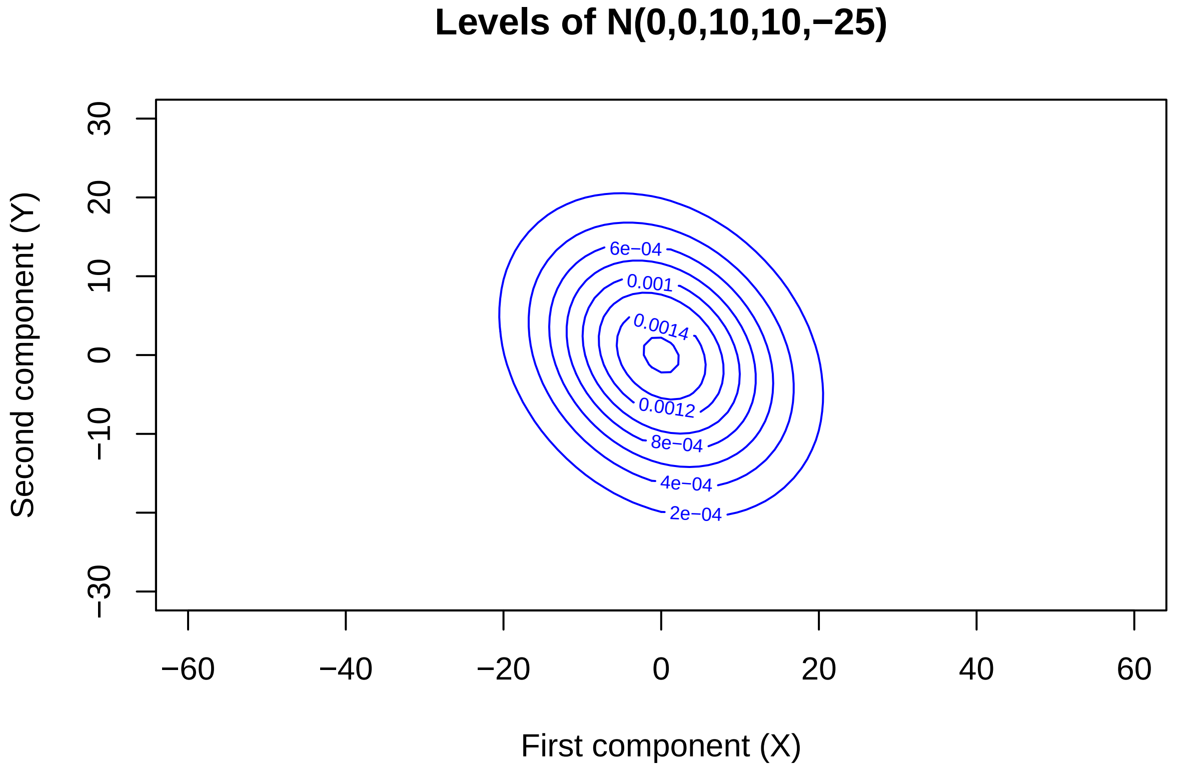
```
#asp = 1 to set the aspect ratio of the plot to 1:1.
#This means that the width and height of the plot will be equal,
#and the plot will not be distorted or stretched in any way
```

The scatter plot shows the realizations of the bivariate Gaussian distribution in the Cartesian plane; most of the points are clustered around the mean (0, 0). There is a limited number of points that deviate from the central part of the distribution. The range of variation is approximately the same for the two components, indicating that the corresponding variances are very similar. The points are arranged in an elliptical-shaped region showing a fairly well-defined orientation. This behavior indicates the presence of a negative covariance, which is however quite small with respect to the variance of the two components.

Contourplots

contour()

```
x1 <- x2 <- seq(-30, 30, length = 51)
dens <- matrix(dmvnorm(expand.grid(x1, x2), sigma = sigma1),
ncol = length(x1))
contour(x1, x2,
        dens,
        main = "Levels of N(0,0,10,10,-25)",
        col="blue",
        xlab = "First component (X)", ylab = "Second component (Y)",
        asp =1)
```



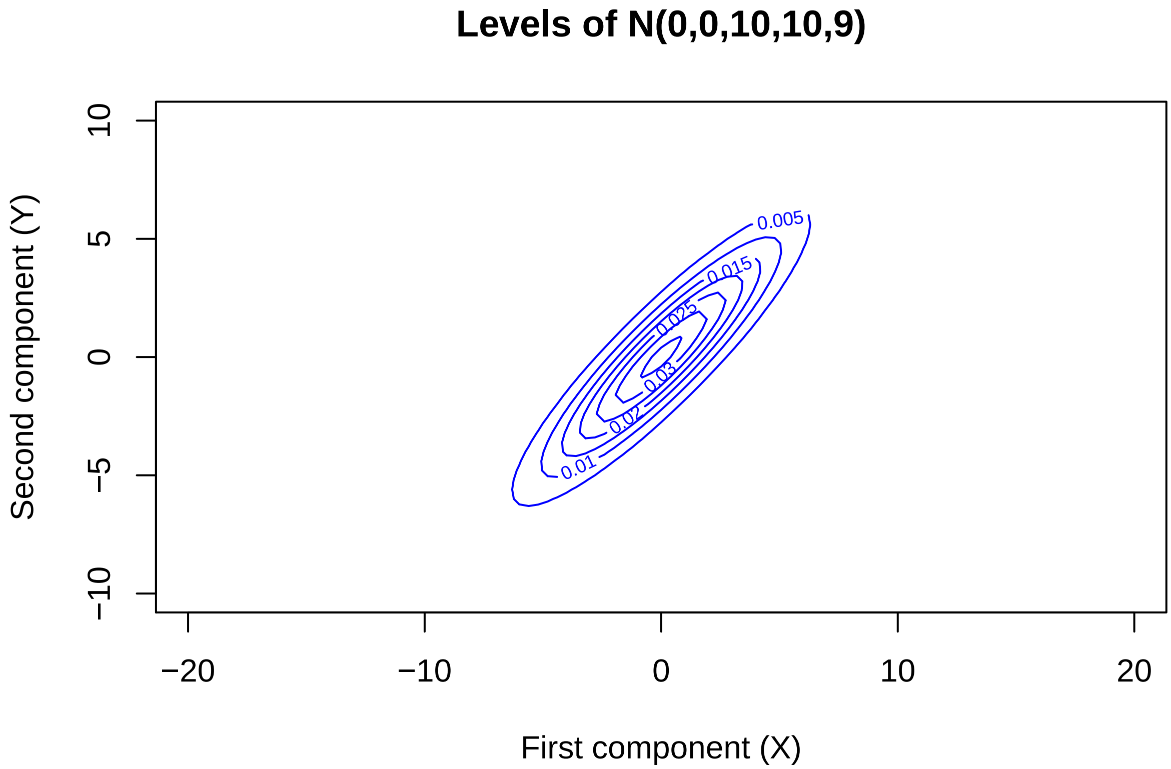
We observe in this plot the elliptical shape of the contour lines and the lack of orientation. This is due to the simultaneous occurrence of two behaviors:

- The approximately equal variability between the two components

- The negative (though small) value of the correlation between the two components $\rho_{xy} = -0.25$. The center of the circle (intended as the intersection point of the two axes) is located at the mean of the distribution.

Different output:

```
sigma3 <- matrix(c(10,9,9,10), ncol=2)
X2 <- rmvnorm(n = 3000, mean = c(0,0), sigma = sigma1)
x1 <- x2 <- seq(-10, 10, length = 51)
dens <- matrix(dmvnorm(expand.grid(x1, x2), sigma = sigma3),
ncol = length(x1))
contour(x1, x2,
        dens,
        main = "Levels of N(0,0,10,10,9)",
        col="blue",
        xlab = "First component (X)", ylab = "Second component (Y)",
        asp =1)
```



As we can see the shape of the ellipse higlights the high positive linear correlation between the two variabiles. The center of the circle (intended as the intersection point of the two axes) is located at the mean of the distribution. This is caused by an high value of correlation between the two variables $\rho_{xy} = 0.9$.

Measures of association

Correlations

`cor()`

```
round(cor(ratings),4)
```

```
##         rating profit capital f_flex
## rating  1.0000 0.5495  0.7558 0.7042
## profit  0.5495 1.0000  0.6180 0.6038
## capital 0.7558 0.6180  1.0000 0.6448
## f_flex  0.7042 0.6038  0.6448 1.0000
```

In this case they are all positive and quite high. We have to compare them to partial correlation that are more informative.

Partialcorrelations

`parcor(cov())`

```
library(ggm)
s <- cov(ratings)
round(parcor(s),4) #inpute variace covariance matrix
```

```
##         rating profit capital f_flex
## rating  1.0000 0.0144  0.5236 0.4081
## profit  0.0144 1.0000  0.3121 0.3060
## capital 0.5236 0.3121  1.0000 0.1235
## f_flex  0.4081 0.3060  0.1235 1.0000
```

There are difference, *rating* is linearly correlation to *profit*, but if we considered also other variables the correlation drop to almost zero. In general there are difference because the *profit* is the interesting variable, the correlation with *rating* drop considering also *capital* and *f_flex*. All correalation drops, this means that when considering the influence of other variables the linear correlation soffer.

Corplot

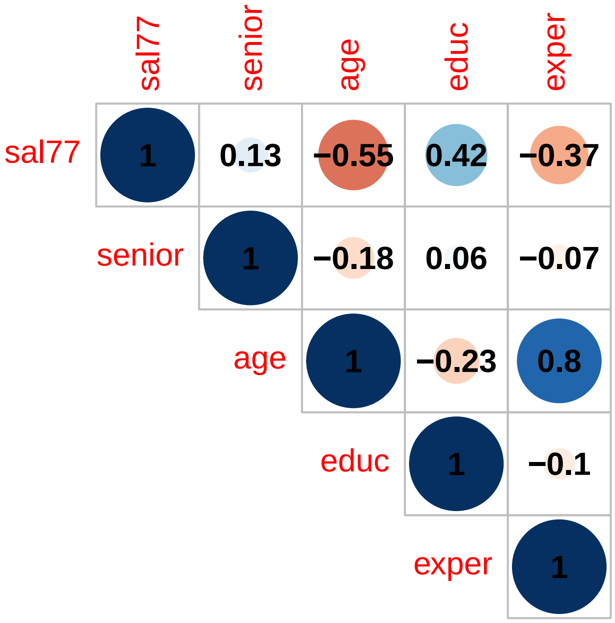
`corrplot(cor())`

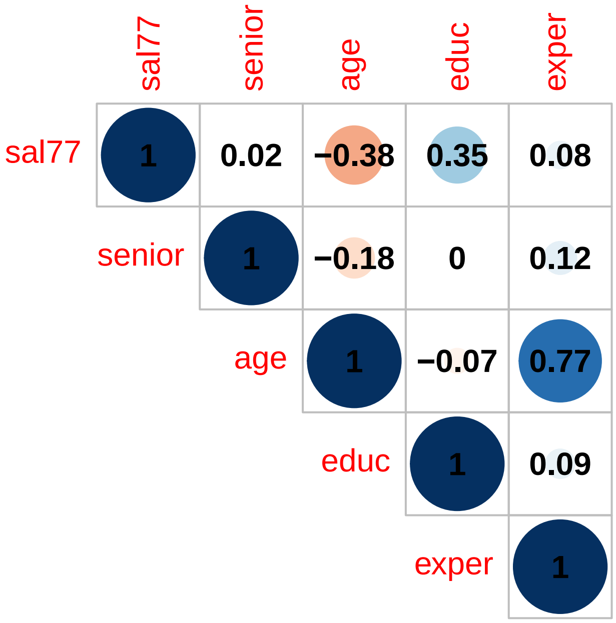
```
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.89 loaded
```

```
require(ggm)
par(mfrow = c(1, 2))
corrplot(cor(bank[, -1]),
         type = "upper",
         method = "circle",
         addCoef.col = TRUE,
         cl.pos = "n")
corrplot(parcor(cov(bank[, -1])),
         type = "upper",
         method = "circle",
         addCoef.col = TRUE,
         cl.pos = "n")
```

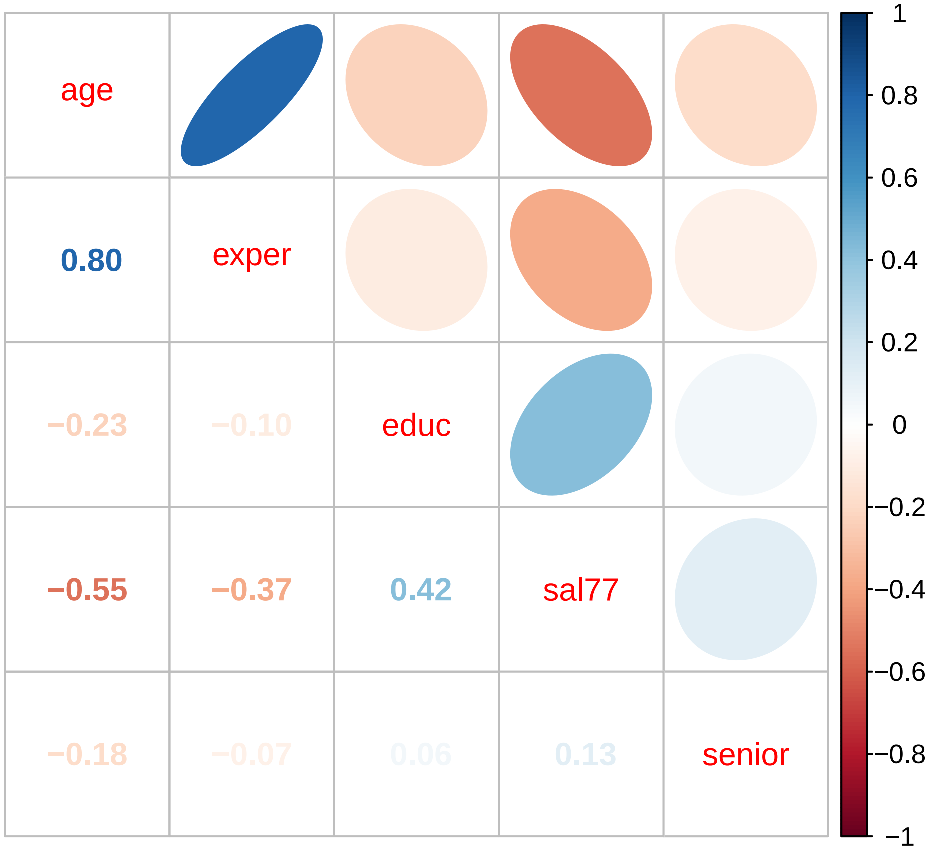




The most significant correlation coefficient is, as easily predictable, that between age and work experience: in fact, it is observed that as age increases, experience also increases (linearly). The amount of salary in 1977 is also moderately linearly associated with other covariates, particularly with age (negatively: older people generally receive lower salaries) and with years of education (longer periods of education correspond to higher salaries). The number of months of previous work experience also appears to be linearly correlated with initial salary (it would seem that as experience increases, salary decreases), but this value tends to be zeroed out considering the partial correlation coefficient: much of this association is therefore due to the interaction of other variables (especially the *age* variable). This value is also the most significant difference between the raw and partial correlation coefficients. The remaining pairs of variables do not appear to be significantly linearly associated, either in terms of raw or partial correlations.

Alternative form:

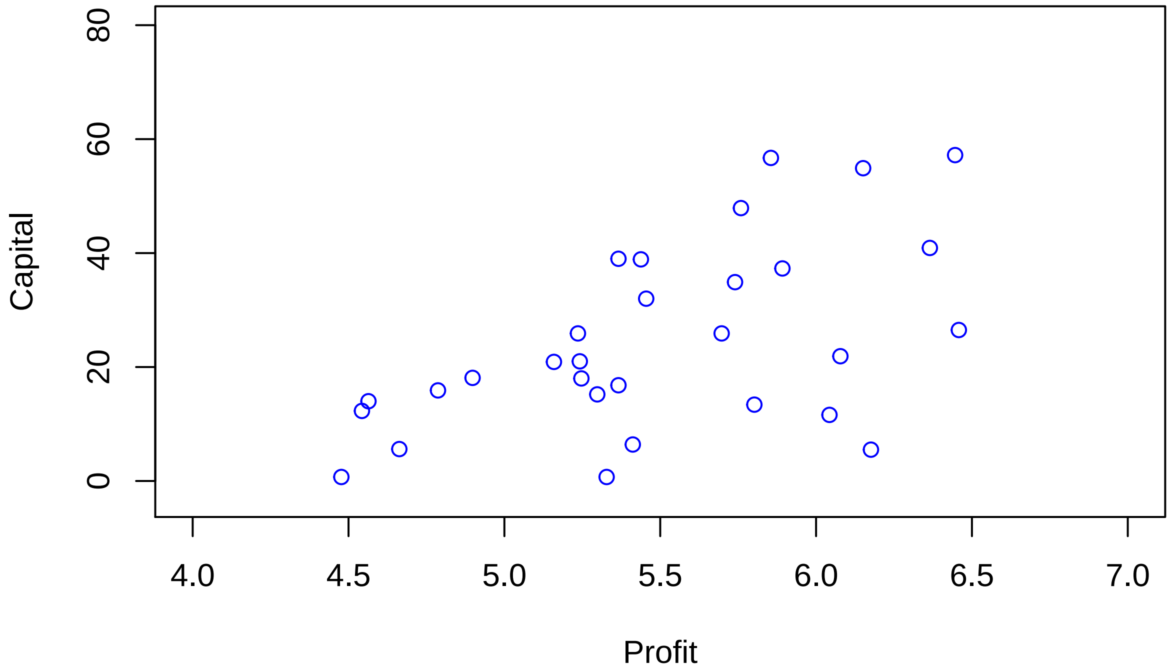
```r
corrplot.mixed(cor(bank[, -1]),
               lower = "number",
               upper = "ellipse",
               order = 'AOE')
```



Scatterplot

`plot()`

```
plot(ratings$profit, ratings$rating,
     xlab='Profit', ylab='Capital',
     ylim = c(-3,80), xlim = c(4,7),
     type = "p",
     col = 'blue')
```

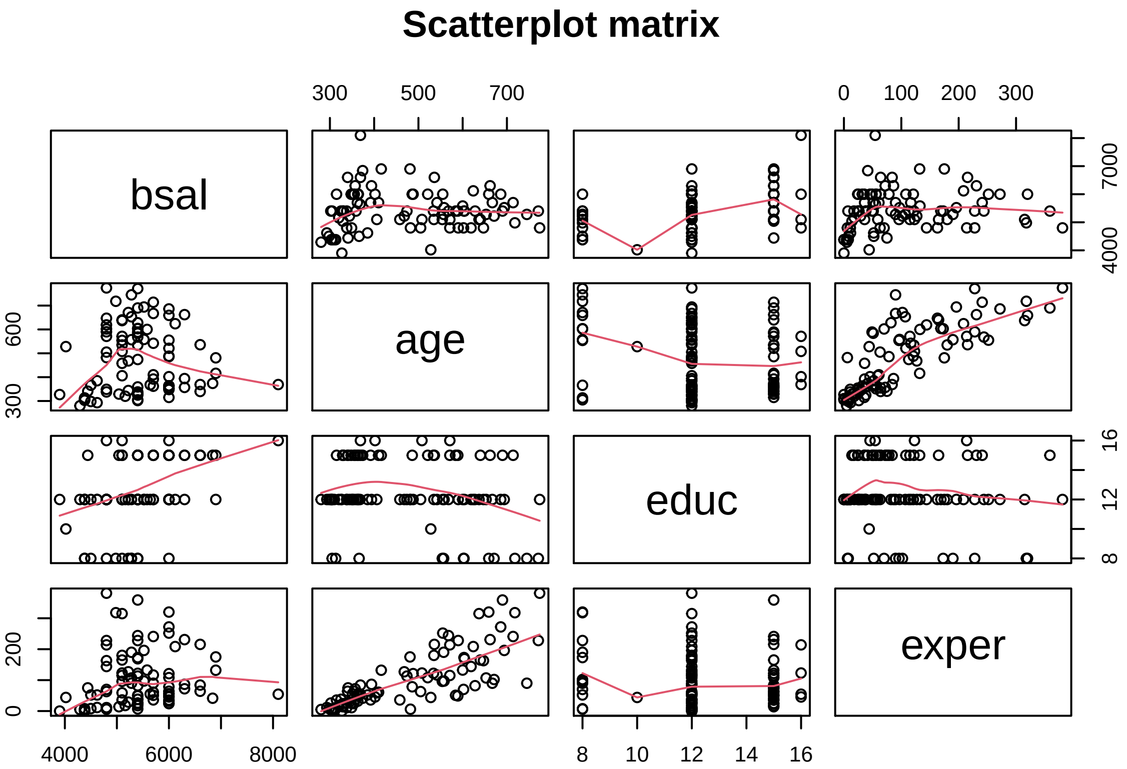


No particular trend is visible, it seams that the two variables are positively correlated with eachother since higher value of *profit* generarly corrispond to higher value of *capital*. There are too few observation to advance any hypotesis, we whould look directly at the correlation value. There i no single poin deviating significantly form all the others.

Scatter plot matrix

pairs()

```
pairs(bank[, c(1, 4, 5, 6)],
      panel = panel.smooth,
      #lower.panel = NULL,
      main = "Scatterplot matrix")
```



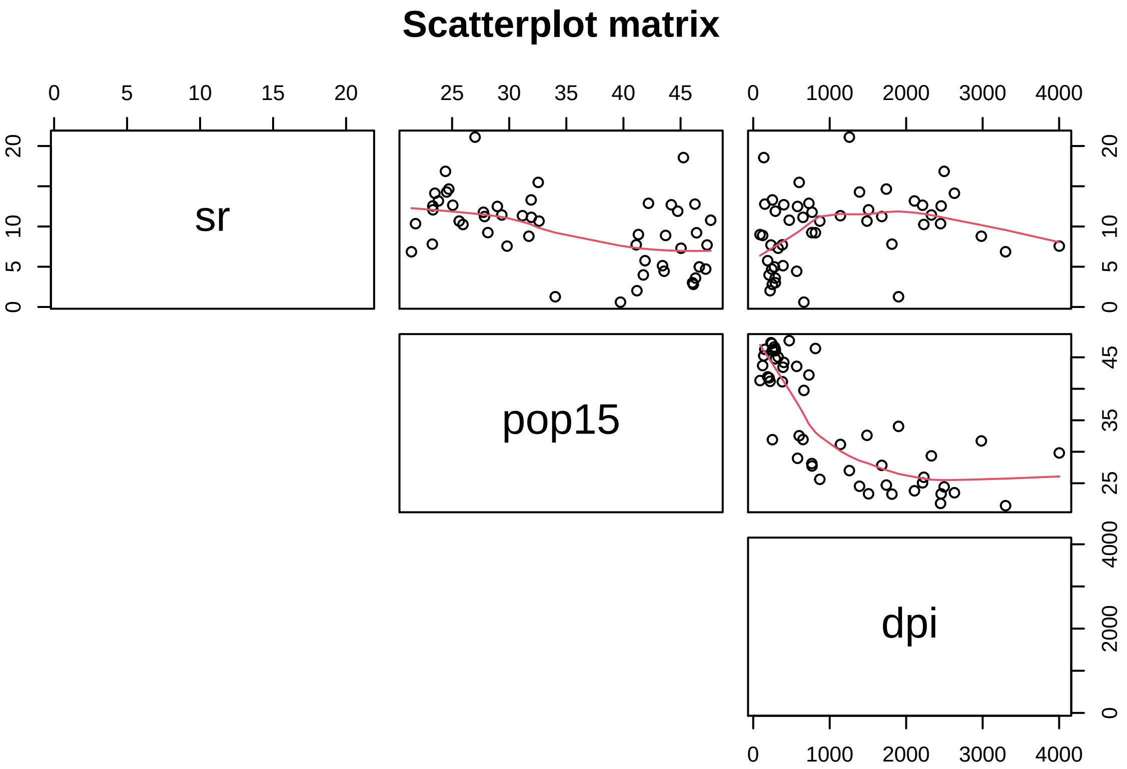
The age (*age* variable) and work experience (*exper* variable) exhibit a very similar behavior. In both cases, the relationship with the initial salary amount (response variable *bsal*) is quadratic. The points are arranged on the Cartesian plane following an approximately parabolic shape (concave downwards): as age and work experience increase, the value of the salary initially tends to increase, and then stabilizes and slightly decreases. In particular, the initial salary growth is significant:

- up to approximately 35 years of age, when considering the *age* variable;
- up to approximately 6 years of work experience, when evaluating the *exper* variable.

Additionally, there is a worker with a particularly high initial salary ($8100), characterized by a rather young age (369 months, about 31 years old), only 4 and a half years of work experience, but a long education phase (16 years). Regarding the years of education, it is evident from the graph that the variable can take a limited number of values (usually 8, 12, or 15 years, in addition to a limited number of workers with 16 years and only one with 10). The corresponding scatter plot suggests a positive association with the initial salary amount.

Another one:

```
pairs(dat,
      panel = panel.smooth,
      lower.panel = NULL,
      main = "Scatterplot matrix")
```



- Considering the two scatter plots related to the *pop15* variable, we preliminary observe that points are divided into two well-separated groups, corresponding to percentages of population under 15 from 20% to 35%, and from 35% to 50%, respectively.
- The presence of this pattern complicates the interpretation of a potential association. We observe that variables *pop15* and *dpi* present a strong but not linear negative association: an increase in the percentage of population under 15 implies a decrease in the values of per capita income, and vice-versa.
- The association between variables *pop15* and *dpi* seems to be slightly negative: an increase in the percentage of population under 15 corresponds to a very slight decrease in the saving rate. However, the points are very disperse around the red trend line, and the association seems not to be linear.
- Finally, also variables *sr* and *dpi* show a non linear association. In particular, the points seems to follow a positive trend for small amounts of per capita income (up to about 1000$), and a negative one for higher amounts (from 2000$ on). It is also important to remark that most of the points are concentrated in the left part of the plot, corresponding to small values of the *dpi* variable.

Nonparametric bootstrap

Bootstrap generation

boot()

```
#parameter of interest
mu_dif =   mean(data1$Server_A) - mean(data1$Server_B)
#Function to calcolate the parameter of interest
mean_diff <- function(data1, indices) {
  mean(data1[indices,"Server_A"])  - mean(data1[indices,"Server_B"])
}
```

```
library(boot)
set.seed(123)
```

```
n_rep <- 1000
Tboot <- boot(data = data1, statistic = mean_diff, R = n_rep)
round(sd(Tboot$t),3)
```

```
## [1] 2.115
```

We have ottened 1000 bootstrap replications to obtain a standard error of 2.115 on the parametrer of interest, it seems to be not very high.

Alternative:

bootstrap()

```
require(bootstrap)
```

```
## Loading required package: bootstrap
```

```
##
## Attaching package: 'bootstrap'
```

```
## The following objects are masked from 'package:faraway':
##
##     diabetes, hormone
```

```
set.seed(130)
n <- length(A)
aux_fun <- function(ind) {
  A.boot <- A[ind]
  B.boot <- B[ind]
  mean(A.boot) - mean(B.boot)
}

mean.boot <- bootstrap::bootstrap(x = 1:n,
                                  nboot = 1000,
                                  theta = aux_fun)

summary(mean.boot$thetastar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   10.62   12.06   12.11   13.50   19.50
```

```
sd(mean.boot$thetastar)
```

```
## [1] 2.118793
```

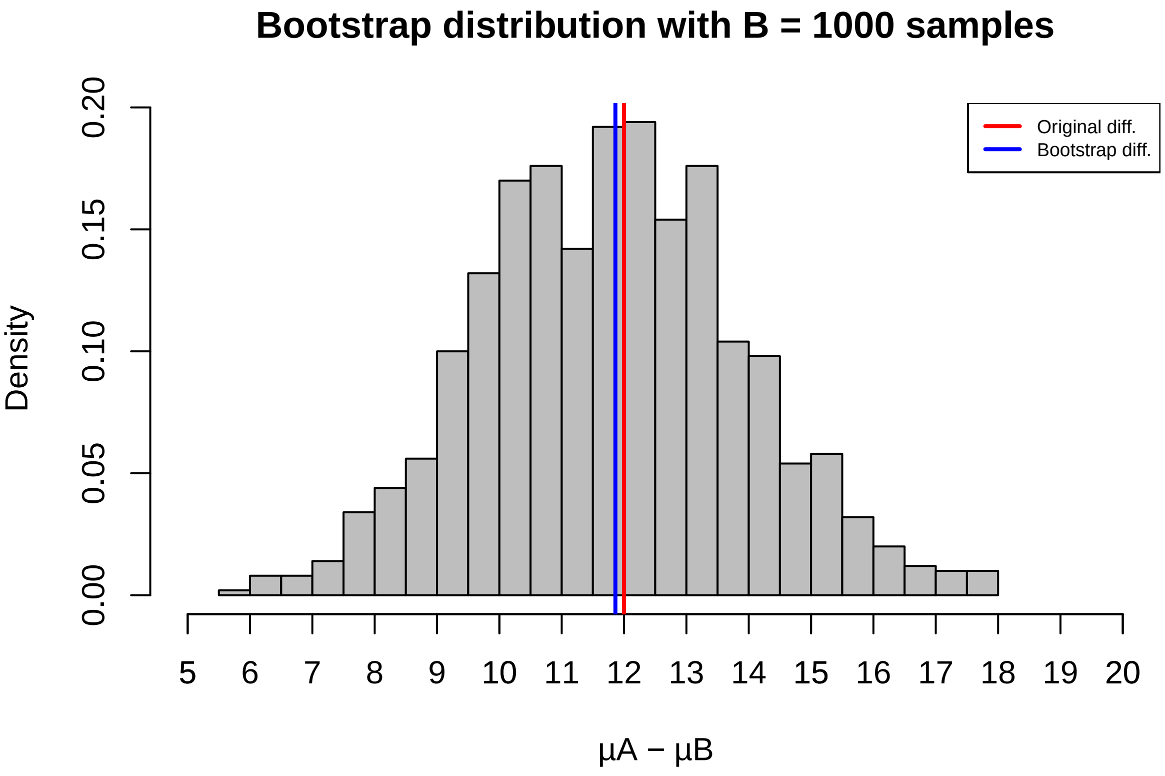
Considering 1000 bootstrap repetitions, the difference between the means with servers A and B varies from a minimum of 6 to a maximum of almost 20. The average value of the difference is instead equal to 12, approximately equal to the median. Both indices coincide with the value of the difference between the means calculated on the original sample. The standard error for the estimated mean difference is calculated as the standard deviation of the 1000 bootstrap estimates. This value is quite low, approximately equal to 2, indicating a good accuracy in the estimation of the correlation coefficient.

Bootstrap confidence intervals

Bootstrap distribution

```
hist(Tboot$t)
abline(v = c(mu_dif, mean(Tboot$t)))
```

```
hist(Tboot$t, breaks=40,
     freq=FALSE,
     main = "Bootstrap distribution with B = 1000 samples",
     xlab = "µA - µB", xlim = c(5,20),
     col= "gray",
     ylab = "Density")
axis(side=1, at=seq(5, 20, by=1), labels=seq(5, 20, by=1))
abline(v = c(mu_dif, mean(Tboot$t)),
       col = c("red", "blue"),
       lwd = c(2, 2))
legend("topright",
       c("Original diff.", "Bootstrap diff."),
       col = c("red", "blue"),
       lwd = c(2, 2),lty = c(1, 1),
       cex = 0.6)
```



As previously noted, the values taken by the bootstrap differences range from a minimum of 4 to a maximum of 20. The distribution is approximately symmetric around a central peak, located at the value of 12, which coincides with both the difference between the means calculated on the original data and the mean of the 1000 bootstrap differences.

Cofidence interval

quantile()

```
Q <- quantile(Tboot$t, c(0.025, 0.975))
Q
```

```
##   2.5%  97.5%
##  7.750 16.125
```

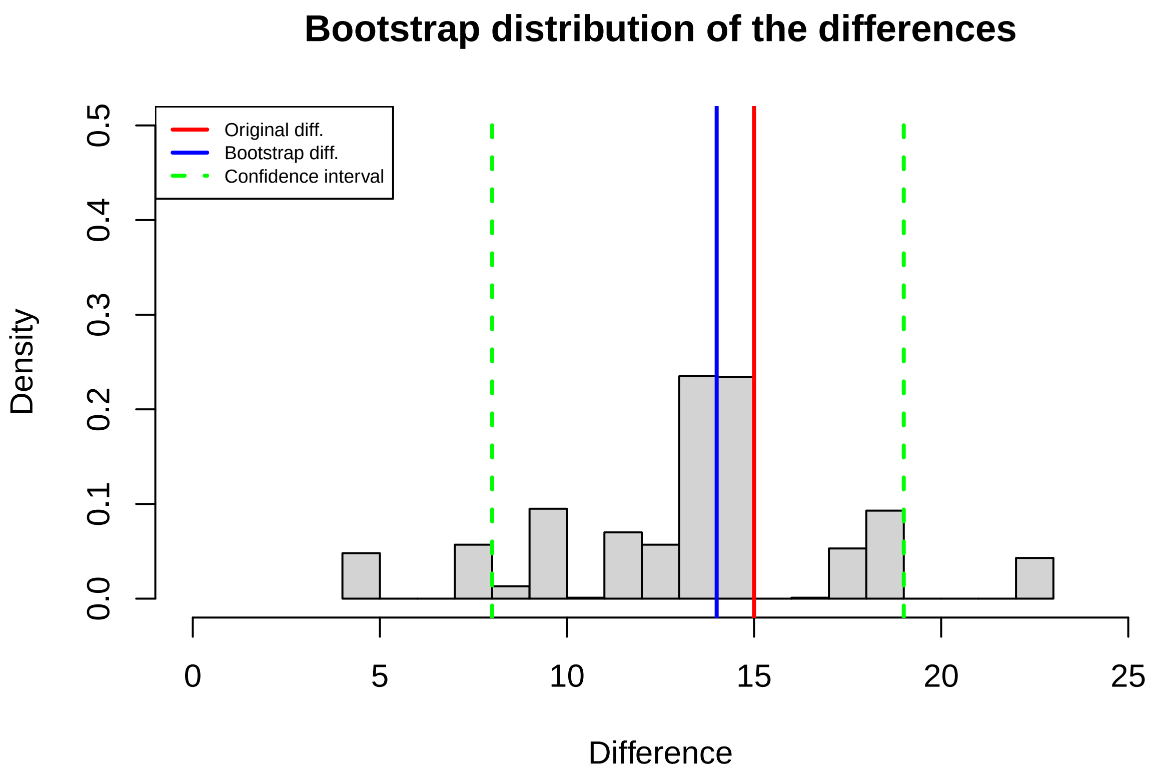
The 95% confidence interval for the difference is therefore [8.13, 16.25]. It represents a range of values centered around the point estimate of the parameter. Note that 0.95 is not the probability that the point estimate falls within the interval: on the contrary, the interval includes the point estimate in 95% of cases (95% of bootstrap replications).

Confidence interval on the distribution

```
hist(Tboot$t)
abline(v = c(median_dif, median(Tboot_2$t), Q2))
```

```
#parameter of interest
median_dif <- median(data1$Server_A) - median(data1$Server_B)
#Function to calcolate the parameter of interest
median_diff <- function(data1, indices) {
  median(data1[indices, "Server_A"]) - median(data1[indices, "Server_B"])
}
set.seed(123)
n_rep <- 1000
Tboot_2 <- boot(data = data1, statistic = median_diff, R = n_rep)
Q2 <- quantile(Tboot_2$t, c(0.05,0.95))
```

```
hist(Tboot_2$t,
     main = "Bootstrap distribution of the differences",
     breaks = 25,
     freq = FALSE,
     ylab = "Density", xlab = "Difference",
     xlim = c(0, 25), ylim = c(0, 0.5))
abline(v = c(median_dif, median(Tboot_2$t), Q2),
       col = c("red", "blue", "green", "green"),
       lwd = c(2, 2, 2, 2),
       lty = c(1, 1, 2, 2))
legend("topleft",
       c("Original diff.", "Bootstrap diff.", "Confidence interval"),
       col = c("red", "blue", "green"),
       lwd = c(2, 2, 2),
       lty = c(1, 1, 2),
       cex = 0.6)
```



The plot confirms the broad range of the confidence interval, which includes most of the bootstrap estimates for the difference between medians: only along the left-hand side of the distribution are some values observed outside the interval, also some values are presented in the right-hand side. Both the value of the difference calculated on the original dataset and the mean of the 1000 differences calculated with the bootstrap method fall in the central area of the interval.

Multiple linear regression model

Summary

`summary(lm())`

```
mod <- lm(data = bank, formula = bsal ~ educ + exper)
summary(mod)
```

```
##
## Call:
## lm(formula = bsal ~ educ + exper, data = bank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1286.42  -404.50    25.66   365.71  2285.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3569.9077   385.6639   9.257 1.01e-14 ***
## educ         134.7096    29.2112   4.612 1.32e-05 ***
## exper          1.6430     0.7331   2.241   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 636.2 on 90 degrees of freedom
## Multiple R-squared:  0.2136, Adjusted R-squared:  0.1961
## F-statistic: 12.22 on 2 and 90 DF,  p-value: 2.011e-05
```

The function summary() shows different information about the estimated model:

- Descriptive statistics about the regression residuals: we observe that the range of the residuals is extremely wide, indicating that for some statistical units, the deviation between the observed value and the value predicted by the model is very high. The median value is slightly different from 0 (as expected based on theoretical assumptions). The quartiles are equidistant from the center of the distribution, which may suggest symmetry. In conclusion, it is possible that the residuals are distributed according to a Gaussian random variable, but the variability appears to be too high. Further analysis of the residuals will be necessary to verify the validity of the assumptions underlying the model.

- Estimates of the regression model parameters: the estimated values of the intercept and each of the covariates included in the model are reported, along with their standard errors and the results of the T test for each regression parameter.

  - The estimated coefficient $\hat{\beta}_0$ for the intercept represents the expected value of the 0 response variable when all explanatory variables have a value of zero. Note that frequently (including in this case), this value may not have practical interpretability; it does not make sense to consider the salary value for a worker with 0 years of education. The T test associated with this coefficient has a test statistic value of 9.26 and a corresponding p-value of the order of $10^{-14}: the null hypothesis $H_0 : \beta_0 = 0$ is rejected at any level of significance.

  - The estimated coefficient $\hat{\beta}_1$ for the number of years of education represents the 1 expected increase of the response variable for a unit increase in the educ variable while holding the remaining covariates constant. In other words, an increase of one year in the education period (while keeping the number of previous work months constant) corresponds to an increase in the initial salary of about 135. The p-value associated with the T test is also sufficiently close to zero to reject the null hypothesis that the estimated coefficient $\hat{\beta}_1$ is zero.

  - The same in terpretation applies to the estimated coefficient $\hat{\beta}_2$ for the exper covariate: considering the number of years of education fixed, for each additional month of previous work experience, the initial salary increases by about 1.5.

- The residual standard error (RSE), which is approximately 636, along with its degrees of freedom: 90, i.e., $n - p - 1 = 93 - 2 - 1 = 90$, where p represents the number of covariates and one is subtracted if the model contains an intercept. This is the square root of the ratio of the sum of squares of residuals and the number of degrees of freedom. The RSE can be interpreted as the average deviation around the mean of residuals, which is assumed to be zero, and thus as the average deviation between observed and corresponding interpolated values. In other words, this value states that, on average, interpolated values deviate from observed ones by 636. Additionally, the percentage error can be obtained: it is sufficient to take the ratio with the sample mean value of the score: 636.2/5420.3 = 0.11, hence an error of 11.7%.

- The multiple R-squared and its adjusted value. The multiple linear determination coefficient represents the ratio of the variance explained by the interpolating plane and the total variance of the response variable. In this case, a (rather low) value of 0.21 is observed, indicating that the interpolating plane explains approximately 21% of the variability of the initial salary. Note that this is a goodness-of-fit index, which cannot detect whether the model has been correctly specified.

- The results of the F-test, i.e., the test statistic value, which is 12.22, and the corresponding p-value (of the order of 10−5). The proximity to 0 of the p-value allows us to reject the null hypothesis that all regression coefficients, except for the intercept, are equal to 0.

Alternative comment:

```
lm1 <- lm(eta ~ ., data = shell)
summary(lm1)
```

```
##
## Call:
## lm(formula = eta ~ ., data = shell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5848 -1.1779 -0.2789  0.8193  8.2534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7049     0.9036   2.994  0.00299 **
## lun           9.0275     7.0593   1.279  0.20198
## dia           1.9965     8.2398   0.242  0.80872
## alt          26.5671     8.9494   2.969  0.00324 **
## pesot        11.5984     2.8257   4.105 5.26e-05 ***
## pesom       -22.8850     3.1683  -7.223 4.43e-12 ***
## pesov       -12.2573     5.0720  -2.417  0.01628 *
## pesog         1.1417     4.3483   0.263  0.79307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.164 on 292 degrees of freedom
## Multiple R-squared:  0.5657, Adjusted R-squared:  0.5553
## F-statistic: 54.34 on 7 and 292 DF,  p-value: < 2.2e-16
```

The descriptive statistics for the residuals show values between -7.5 and 8.2, this values are over the [-3,3] interval of the normal distribution, however further analysis are required to verify it. The median is very close to zero, and the overall distribution seems to be symmetrical around it.

The estimated value for the intercept term is equal to $\hat{\beta}_0 = 2.7$: this means that according to the model, if all the variable are fixed to zero the abalone will have 2.7 years, this is obviously a theoretical event. The corresponding standard error is equal to $SE_{\hat{\beta}_0} = 0.9$, it is used to asses the significantly of the estimated parameter through the T test. The test statistics is equal to 2.99 and the relative p-value is extremely small (under 0.003). In this case we can reject the null hypothesis $H_0 : \hat{\beta}_0 = 0$, concluding that the intercept term provide a significant contribution in explaining the response variable. However is important to remember that this type of test is not significant for feature selection so other analysis should be done in this sense, like the AIC or the BIC.

The same conclusion holds for all the explanatory variables: the very small values of the p-values for *pesot* and *pesom* lead us to reject the null hypothesis of non significance at every confidence level, we can also reject it for *alt* and *pesov* but with a confidence level slightly inferior. For the other variables we can not reject they null hypothesis, however as seed before this is not sufficient for not using them in our model.

The residual standard error (RSE) is equal to 2.164; it means that, on average, the observed and the estimated values for each sample unit differ by 2.164. Although the value seems small, it needs to be further evaluated with respect to a centrality index (e.g., the mean) of the response variable. Computing the ratio between the RSE and the mean of the *eta*, we obtain the percentage error, equal to $2.164/11.3 = 0.19$: on average the percentage error between observed and fitted values is equal to 19%.

The multiple R^2 shows an weak fit of the model to the data; being equal to 0.57 (with an adjusted value approximately equal). This means that only 57% of the total variability of the response variable can be explained by the estimated model. Further analysis are necessary to improve this value.

Finally, let us consider the F test having the joint non-significance of all regression coefficients as null hypothesis (the alternative hypothesis considers instead that at least one coefficient is statistically significant). The value of the F statistics is equal to 54.34; the associate p-value is very close to 0 ($<$ 2.2e-16), leading us to reject the null hypothesis at each confidence level.

In general the results of this model are not good, since the level fo R^2 is pretty low. We should check for a feature selection that could improve the model. In addition checking for outliers or influence points can be also useful to improve the model.

Fitted values and residuals

```
mod$fitted[]
mod$residuals[]
```

```
mod <- lm(data = dat, formula = sr ~ pop15 + dpi)

# observed value
dat["Italy", 1]
```

```
## [1] 14.28
```

```
#fitted value
round(mod$fitted["Italy"], 2)
```

```
## Italy
## 12.79
```

```
#residual
round(mod$residuals["Italy"], 2)
```

```
## Italy
##  1.49
```

- The estimated and observed values for the Italian saving rate are equal to $y = 12.28$ and $\hat{y} = 12.79$, respectively.
- The corresponding residual, computed as their difference, is therefore $r = y - \hat{y} = 1.49$. It represents the distance between the true and the estimated value for the response variable (computed for a specific single unit).
- Comparing this value with the descriptive statistics of the residuals (obtained through the summary of the model, see point 12.7), we can observe that it corresponds almost perfectly with the third quartiles of the residuals distribution: the residual value com- puted for Italy is greater than 75% of the residuals of the other countries.

Estimate for variance

```
sum(mod$residuals2)/(n-p)
```

```
n <- dim(dat)[1]
p <- dim(dat)[2]
s2 <- sum(mod$residuals^2)/(n-p); s2
```

```
## [1] 15.83242
```

```
s <- sqrt(s2); s
```

```
## [1] 3.978997
```

The value of s, which is also reported in the summary of the model (see point 12.7), represents the average variability of the residuals around their mean, which is equal to 0 by hypothesis. Hence, s represents the average distance between the observed and estimated values. A low value of denotes a good adaption of the model to the the given data. More specifically, we can use s^2 to decompose the total variability of the response variable between explained and residual components.

Variance decomposition:

```
TSS <- (n-1) * var(dat[, 1]); TSS
```

```
## [1] 983.6282
```

```
SSR <- (n-p) * s2; SSR
```

```
## [1] 744.1237
```

```
SSE <- TSS - SSR; SSE
```

```
## [1] 239.5045
```

We observe that the total variability (measured through the deviance indicator) is equal to 983.63; the estimated model is able to explain an amount equal to 239.51, while the remaining 744.12 is residual. This way we can also compute again the multiple R-squared index, as ratio between explained and total variability.

```
R2 <- SSE/TSS; R2
```

```
## [1] 0.2434909
```

Confidence interval for parameters

```
confint(mod, level = 0.9))
```

```
mod <- lm(data = diamonds, formula = carat ~ .-width)
```

```
round(confint(mod, level = 0.9), 3)
```

```
##                5 %   95 %
## (Intercept) -1.206 -1.088
## cut2        -0.065 -0.007
## cut3        -0.075 -0.019
## cut4        -0.065 -0.010
## cut5        -0.085 -0.031
## price        0.027  0.032
## length       0.124  0.194
## depth        0.214  0.328
```

We first observe that none of the above confidence interval contains 0, so that we can consider each explanatory variable (as well as the intercept) significant at a confidence level of 90%. This is the exact same result obtained through the significance T test (with significance level equal to 0.9). As an example, we notice that with probability 0.9 the increase in the carats corresponding to a unitary increase in the price (with all the others covariates kept fixed) lies between 0.027 and 0.032. Similarly, considering the categorical variable cut, the decrease in the response variable corresponding to a change from worst to best quality cut (maintaining all the other variables constant) falls between -0.085 and -0.031 with confidence level of 0.9.

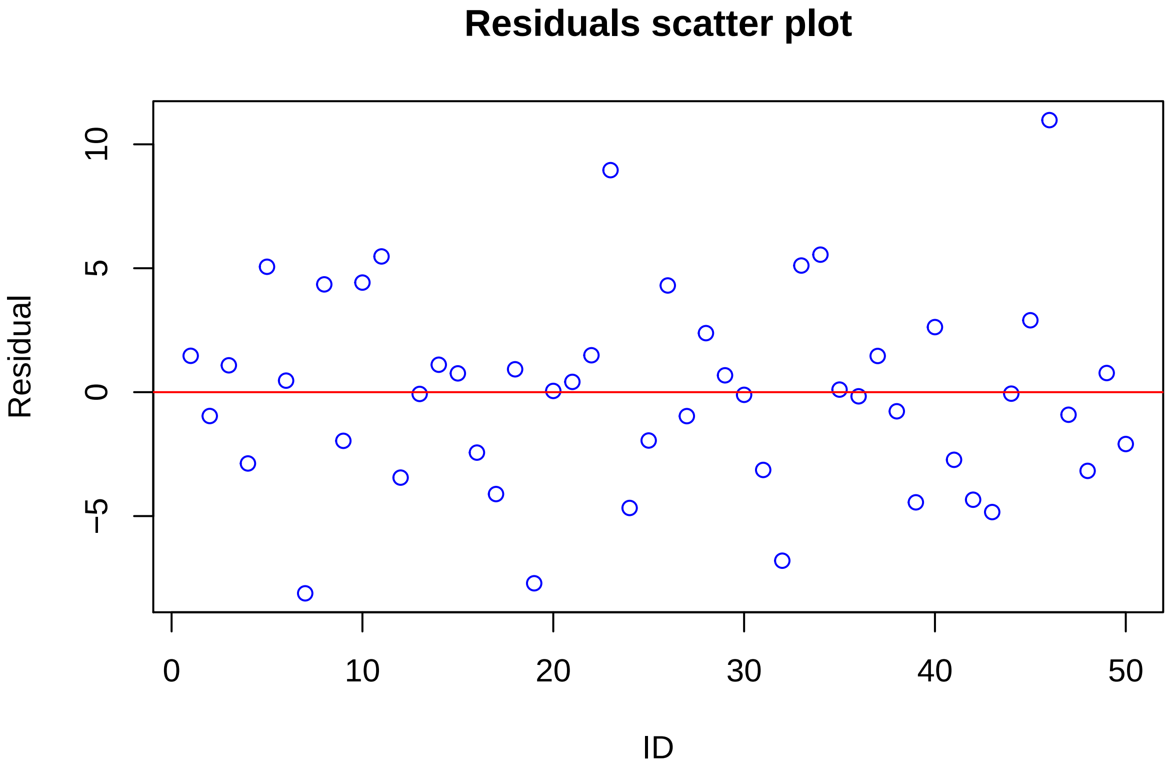
Graphical inspection of residuals

```
mod <- lm(sr ~ dpi + pop15, dat)
```

Residuals scatter plot

```
plot(mod$residuals)
```

```
plot(mod$residuals,
     main = "Residuals scatter plot",
     xlab = "ID", ylab = "Residual",
     col = "blue")
abline(h = 0, col = "red")
```

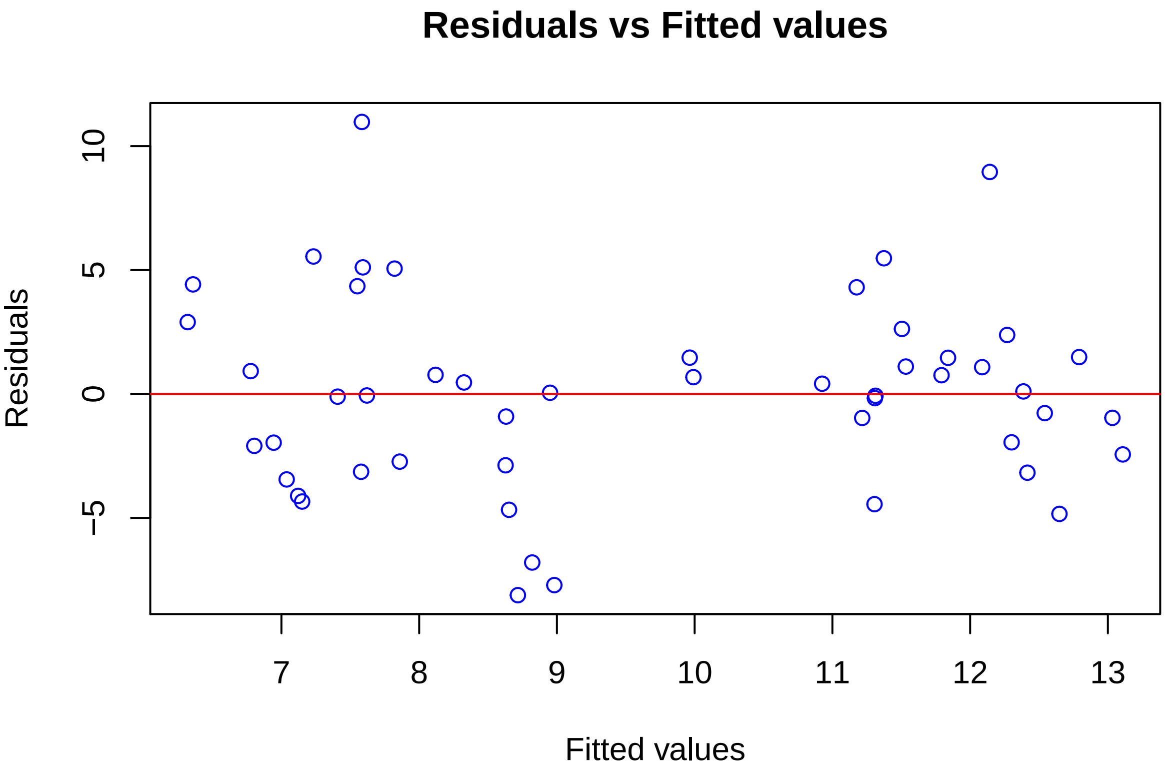


Firstly, as already observed by the descriptive statistics returned by the lm() function, we notice that, although slightly broader than the expected, the range of variation of the residuals is quite limited, ranging approximately from a minimum of -10 to a maximum of 10. There are, therefore, no residuals that assume particularly high values, neither positive nor negative. However, we note the presence of a small number of values that slightly deviate from the rest of the point cloud (see, e.g., observation with ID 46 in the upper part of the plot). We then observe that there are no particular trends or patterns in the arrangement of the points in the plane; they seem to be arranged in a substantially random manner. Moreover, we can note a concentration of points that is approximately equivalent between the positive and negative half-plane, with positive and negative residuals alternating randomly. We can, therefore, validate the hypothesis of the lack of correlation among the residuals (i.e., that they are uncorrelated random variables and therefore independent) and the hypothesis of constant variance of the residuals as the units vary (i.e., that phenomena of the type of residuals of the first units being very concentrated around zero and those of subsequent observations being very dispersed do not occur).

Residuals vs fitted values

```
plot(mod$fitted, mod$residuals)
```

```
plot(mod$fitted,
     mod$residuals,
     main = "Residuals vs Fitted values",
     xlab = "Fitted values", ylab = "Residuals",
     col = "blue")
abline(h = 0, col = "red")
```

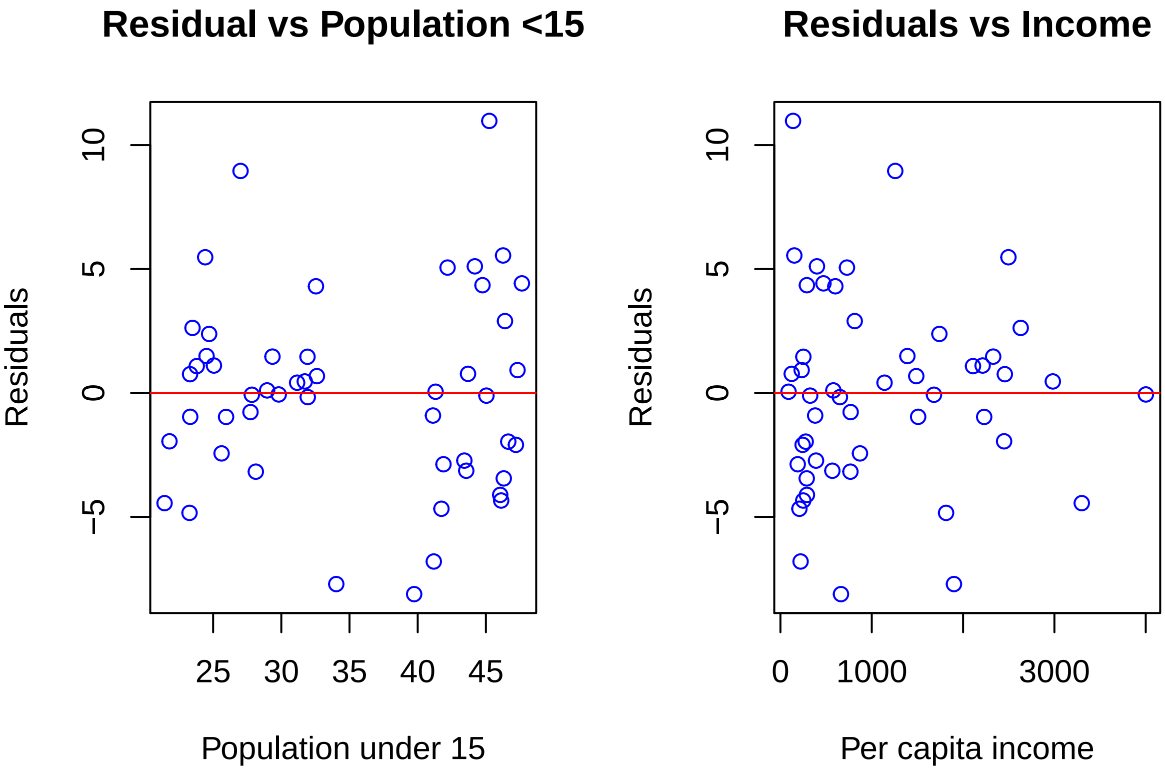


The points appear to be clustered into two distinct and well-separated groups. However this behavior seems to be due to the peculiar distribution of the fitted values (x-axis): in both groups residuals are distributed quite randomly, without the presence of particular systematic trends, and tend to arrange themselves in a cloud of points that is generally elliptical in shape. Moreover, as already noticed, we observe that the positive and negative residuals alternate randomly and are divided almost equally. Again, we notice the presence of some points that deviate from the two clouds of points, especially in the left part of the graph.

Residuals vs covariates

```
plot(dat[,1], mod$residuals)
```

```
par(mfrow = c(1, 2))
plot(dat[, 2], mod$residuals,
     main = "Residual vs Population <15",
     xlab = "Population under 15", ylab = "Residuals",
     col = "blue")
abline(h = 0, col = "red")
plot(dat[, 3],
     mod$residuals,
     main = "Residuals vs Income",
     xlab = "Per capita income", ylab = "Residuals",
     col = "blue")
abline(h = 0, col = "red")
```

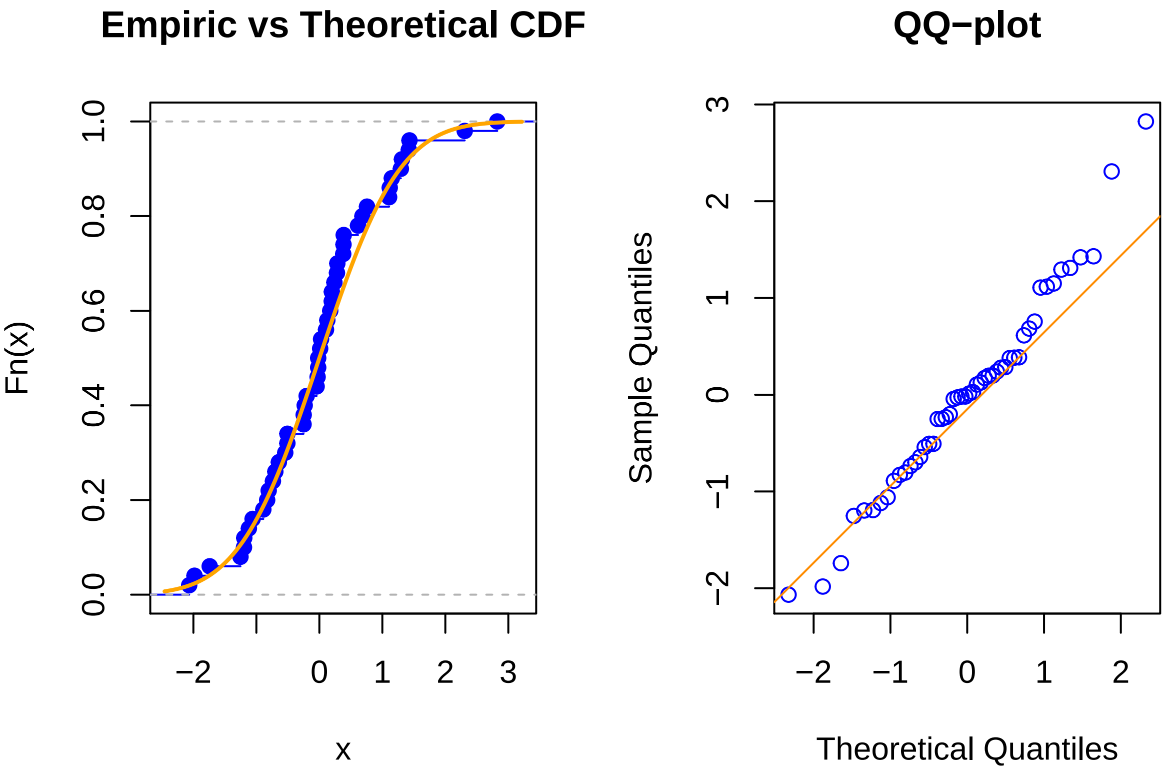


The first plot, regarding the percentage of population under 15, shows the same behavior highlighted in the previous plot: points are separated into two distinct groups, but again this is due to the distribution of the considered covariates. Once more, we remark the presence of a small number of residuals that deviates from the main portion of the points, especially in the right part of the plot. The second plot, concerning the per capita income, shows instead a substantially opposite situation. In this case, the arrangement of points is completely asymmetric: we observe, in the left portion of the graph, a very high density of points, with quite a long tail of values that deviate as we move towards high values of the income. We therefore highlight a systematic arrangement of points, without the presence of an elliptical arrangement of points. This plot suggests evidence against the hypothesis of a linear association between the response variable and this covariate: it may be necessary to perform a transformation (in this case, typically logarithmic) of this explanatory variable.

Empirical vs teoretical and Q-Q plot

```
plot(ecdf(stand_res))
qqnorm(stand_res)))
```

```
stand_res <- rstandard(mod)

par(mfrow = c(1, 2))
plot(ecdf(stand_res),
     main ='Empiric vs Theoretical CDF',
     col = "blue")
curve(pnorm(x),
      col = 'orange',
      lwd = 2,
      add = TRUE)

qqnorm(stand_res,
       main = "QQ-plot",
       col = "blue")
qqline(stand_res, col = "darkorange")
```

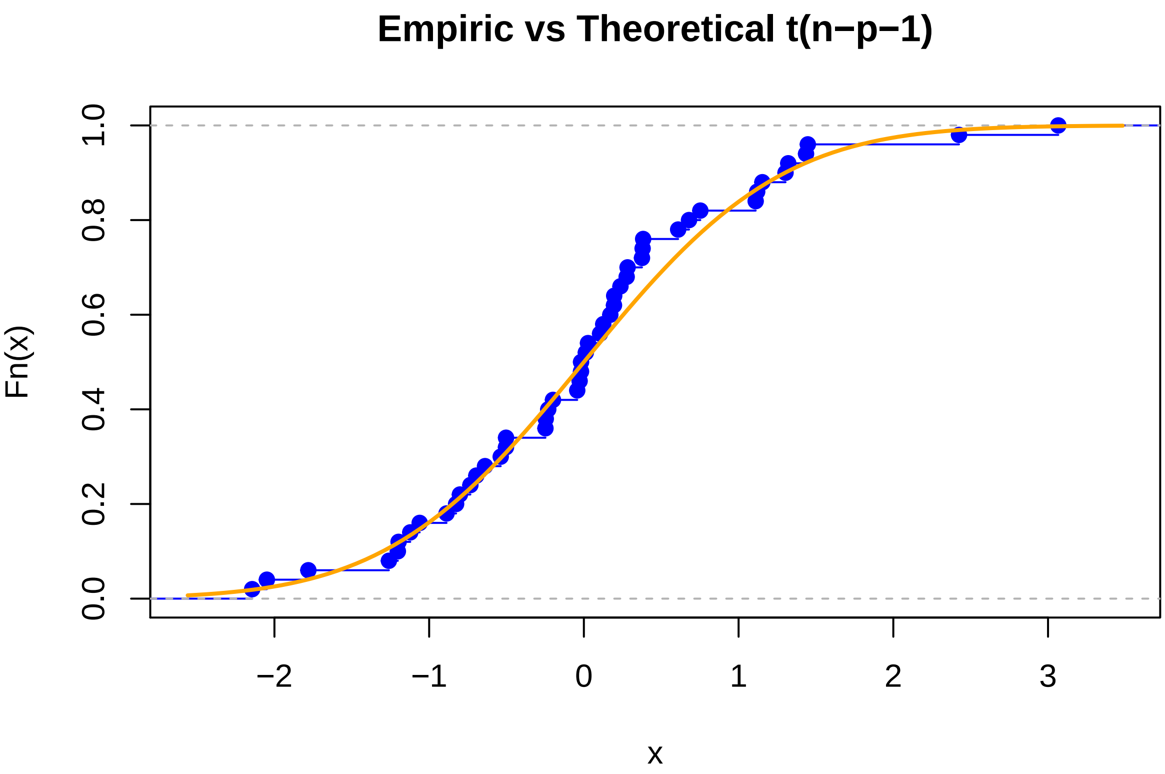


We observe that the distribution of the standardized residuals is approximately normal; in both plots the empirical values (represented as blue points) tend to follow the reference line of the theoretical values; only the QQ-plot (which is much more sensible) shows some points that slightly deviates on the right part of the graph, suggesting a right tail that is heavier than the expected one (under the hypothesis of normality). Summing up all the results obtained through the graphical analysis: residuals seems to follow approximately a Gaussian distribution, although presenting a right tail that is slightly heavier than the expected (i.e., we have a few positive residuals slightly bigger than the expected). The mean (and the median) of the distribution is equal to 0; the variance is surely greater than 1, but may be assumed as constant along all sample units. Residuals are also uncorrelated with the response variable. The dpi explanatory variable is likely to have a non-linear association with the response variable.

Alternative with studentize residual:

```
plot(ecdf(stud_res))
```

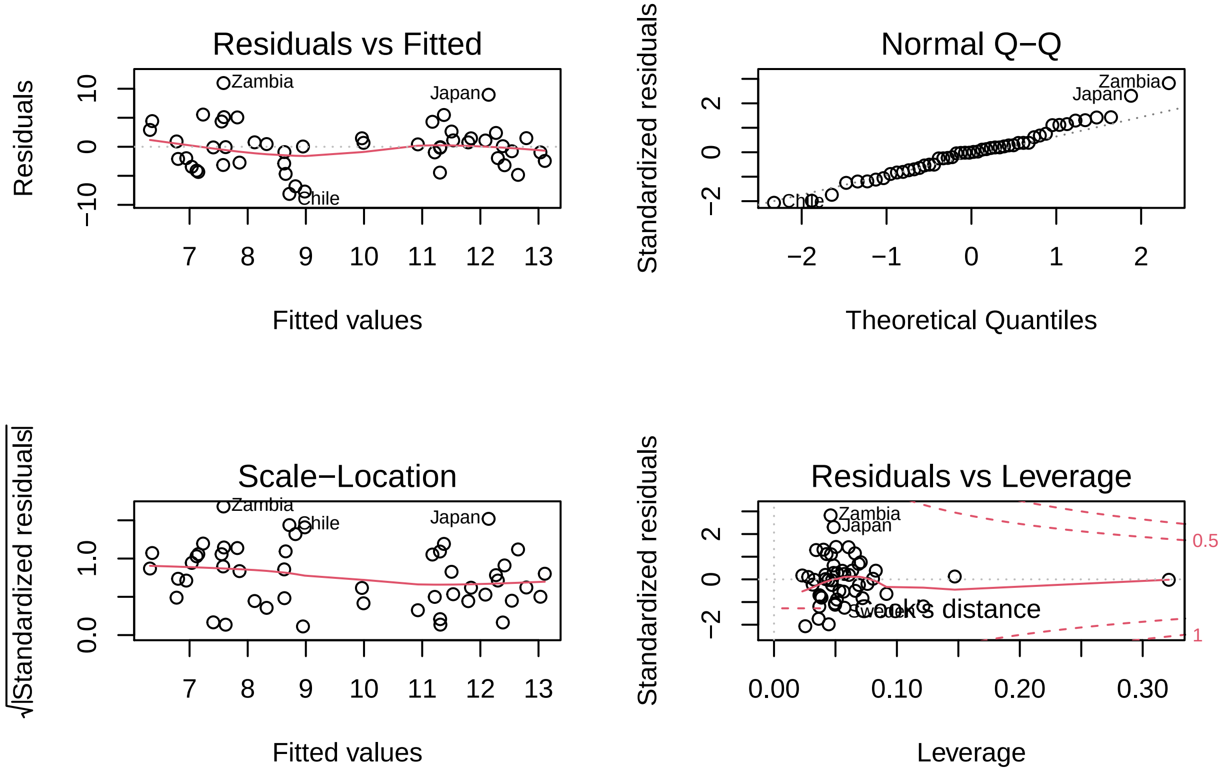
```
stud_res <- rstudent(mod)

plot(ecdf(stud_res),
     main ='Empiric vs Theoretical t(n-p-1)',
     col = "blue")
df <- dim(dat)[1] - dim(dat)[2]-1
curve(pt(x,df),
      col = 'orange',
      lwd = 2,
      add = TRUE)
```



Some alternative:

```
plot(mod)
```

```
par(mfrow = c(2, 2))
plot(mod)
```

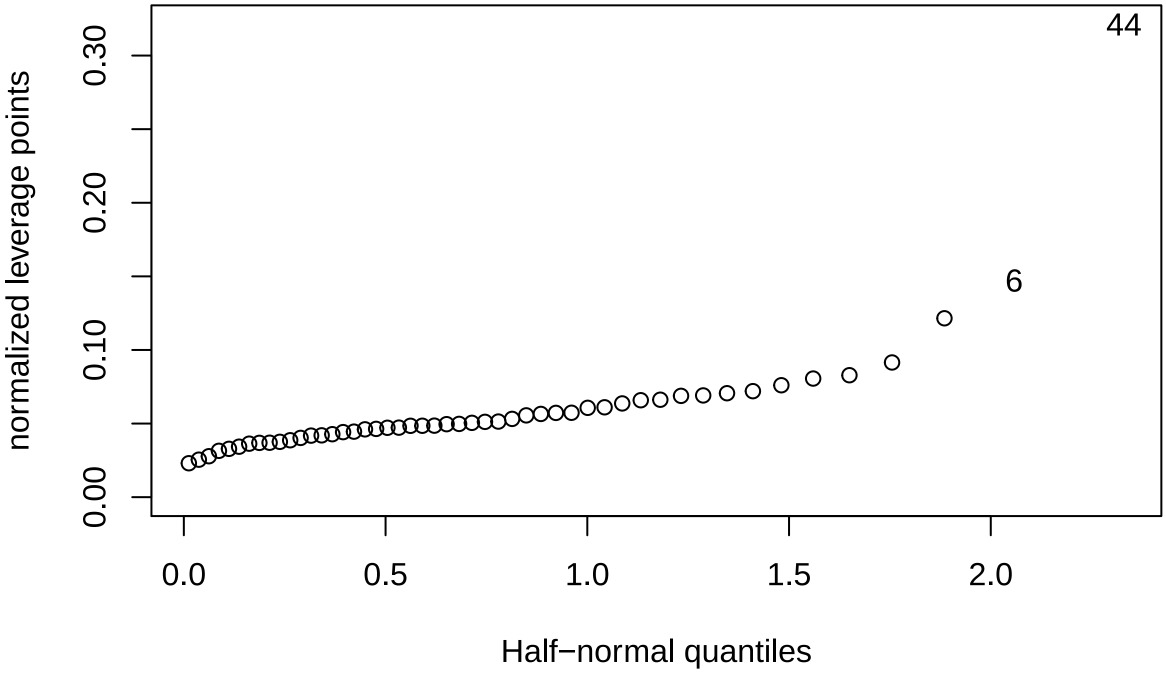


Decting unusual and influential observations

Leverage points

hlfnorm(hatvalues(mod))

```
lev <- hatvalues(mod)

library(faraway)
halfnorm(lev,
         ylab = "normalized leverage points")
```

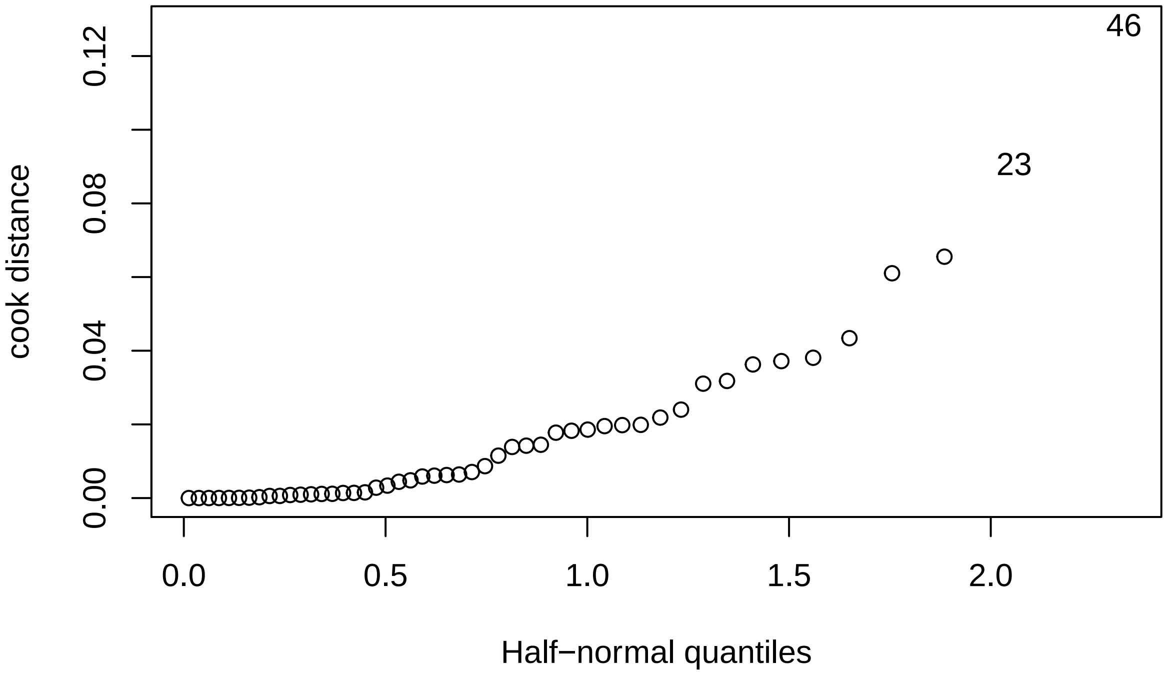


This function mesures the value of h_{ii} to evaluate those observations having a potential leverage effect, i.e., an high influence on the fit. In this case we notice that there are two observations which could potentially influence the fit: 6 and 44. However, all the values are very low so they are not problematic.

Influential values

hlfnorm(cooks.distance(mod))

```
cook <- cooks.distance(mod)
halfnorm(cook, ylab = "cook distance",
         pch=16, cex=0.4)
```



No point is identified as an ifluent value, since the Cook's distance is very low. Usually the problematic point has a value over 0.5 or over 1.

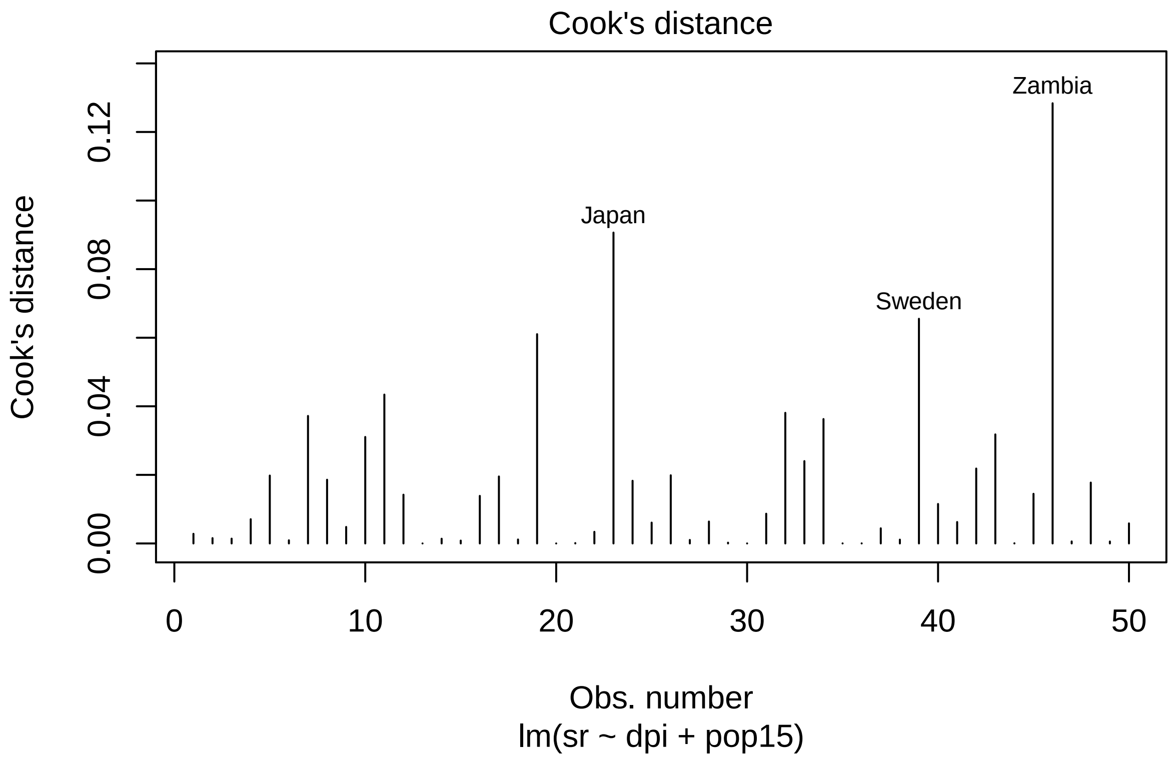
If we want to remove the observation we could write **dat**[**-46,**] in order to see if the model without this obs fit well, to do that we should check the two model's R^2.

Alternative graphs

Alternative 1:

```
plot(mod, which=4)
```

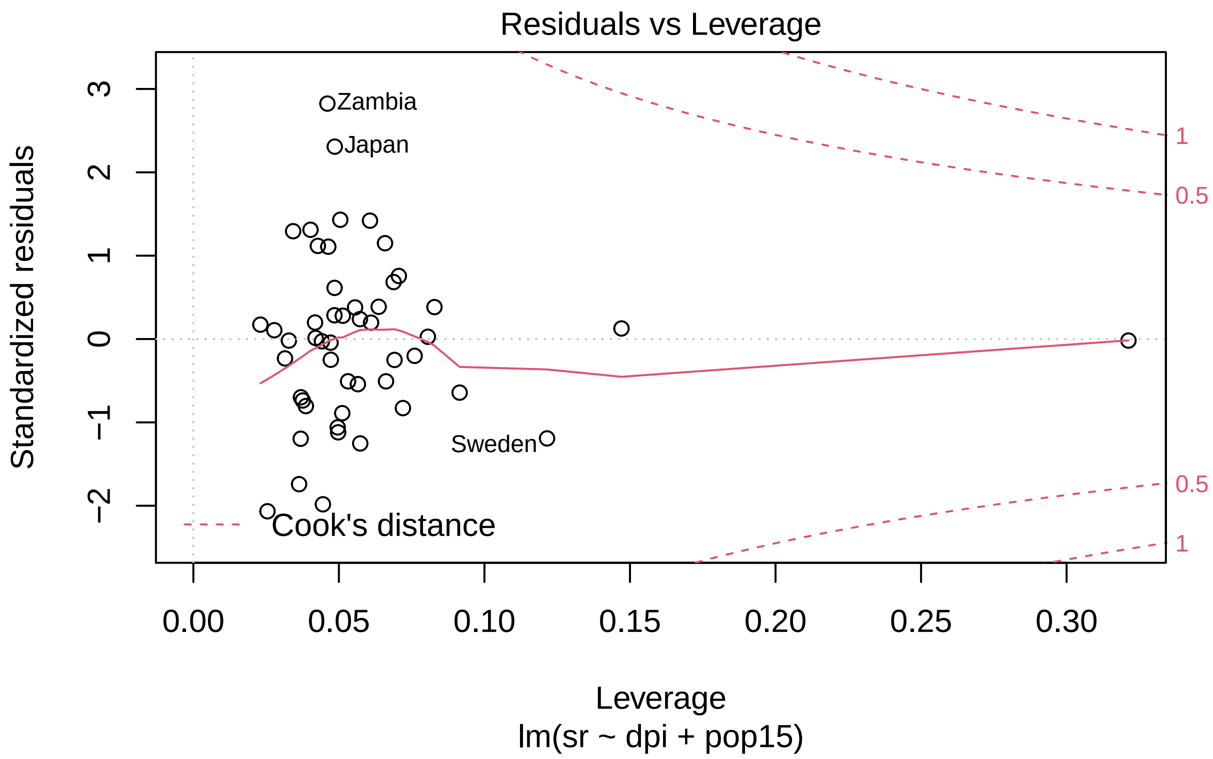
```
plot(mod, which=4)
```



Alternative 2:

```
plot(mod, which=5)
```

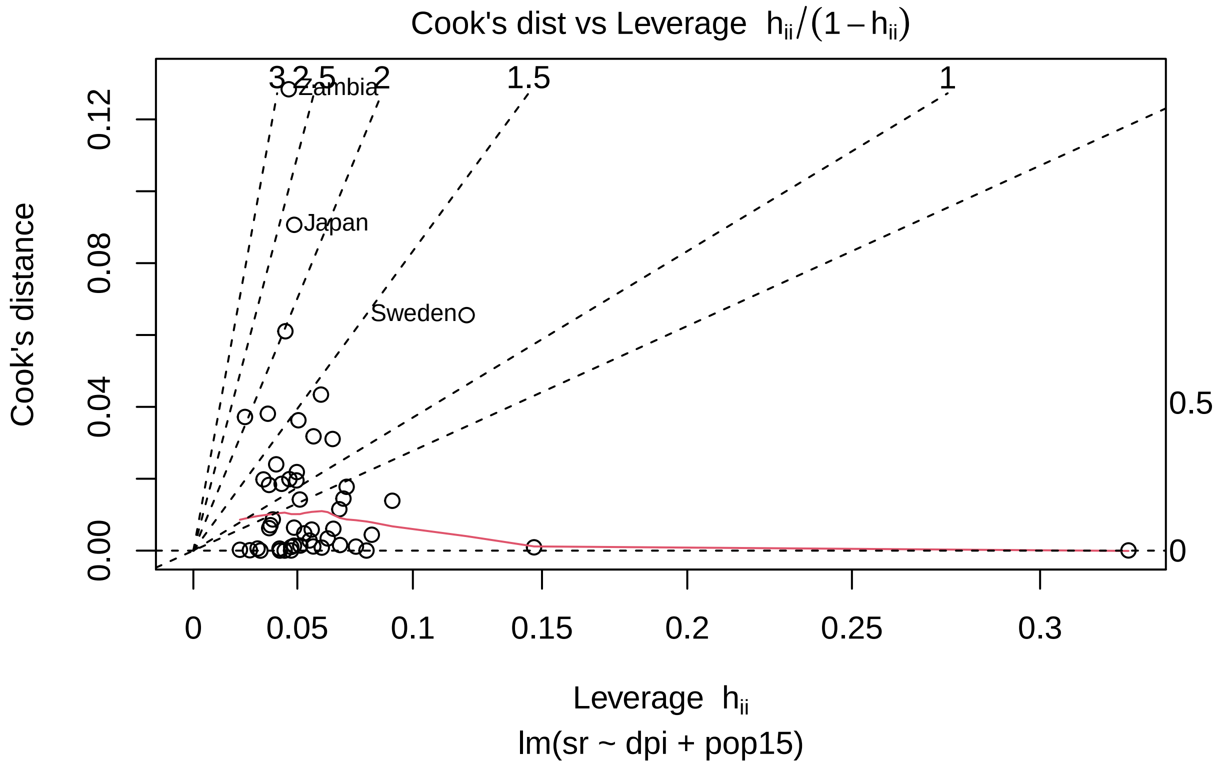
```
plot(mod, which=5)
```



Alternative 3:

```
plot(mod, which=4)
```

```
plot(mod, which=6)
```



Multicollinearity: Variance inflation factor

```
vif(mod)
```

```
mod.C <- lm(data = ratings, formula = rating ~ .)
```

```
library(faraway)
vif(mod.C)
```

```
##   profit  capital   f_flex 
## 1.831589 1.992200 1.937912
```

We observe that each explanatory variable is associated with a measure of VIF, calculated as the reciprocal of 1 minus the multiple linear regression coefficient of determination obtained by excluding the corresponding variable. A particularly high value for a certain variable (a possible criterion, but not the only nor binding one, suggests considering it as such if it exceeds 10) indicates the presence of excessive collinearity. The practical effect is that the standard error corresponding to that variable is higher compared to what it would be in the absence of collinearity; this leads to potential inaccuracies in the results of the T-test and in the computation of the confidence intervals for the regression coefficients. In the present case, all the VIF values are very small, ensuring that the corresponding explanatory variables are not collinear.

Model selection

```
step(mod)
```

```
library(stats)
step(mod.C)
```

```
## Start:  AIC=142.64
## rating ~ profit + capital + f_flex
##
##           Df Sum of Sq    RSS    AIC
## - profit   1      0.55 2669.0 140.65
## <none>                 2668.4 142.64
## - f_flex   1    533.32 3201.7 146.11
## - capital  1   1007.66 3676.1 150.25
##
## Step:  AIC=140.65
## rating ~ capital + f_flex
##
##           Df Sum of Sq    RSS    AIC
## <none>                 2669.0 140.65
## - f_flex   1    617.18 3286.1 144.89
## - capital  1   1193.52 3862.5 149.74

##
## Call:
## lm(formula = rating ~ capital + f_flex, data = ratings)
##
## Coefficients:
## (Intercept)      capital       f_flex
##     -27.592        3.946       19.887
```

```
#direction = c("both", "backward", "forward")
```

In our case, the procedure starts with the complete model, which includes the variables *profit*, *capital*, and *f_flex*. The corresponding AIC index value is 142.64. The first step shows that eliminating the *profit* variable leads to an improvement in the AIC value, reducing it to 140.65. Note that eliminating a variable always results in a decrease in the amount of explained deviance and an increase in the amount of residual deviance. In this case, eliminating the *profit* variable results in a reduction of 0.55 in explained deviance, which is relatively low compared to the total deviance. Eliminating the other two variables does not result in any improvement. At this point, the new reference model includes only the *capital* and *f_flex* covariates. The proce- dure then attempts to further eliminate each of these two remaining variables, but none of these further eliminations results in an improvement in the AIC index. Therefore, this model is the optimal one.

Predictions

Prediction of future observation

predict(mod, new_x, interval = "prediction", level = 0.95)

```
new_x <- c(1, 3.9, 6.02, 1.43)
names(new_x) <- names(coefficients(mod.C))
new_x <- data.frame(t(new_x))
```

```
predict(mod.C, new_x, interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 24.07958 -1.667204 49.82637
```

The point prediction is the same with both specifications: for a new observation with specified covariate values, a rating value of about 24 is predicted; we observe that this result is approximately coincident with the mean of the variable.

Regarding the interval estimation, the prediction interval is extremely wide, with a lower bound of -1.7 and an upper bound of 49.8. We can state, with a 95% confidence level, that the value of the response for the new observation falls within these values. Note that this interval goes beyond the range of variation of the response variable.

Prediction of mean response

```
predict(mod, new_x, interval = "confidence", level = 0.95)
```

```
predict(mod.C, new_x, interval = "confidence", level = 0.95)
```

```
##        fit      lwr     upr
## 1 24.07958 8.938264 39.2209
```

The confidence interval is much narrower (it does not take into account the variability of the new observation). In this case, the mean of the response variable (including the new observation) falls between 8.9 and 32.2 with a 95% confidence level.

MLR model whith categorical covariate

Dummy covariate

```
library(faraway)
data("sexab", package='faraway')
```

Change the reference class

```
relevel()
```

```
factor(sexab$csa[1:10])
```

```
##  [1] Abused Abused Abused Abused Abused Abused Abused Abused Abused Abused
## Levels: Abused
```

```
as.integer(sexab$csa[1:10])
```

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

We want our analysis to be centered on the category “NotAbused”, since the reference category as we can see is “Abused” we want to change it.

```
sexab$csa <- relevel(sexab$csa, ref="NotAbused")
as.integer(sexab$csa[1:10])
```

```
##  [1] 2 2 2 2 2 2 2 2 2 2
```

As we can se now the “Abused” observation are printed as category 2, so the reference category is “NotAbused”.

```
table()
table(sexab$csa)
```

```
##
## NotAbused    Abused
##        31        45
```

The two class are not balanced, this could bring some problem when training a model on this data, but since this data is used only for explanatory purpuses it work just fine.

Summary subsetted by the categorical variable

`by()`

```
by(sexab, sexab$csa, summary)
```

```
## sexab$csa: NotAbused
##       cpa               ptsd               csa
##  Min.   :-3.1204   Min.   :-3.349   NotAbused:31
##  1st Qu.:-0.2299   1st Qu.: 3.544   Abused   : 0
##  Median : 1.3216   Median : 5.794
##  Mean   : 1.3088   Mean   : 4.696
##  3rd Qu.: 2.8309   3rd Qu.: 6.838
##  Max.   : 5.0497   Max.   :10.914
## ------------------------------------------------------------
## sexab$csa: Abused
##       cpa              ptsd               csa
##  Min.   :-1.115   Min.   : 5.985   NotAbused: 0
##  1st Qu.: 1.415   1st Qu.: 9.374   Abused   :45
##  Median : 2.627   Median :11.313
##  Mean   : 3.075   Mean   :11.941
##  3rd Qu.: 4.317   3rd Qu.:14.901
##  Max.   : 8.647   Max.   :18.993
```

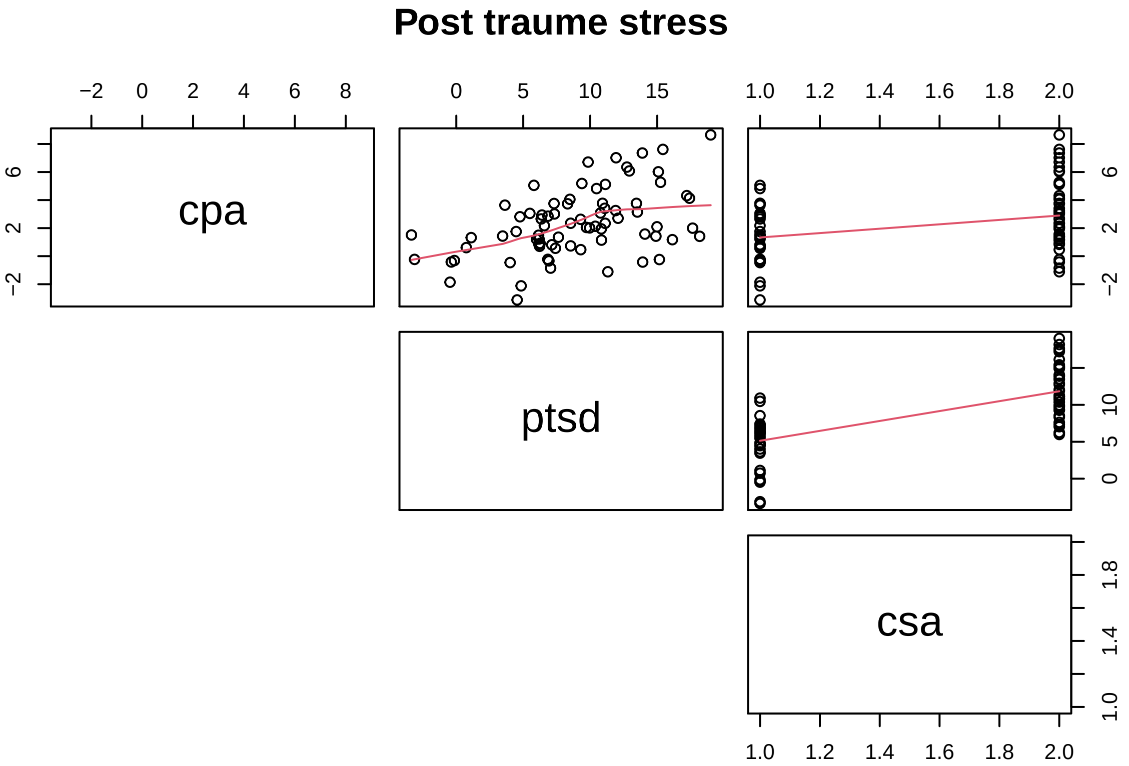
The range of a variable refers to the difference between the minimum and maximum values observed in the data. In this case, for the *cpa* variable, the range is from -3.1204 to 5.0497, while for the *ptsd* variable, the range is from -3.349 to 10.914. It's worth noting that these ranges are different in magnitude and direction, indicating that the variables have different distributions. The median of a variable is the middle value when all of the values are arranged in order. In this output, the median for *cpa* is 1.3216, and the median for *ptsd* is 5.794. This means that half of the observations for each variable fall above the median, and half fall below. The mean of a variable is the arithmetic average of all the values. In this output, the mean for *cpa* is 1.3088, and the mean for *ptsd* is 4.696. Compared to the medians, the means are slightly lower for both variables, which could indicate that there are some extreme values pulling the means downward.

Finally, the output presents a comparison between two classes: "Abused" and "NotAbused", as indicated by the *csa* variable. The summary statistics for the two groups reveal that, on average, those who have experienced childhood sexual abuse tend to have higher scores for both *cpa* and *ptsd* compared to those who have not experienced childhood sexual abuse. For example, the mean score for *cpa* is 3.075 for the "Abused" group, compared to a mean score of -1.115 for the "NotAbused" group. Similarly, the mean score for *ptsd* is 11.941 for the "Abused" group, compared to a mean score of 5.985 for the "NotAbused" group. These differences suggest that childhood sexual abuse may be associated with higher levels of both childhood physical abuse and post-traumatic stress disorder.

Plots

`pairs()`

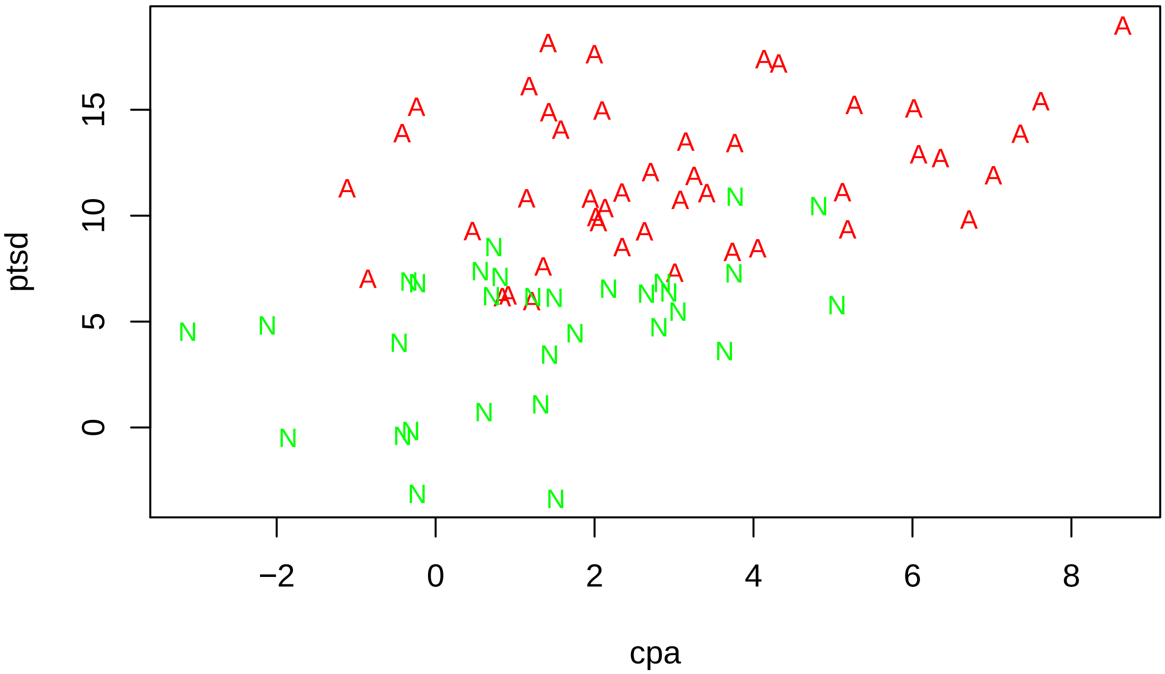
```
pairs(sexab, panel = panel.smooth,
      main = "Post traume stress",
      lower.panel = NULL)
```



By examining this graph, it is evident that there exists a positive correlation between the variables of *cpa* and PTSD. Moreover, the *csa* variable seems to be correlated with both of the other two variables. Notably, we can observe a discernible difference between the values of *cpa* for the two classes, with the “Abused” class exhibiting a higher value of *cpa* This distinction becomes even more pronounced when considering the variable of *ptsd*.

plot()

```
col <- c("green", "red")

plot(ptsd ~ cpa,
     pch = as.character (csa),
     sexab,
     col = col[sexab$csa],
     cex = 0.8)
```



This scatterplot with respect to the physical abuse score is described by units with respect to the third variable ("Abused" or "NotAbused") drqwn with a different symbol and colors. It is evident form the figure that the units indicated with A are almos always above those inidcated with N, i.e. with the same score for physical abuse, the stress reported by women who were also sexyally abused at school age is higher. Furthermore, the highest *cpa* value are observed for those who are abused.

ANCOVA model

`summary()`

```
lm1 <- lm(ptsd ~ cpa + csa, sexab)
summary(lm1)
```

```
##
## Call:
## lm(formula = ptsd ~ cpa + csa, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1567 -2.3643 -0.1533  2.1466  7.1417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9753     0.6293   6.317 1.87e-08 ***
## cpa           0.5506     0.1716   3.209  0.00198 **
## csaAbused     6.2728     0.8219   7.632 6.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.273 on 73 degrees of freedom
## Multiple R-squared:  0.5786, Adjusted R-squared:  0.5671
## F-statistic: 50.12 on 2 and 73 DF,  p-value: 2.002e-14
```

- Descriptive statistics about the regression residuals: we can see that the range of the residual is not very wide, although is over the range of +-3 o the normal residuals. The median is almost 0 (as expected based on theoretical assumptions), The quartiles are equidistant from the center of the distribution, which may suggest symmetry. So looking at residual distribution we can assume that they follow a Gaussian distribution.

- Estimates of the regression model parameters: the estimated values of the intercept and each of the covariates included in the model are reported, along with their standard errors and the results of the T test for each regression parameter. In this case all the estimates are considered important for the model.

    - The estimated coefficient $\widehat{\beta}_0$ for the intercept represents the expected value of the 0 response variable when all explanatory variables have a value of zero, considering the "NotAbused" category.

    - The estimated coefficient $\widehat{\beta}_1$ for *cpa* of education represents the 1 expected increase of the response variable for a unit increase in the *cpa* variable.

    - We can see form the estimate of *csaAbused* that if the value of *csa* is not "Abused" we have an increment of 6.27 in the explanatory variables keeping fixed the other covariate.

It can be seen that the association between *cpa* test score with the *ptsd* score is positive net of *csa*: the test score increase by 0.55 points for each unit increas in physical abuse whether the female has (or has not) been sexually abused.

The slope of both lines si 0.551, but the "Abused" line is 6.273 higher than the "NotAbused". The regression line for "NotAbused" is the following

$$\widehat{Y}_N = 3.97 + 0.55 * cpa$$

And those for "Abused" is

$$\widehat{Y}_N = (3.97 + 6.27) + 0.55 * cpa$$

$$\widehat{Y}_N = 10.24 + 0.55 * cpa$$

- The residual standard error (RSE), which is approximately 3.3, along with its degrees of freedom: 73, i.e., $n - p - 1 = 76 - 2 - 1 = 73$, where p represents the number of covariates and one is subtracted if the model contains an intercept. This is the square root of the ratio of the sum of squares of residuals and the number of degrees of freedom. The RSE can be interpreted as the average deviation around the mean of residuals, which is assumed to be zero, and thus as the average deviation between observed and corresponding interpolated values. In other words, this value states that, on average, interpolated values deviate from observed ones by 3.3.

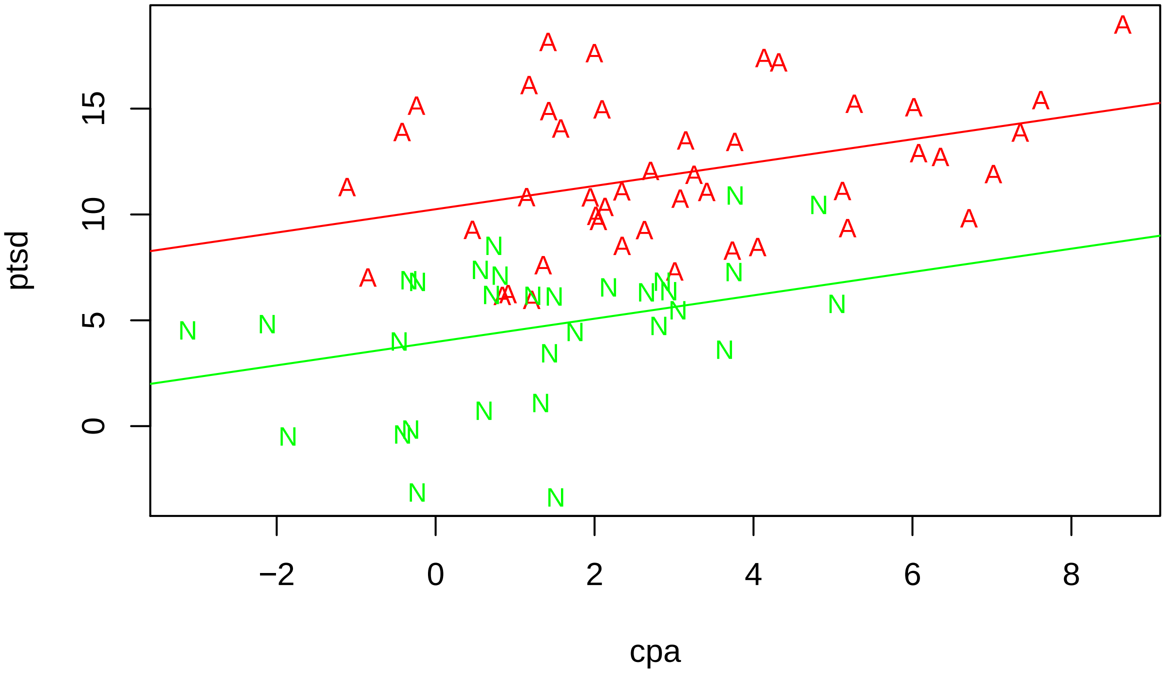
- The multiple R-squared and its adjusted value. The multiple linear determination coefficient represents the ratio of the variance explained by the interpolating plane and the total variance of the response variable. In this case, a value of 0.56 is observed, indicating that the interpolating plane explains approximately 56% of the variability of the initial salary. This value is not considered sufficent high, may be other specification of the model are needed.

- The results of the F-test, i.e., the test statistic value, which is 50.12, and the corresponding p-value (of the order of $10e^{-14}$). The proximity to 0 of the p-value allows us to reject the null hypothesis that all regression coefficients, except for the intercept, are equal to 0.

From this output we can add two parallel lines to the plot:

plot()

```
col <- c("green", "red")

plot(ptsd ~ cpa,
     pch = as.character (csa),
     sexab,
     col = col[sexab$csa],
     cex = 0.8)
abline(10.248, 0.551, col="red")
abline(10.248-6.273, 0.551, col="green")
```



It can be seen that the intercepts are different but the slopes of the two straight lines are the same. The vertical distance between the two lines is the estimated coefficent. Since the lines are parallel, the value 6.27 is the difference between the main effects of the dichotomus variable and represents the vertical distance between the two regression lines form each fixed value of the continuous variable. The value of 0.55 is the main effect of the continuous covariate *cpa* whatever the level of the dichotomous variable. It can be seen that the association between *cpa* test score with the *ptsd* score is positive net of *csa*: the test score increase by 0.55 points for each unit increas in physical abuse whether the female has (or has not) been sexually abused. So the expected value of the response when there is no physical abuse is much higher for those who have benn sexually abused.

Interaction terms

lm()

```
lm2 <- lm(ptsd ~ cpa*csa, sexab)
summary(lm2)
```

```
##
## Call:
## lm(formula = ptsd ~ cpa * csa, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1999 -2.5313 -0.1807  2.7744  6.9748
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.6959     0.7107   5.201 1.79e-06 ***
## cpa              0.7640     0.3038   2.515   0.0142 *
## csaAbused        6.8612     1.0747   6.384 1.48e-08 ***
## cpa:csaAbused   -0.3140     0.3685  -0.852   0.3970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.279 on 72 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5654
## F-statistic: 33.53 on 3 and 72 DF,  p-value: 1.133e-13
```

Note that the estimated regression coefficient is negative and it should be interpreted by comparing slopes under abused and not abused conditions. For $\widehat{\beta}_3 = -0.314$, we say that the expected score on post-traumatic stress disorders 3 under physically and sexually abused conditions is $0.45 = (0.764 - 0.314)$ compared to 0.764 under only physically abused but not sexually abused conditions. According to the t-test we have to reject the null hypothesis that the parameter referred to the interaction term is equal to zero at each significance level. However, for model selection it is better to use AIC index.

Categorical covariate with more than two levels

Summary subsetted by the categorical variable

summary()

```
mod <- lm(data = diamonds, formula = carat ~ .-width)
summary(mod)
```

```
##
## Call:
## lm(formula = carat ~ . - width, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24794 -0.04288 -0.01309  0.03910  0.53387
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.146856   0.035857 -31.984  < 2e-16 ***
## cut2        -0.035797   0.017596  -2.034 0.042249 *
## cut3        -0.046812   0.016872  -2.775 0.005658 **
## cut4        -0.037563   0.016923  -2.220 0.026725 *
## cut5        -0.057626   0.016433  -3.507 0.000479 ***
## price        0.029657   0.001618  18.328  < 2e-16 ***
## length       0.159147   0.021192   7.510 1.60e-13 ***
## depth        0.270905   0.034421   7.870 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08218 on 792 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9707
## F-statistic:  3785 on 7 and 792 DF,  p-value: < 2.2e-16
```

The descriptive statistics for the residuals show very small values, being minimum and maximum values equal to -0.25 and 0.53, respectively. The median is very close to zero, and the overall distribution seems to be quite symmetrical around it. In genaral, even though further (graphical) analysis are required, the theoretical assumprions seem to be fullfilled.

The estimated value for the intercept term is equal to $\hat{\beta}_0 = -1.15$: according to the model 0 the carats of a diamonds with dimensionse equal to 0 mm, with a price of 0$ and with cut of the worst quality (type 1) has -1.15 carats (this interpretation has obviously no meaning in the practical context). The corresponding standard error, equal to $\hat{SE}_{\beta_0} = 0.036$, is used to asses the significativity of the estimated parameter through the T test. The test statistics is equal to $\frac{\hat{\beta}_0 - 0}{\hat{SE}_{\beta_0}} = -31.98$ and the p-value is extremly small ($\approx 10^{-16}$). We can reject the null hypothesis $h_0 : \beta_0 = 0$, concluding that the intercept term provide a significant contribution in explaining the response variable.

The same conclusion holds for all the continuous explanatory variables: the very small values of the p-values lead us to reject the null hypothesis that $\beta_j = 0$ ($j = 5, 6, 7$): all three variables have a significant effect in the estimation of the diamonds' carats. Regarding the interpretation of the corresponding estimated coefficients, we can for example state that the increase in the value of carats, corresponding to a unitary increase in the price (+1000$) is equal to 0.030 (if the other variables are kept fixed). Simalrly, an increase of 1 mm in length or depth leads to a growth in the carats equal to 0.159 and 0.271, respectively.

Finally, considering the categorical variable cut and considering the worst-quality level (1) as a baseline, we observe that each one of the other quality levels (from 2 to 5) has a negative effect on the carats. For example, a diamond with the best-quality level of cut has on average 0.058 carats less than a diamonds with quality cut of the first category (worst quality). Similarly for quality levels 2, 3, and 4. All these effects are significant (even though with different confidence levels), as shown by the p-values of the associated significative T test.

The residual standard error (RSE) is equal to 0.082; it means that, on average, the observed and the estimated values for each sample unit differ by 0.082. Although the value seems small, it needs to be further evaluated with respect to a centrality index (e.g., the mean) of the response variable. Computing the ratio between the RSE and the mean of the carats, we obtain the percentage error, equal to $0.082/0.84 = 0.098$: on average the percenatege error between observed and fitted values is equal to 9.8%.

The multiple R^2 shows an optimal fit of the model to the data; being equal to 0.971 (with an adjusted value approximately equal), we can conclude that aroung 97% of the total variability of the response variable can be explained by the estimated model.

Finally, let us consider the F test having the joint non-significance of all regression coef- ficients as null hypothesis (the alternative hypothesis considers instead that at least one coefficient is statistically significant). The value of the F statistics is equal to 3785; the associate p-value is very close to 0 ($< 10^{-16}$), leading us to reject the null hypothesis at each confidence level.

Plot

The estimated model cannot be represented into a 2D scatterplot, due to the excessive number of explanatory variables. Therefore, we only report here the equations of the model. Denoting by Y the response variable carat and by $X1$, $X2$, and $X3$ the continuous explanatory variables price, length, and depth, respectively, we obtain the following equations:

$$\hat{Y} = -1.147 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3, \text{ if cut quality is 1}$$

$$\hat{Y} = -1.147 - 0.036 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3, \text{ if cut quality is 2}$$

$$\hat{Y} = -1.147 - 0.047 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3, \text{ if cut quality is 3}$$

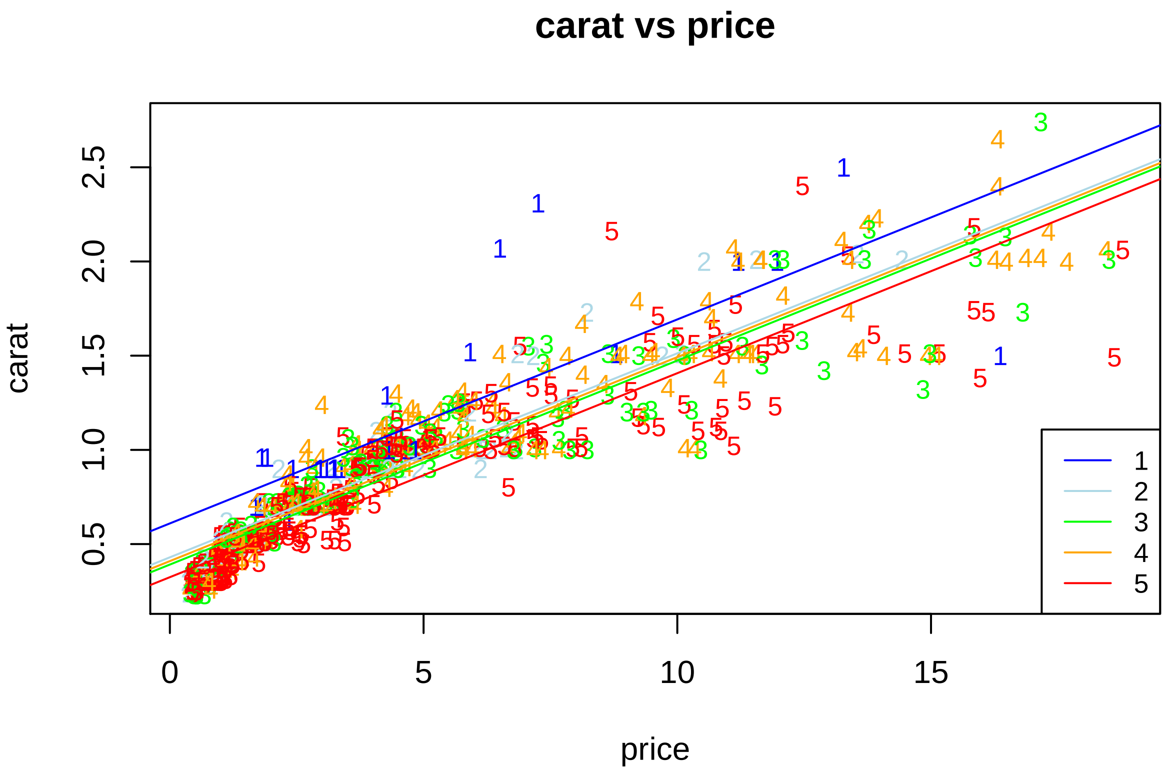
$$\hat{Y} = -1.147 - 0.038 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3, \text{ if cut quality is 4}$$

$$\hat{Y} = -1.147 - 0.058 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3, \text{ if cut quality is 5}$$

Consideting only 2 covariate:

```
plot()
abline()
```

```
lma <- lm(carat ~ price + cut, data = diamonds)

col <- c("blue", "light blue", "green", "orange", "red")
names(col) <- levels(diamonds$cut)

plot (carat ~ price, pch=as.character (cut), diamonds, col = col[diamonds$cut], cex = 0.8)

abline (lma$coefficients[1], lma$coefficients[2], col = "blue")
abline (lma$coefficients[1] + lma$coefficients[3], lma$coefficients[2], col = "light blue")
abline (lma$coefficients[1] + lma$coefficients[4], lma$coefficients[2], col = "green")
abline (lma$coefficients[1] + lma$coefficients[5], lma$coefficients[2], col = "orange")
abline (lma$coefficients[1] + lma$coefficients[6], lma$coefficients[2], col = "red")
legend("bottomright", 2, c(1,2,3,4,5), col = c("blue", "light blue", "green", "orange", "red"), lty= 1,
title("carat vs price")
```



Upon analyzing these scatter plots, we can observe some interesting patterns. Firstly, it is apparent that the straight lines for each plot have different intercepts, but they share the same slope. The vertical distance between the two lines represents the estimated coefficient. Moving onto the regression line for the cut variable, we can see that the intercept values vary across the different cuts. Specifically, the expected value for the carat variable is higher for cut type 1 compared to cuts 2, 3, and 4, while cut 5 has the lowest intercept value.

Logistic regression model for binary response

Relevel of the respone

as.factor()

```
r <- as.factor(0 + (urine$r == "1"))
urine$r <-  r
```

In this case we had the variable *r* which even if it is a binary variable has not been considered as a factor, to proceed with further analysis we must transform it into a factor and change it in the dataset.

Another example:

```
table(Womenlf$partic)
```

```
##
## fulltime not.work parttime
##       66      155       42
```

```
Y <- as.factor(0 + (Womenlf$partic == "not.work"))
table(Y)
```

```
## Y
##   0   1
## 108 155
```

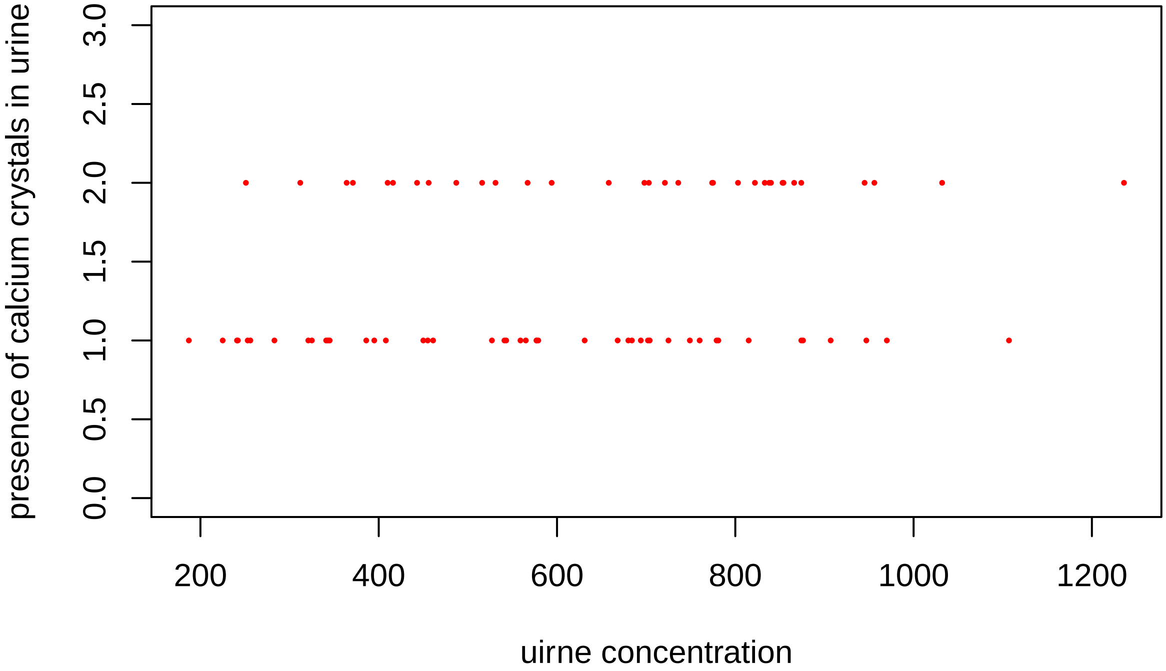
The variable *partic* had three category (“fulltime”, “not.work”,“fulltime”) but we are only interested if a woman work or not, so we collpase the “fulltime” and “fulltime” category in to a single one. The new varible *Y* get value 0 if the woman work and 1 else.

Explanatory analyses

Plots

plot()

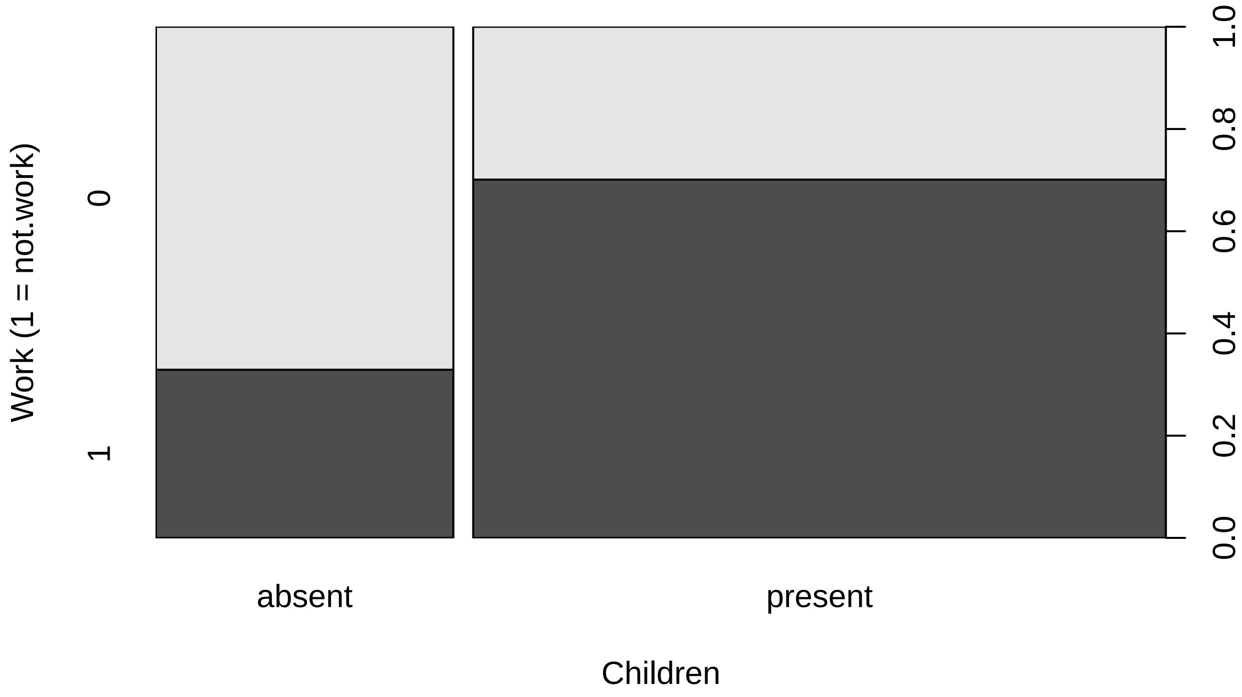
```
plot(urine$osmo, urine$r,
     main = " ",
     xlab = "uirne concentration",
     ylab = "presence of calcium crystals in urine",
     pch = 16, cex = 0.4, col = "red", ylim = c(0,3))
```



We plotted the categorical variable indicating the presence of calcium in the urine on the y-axis (1 for the absence of “0” and 2 for the presence of “1”) and the urine concentration on the x-axis. No particular differences are visible. The only visible thing is that it seems that for class “1” the concentration in the urn is less present below the value 400 than for class “0”.

`plot()`

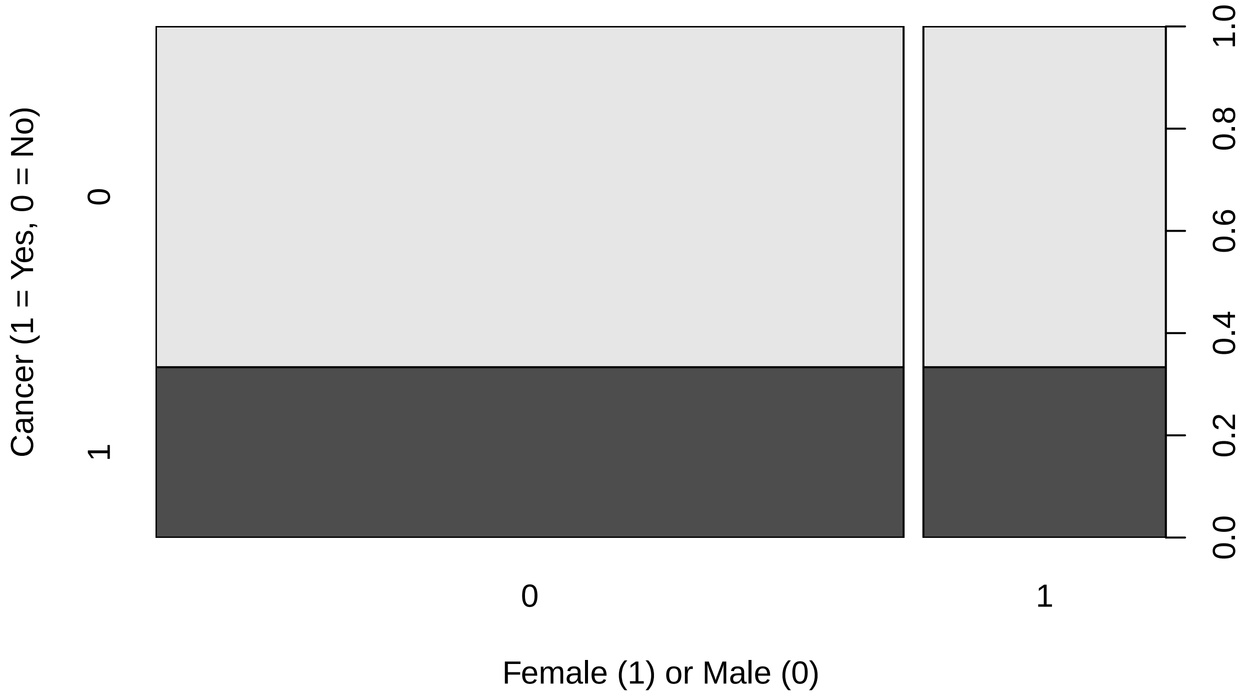
```
plot(Womenlf$children, Y,
     main = " ",
     xlab = "Children", ylab = "Work (1 = not.work)")
```



This graph is referred to the response variable with respect to the presence or absence of children in the family. We can see that the presence of children make the proportion of those who do not work much higher that the observed when there are no children, the proportion is 0.7 vs 0.33 respectively.

Another example:

```
plot(birds$female, birds$cancer,
     xlab = "Female (1) or Male (0)",
     ylab = "Cancer (1 = Yes, 0 = No)")
```



From the length of the base of the rectangles, we observe that the number of females is much lower than that of males. Additionally, we notice that the proportion of individuals with cancer is the same between men and women, slightly higher than 30%.

Tables

`table()`

```
table(Womenlf$children, Y)
```

```
##          Y
##              0   1
##   absent    53  26
##   present   55 129
```

We can see that for the class of workers the presence and absence of children si quite balanced with 55 and 53 respectively, but looking at the class of non workers it is visible how the observations with the presence of children are the big majority.

`prop.table(table)`

```
round(prop.table(table(Womenlf$children, Y)), 1)
```

```
##          Y
##              0   1
##   absent   0.2 0.1
##   present 0.2 0.5
```

By looking at the table of propotrion this difference is even more visible. We can see that the women with the presence of children usually do not work.

Odds and logit

```
Odds_CancerMen <- 37/74; Odds_CancerMen
```

```
## [1] 0.5
```

```
Odds_CancerWomen <- 12/24; Odds_CancerWomen
```

```
## [1] 0.5
```

```
Odds_ratio <- (37/74)/(12/24); Odds_ratio
```

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

As expected, the odds of having cancer is exactly the same for men and women in the sample, resulting in a value of the odds ratio equal to 1. The binary variable *female* does not seem to be associated to the response variable. The same procedure is now repeated for the binary explanatory variable bird, measuring the presence or absence of caged birds.

```
table(Birds = birds$bird, Cancer = birds$cancer)
```

```
##      Cancer
## Birds  0  1
##     0 64 16
##     1 34 33
```

```
round(prop.table(table(Birds = birds$bird, Cancer = birds$cancer)), 4)
```

```
##      Cancer
## Birds      0      1
##     0 0.4354 0.1088
##     1 0.2313 0.2245
```

The number of individuals having caged birds and also having cancer is equal to 33, corre- sponding to around 23% of the total. Only 16 subjects (about 11% of the total) have cancer without having caged birds.

```
round(prop.table(table(Birds = birds$bird, Cancer = birds$cancer), margin = 1), 4)
```

```
##      Cancer
## Birds      0      1
##     0 0.8000 0.2000
##     1 0.5075 0.4925
```

Considering the row-conditioned relative frequencies, we notice that among the individuals who do not have caged birds (first row) “only” 20% has lung cancer. On the contrary, this proportion noticeably increases considering subjects who have caged birds; in this case the proportion of having lung cancer is almost equal to 50%.

To better quantify the effect of having or not caged birds, we compute the odds and odds ratio.

```
Odds_CancerBirds <- 33/34; round(Odds_CancerBirds, 4)
```

```
## [1] 0.9706
```

```
Odds_CancerNoBirds <- 16/64; round(Odds_CancerNoBirds, 4)
```

```
## [1] 0.25
```

```
Odds_Ratio <- (33/34)/(16/64); round(Odds_Ratio, 4)
```

```
## [1] 3.8824
```

Odds of having cancer is almost 4 times greater for individuals having caged birds. From this preliminary description of the sample data, therefore, the presence or absence of caged birds is associated to the incidence of cancer.

Estimation of the model

`glm()`

```
glm1 <- glm(formula = cancer ~ yrsmoke + bird, family = binomial, data = birds)
summary(glm1)
```

```
##
## Call:
## glm(formula = cancer ~ yrsmoke + bird, family = binomial, data = birds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.18016    0.63640  -4.997 5.82e-07 ***
## yrsmoke      0.05825    0.01685   3.458 0.000544 ***
## bird1        1.47555    0.39588   3.727 0.000194 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 158.11  on 144  degrees of freedom
## AIC: 164.11
##
## Number of Fisher Scoring iterations: 4
```

The range of variation of the residuals is quite narrow, going from a minimum equal to -1.6 to a maximum of 2.1. Non of the residuals assumes particularly high values; the median is quite close to zero (-0.5) and the overall distribution seems to be approximately symmetric.

The Residual deviance value, equal to 158.11 represents the amount of variability of the response variable that the considered model is not able to explain. The Null deviance is instead the amount of variability of the response variable that the null model is not able to explain. Therefore the selected model explains an additional amount of variability of about 29.03.

As already mentioned, the AIC index is equal to 164.11. The number of Fisher scoring iterations (equal to 4) is the number of steps that the estimation iterative algorithm requires to converge to a maximum of the log-likelihood function.

Equation of the model: Based on the estimated coefficients, we can write the analytical expression of the logit of $\hat{p}_i$, where $p_i = P(Y_i = 1)$ is the estimated probabilit of having cancer. We have the following expression:

$$logit(\hat{p}_i) = -3.18 + 0.06 \cdot yrsmoke + 1.48 \cdot bird : 1$$

or, after applying the exponential function:

$$\frac{P(Y_i = 1)}{P(Y_i = 0)} = e^{-3.18} \cdot e^{0.06 \cdot yrsmoke} \cdot e^{1.48 \cdot bird : 1}$$

Parameter interpretation: Firstly, we observe that both estimated coefficients (except the intercept) are positive; therefore, the following general comments hold:

- the probability of having lung cancer increases with the number of past years of smoking;
- the probability of having lung cancer increases for a subject who has caged birds, com- pared to one who does not have them.

Going into more detail, and based on the equations of the estimated model we have written above, it is appropriate to calculate the exponential of the estimated coefficients.

```
exp(coef(glm1))
```

```
## (Intercept)     yrsmoke       bird1
##  0.04157919  1.05997966  4.37344710
```

- The estimated parameter for the intercept term is equal to $\hat{\beta}_0 = -3.18$, with $exp(\hat{\beta}_0) = 0.042$. This vlaue represents the odds of having cancer (i.e., the ratio between the probability of having cancer and the probability of not having cancer), when both explanatory variables are equal to 0:

  $$\frac{P(Y_i = 1)}{P(Y_i = 0)} = e^{-3.18} = 0.042 \Rightarrow P(Y_i = 1) = 0.042 \cdot P(Y_i = 0)$$

  Therefore, cosnidering a subject who has never smoked and without caged birds, the probability of having cancer is about 0.042 times the probability of not having cancer (around 24 times smaller).
- The estimated parameter for the *yrsmoke* covariate is equal to $\hat{\beta}_1 = 0.058$, with $exp(\hat{\beta}_1) = 1.060$. This value represents the (multiplicative) effect on the odds of having 1 cancer, due to a unitary increase

in the past smoking years, when the other explanatory variable is held fixed. In other words, as the number of past smoking years increases by one year, the odds of having cancer is multuplied by 1.060, thus slightly increasing.

- The estimated parameter for the *bird* variable is equal to $\hat{\beta}_2 = 1.476$, with $exp(\hat{\beta}_2) = 4.374$. It has a similar interpretation: this value represents the (multiplicative) effect of the presence of caged birds on the odds of having cancer (keeping the other covariate fixed). In other words, the odds of having cancer for an individual with caged birds is 4.4 times greater with respect to an individual who has not caged birds.

All the estimated coefficient are statistically significant according to significance test. For each coefficient, we have enough evidence to reject the null hypothesis $H_0 : \beta_j = 0$ at each significance level.

Alternative:

```r
mod1 <- glm(Y ~ Womenlf$hincome + Womenlf$children,
            family = binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = Y ~ Womenlf$hincome + Womenlf$children, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9970  -0.9292   0.7768   0.8652   1.6767
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.33583    0.38376  -3.481   0.0005 ***
## Womenlf$hincome          0.04231    0.01978   2.139   0.0324 *
## Womenlf$childrenpresent  1.57565    0.29226   5.391    7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 356.15  on 262  degrees of freedom
## Residual deviance: 319.73  on 260  degrees of freedom
## AIC: 325.73
##
## Number of Fisher Scoring iterations: 4
```

From the model we have

$$logit(\hat{p}_i) = -1.3358 + 0.0423(inc) + 1.5756(chil)$$

where $\hat{p}_i$ is the probability of not working.

Please note that the logistic model must be interpreted as a multiplicative model for the $p/(1-p)$ (odds). Therefore $exp(\hat{\beta}_1) = exp(0.0423) = 1.0432$ is the ratio of the odds of group A (not 1 working) vs group (B working) for a unit increase of husband income held fixed the presence or absence of children. It can also be interpreted as follows: for each additional 1000$s of the husband income the females unemployment increases by 4%. We also see that for a fixed husband income the odds of working for a female the odds of not working increases by $exp(\hat{\beta}_2) = exp(1.5756) = 4.8336$. The probability of not working for mothers is about five times that of childless women with the same household income. Thus, family income is also a risk factor for non-work, although much less so than the presence of children.

- Residuals are not very high.

- The sum of residuals is the residual deviance of the estimated which is equal to 319.73. A model with small residual deviance is preferred. Therefore it is compare with the deviance of the model which does not include covariate null model.
- Residual deviance is compare with the null deviance to assess the model since
- Wald test for each regression coefficient produces a highly significant p-value for the coefficient referred to children: the realized value of the test statistic is

  $$Z = \frac{1.5756}{0.2922} = 5.391$$

  indicating that the null hypothesis $H_0 : \beta_2 = 0$ can be rejected and data provide a significant evidence that support for the not work increase with children in the family after adjusting for the husband income.
- Wald test for the regression coefficient related to the husband income lead us to reject the null hypothesis at the significant level of 95%. There is moder- ate evidence that support not work increase with higher husband incomes after adjusting for the presence of children.

The deviance of the null model is greater than the residual deviance so the covariates are important in explaining the probability of not working.

The Akaike index of the full model is 325.7325.

The logit model can be written in probability form

$$\hat{p}_2 = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 * income + \hat{\beta}_2 * chil)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 * income + \hat{\beta}_2 * chil)}$$

Fitted values (Estimated probabilities)

mod$fitted.values

```
prob <- glm1$fitted.values
round(head(prob), 3)
```

```
##     1     2     3     4     5     6
## 0.355 0.396 0.112 0.424 0.525 0.144
```

For instance, looking at the first six values, we observe that subject 5 has probability of having cancer equal to 0.53. This is a 49-year-old male individual, who has been smoking for 31 years and currently smokes 20 cigarettes per day; he has caged birds. On the contrary, subject 3 has a much lower probability of having cancer. This is a 43-year-old male individual, smoking by 19 years (15 cigarettes per day) without caged birds.

summary()

```
summary(prob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03992 0.15386 0.31176 0.33333 0.48161 0.76992
```

According to the model, the probabilities of having cancer range from 0.04 to 0.77. On average, the probability is equal to 0.33, and 50% of individuals have a probability inferior to 0.31.

which()

```
birds[which.min(prob), ]
```

```
##    female age highstatus yrsmoke cigsday bird cancer
## 88      0  60          1       0       0    0      0
```

```
birds[which.max(prob), ]
```

```
##    female age highstatus yrsmoke cigsday bird cancer
## 36      0  66          0      50      25    1      1
```

The subject with the lowest estimated probability to have cancer is a 60-year-old male individual who has never smoked and who has not caged birds; actually he has not cancer. On the contrary, the individual with the highest estimated probability is a 66-year-old male subject smoking by 50 years and currently smoking 25 cigarettes per day. He has caged birds and actually he has lung cancer.

Confidence interval

`confit(mod)`

```
confint(mod, parm = 3) #only the third estimated coefficient
```

```
##              2.5 %      97.5 %
## cut3 -0.07993053 -0.01369279
```

We observe that the confidence interval does not include 0, highlighting that the correspond- ing variable. We can also compute the exponential of the estimated confidence interval.

Confidence interval for the estimated odds ratio:

`exp(confit(mod))`

```
exp(confint(glm1, parm = 3))
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 %
## 2.050761 9.748175
```

At confidence level equal to 0.95, individuals with caged birds have a multiplicative increase in the odds of having cancer that goes from 2.05 to 9.75. This interval does not contain 1, so the presence of caged birds has a significant effect in estimating the probability to have cancer.

If 1 is included suggests that the odds of the outcome variable being in the reference category is not significantly different from 1.

Multinomial logit model

Explanatory analyses

Table

`prop.table(table())`

```
prop.table(table(Status = hsb$ses, Program = hsb$prog), margin = 1)
```

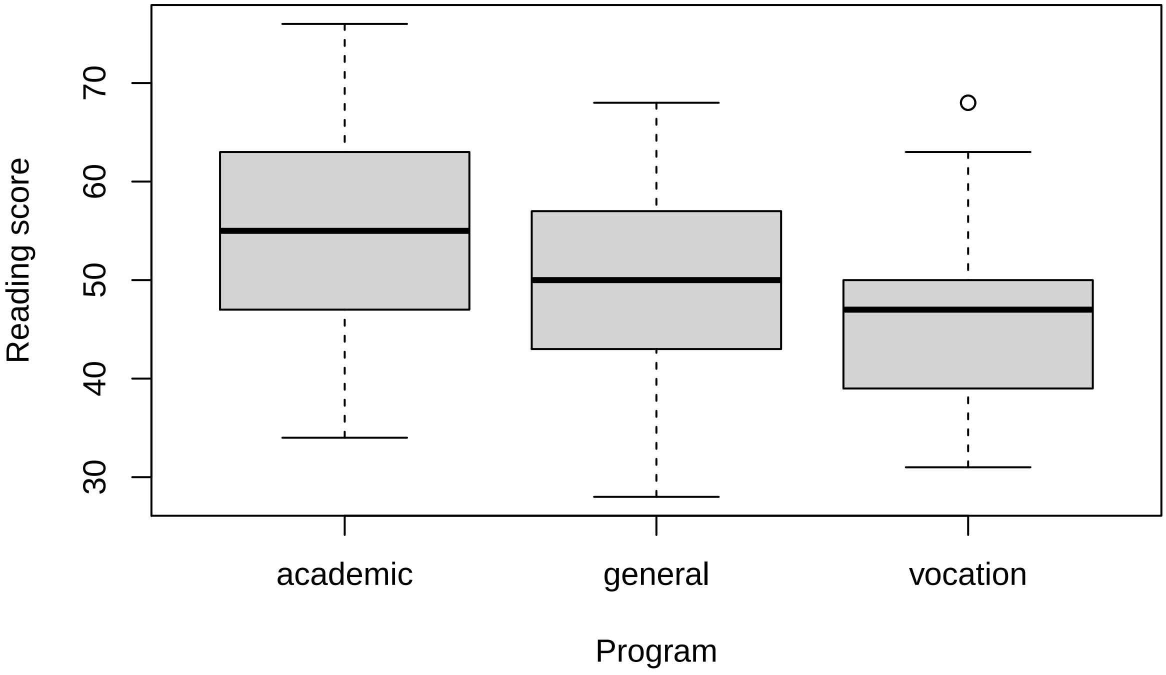
```
##         Program
## Status    academic   general  vocation
##   high   0.7241379 0.1551724 0.1206897
##   low    0.4042553 0.3404255 0.2553191
##   middle 0.4631579 0.2105263 0.3263158
```

We observe that students with a high socio-economic status tend to prefer an academic program (over 72%), with a minority of students selecting the general (15.5%) and vocation (12.1%) programs. On the other hand, considering students with a medium and low socio- economic status, the academic program remains the most popular, but there is no longer such a pronounced predominance. In particular, it is selected by 46.3% of students with a medium socio-economic status (the percentage of students selecting a vocation program increases significantly) and by 40.4% of students with a low socio-economic status (very similar to the proportion of students selecting the general program).

Plots

plot()

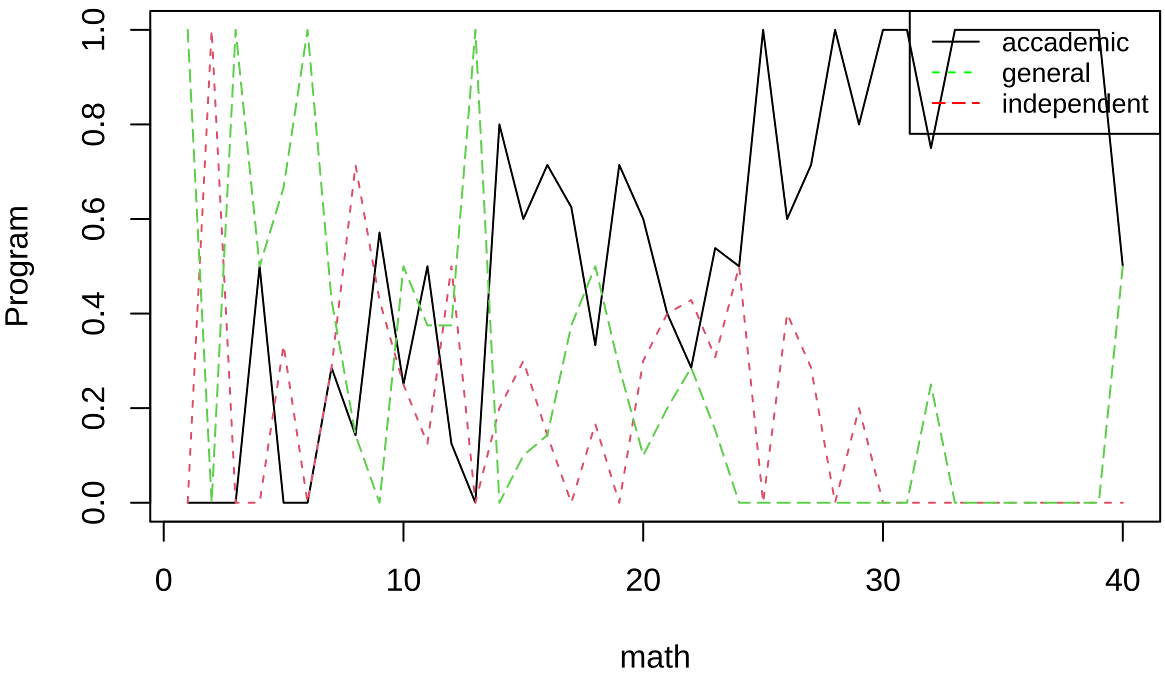
```
plot(hsb$prog, hsb$read,
     xlab = "Program", ylab = "Reading score")
```



The reading score appears to be strongly associated with the chosen study program. Stu- dents enrolled in the academic program have the highest reading level: values range from a minimum of 34 to a maximum of 76 (which represents the highest reading score for the entire dataset); furthermore, half of the students have a score higher than 55. The reading level is generally lower for those in the general program. From the median, we observe that half of these students have a score not exceeding 50; none of them reaches 70. In this class, we also find the student with the lowest level in the entire dataset, which is below 30. Finally, subjects belonging to the vocation program show an even lower reading level, with three-quarters of the students not reaching a score of 50.

plot(prop.table(table(),1))

```
matplot(prop.table(table(hsb$math,hsb$prog),1),
        type="l",
        xlab="math",
        ylab="Program", lty=c(1,2,5))
legend("topright", 2,
       c("accademic", "general", "independent"),
       col = c("black", "green", "red"),
       lty = c(1,2,5), cex = 0.8)
```



In the first part of the grades we can see that there is no difference between the three class, and frequencies alternate, there are little more green pick than other. In the second part expecially over 25 we can see how accademic program outclas the other two with resspect every grade.

In generale, specificare "margin = 1" o "margin = 2" può cambiare l'interpretazione del grafico risultante e delle informazioni che fornisce. In questo caso, essendo "math" l'argomento lungo le righe della tabella di contingenza, la specifica di "margin = 1" permette di vedere come le frequenze relative delle tre categorie di "prog" variano al variare dei valori di "math". Se avessimo specificato "margin = 2", avremmo visto invece come le frequenze relative dei diversi valori di "math" si distribuiscono tra le tre categorie di "prog".

Estimation of the model

multinom()

```
library(nnet)
mod.S <- multinom(formula = prog ~ ses + schtyp + math + science + socst,
                  data = hsb)
```

```
## # weights:  24 (14 variable)
## initial  value 219.722458
## iter  10 value 171.169761
## iter  20 value 157.775586
## final  value 157.775540
## converged
```

The output presents the initial value of the objective function (initial value) which indicates how far the model is from being optimal before starting the optimization process. Next, the value of the objective function after each iteration (iter value), i.e. the value of the objective function calculated during the optimization of the model in each iteration, is presented. The model has converged, as indicated by "converged" at the end of the output, which means that the optimization process has completed successfully and the model has reached a stable value of the objective function. The final value of the objective function (final value) indicates how close the model is to being optimal after optimization.

```
summary(mod.S)
```

```
## Call:
## multinom(formula = prog ~ ses + schtyp + math + science + socst,
##     data = hsb)
##
## Coefficients:
##          (Intercept)      seslow sesmiddle schtyppublic       math    science
## general     2.587029  0.87607389 0.6978995    0.6468812 -0.1212242 0.08209791
## vocation    6.687272 -0.01569301 1.2065000    1.9955504 -0.1369641 0.03941237
##                socst
## general  -0.04441228
## vocation -0.09363417
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic       math    science
## general     1.686492 0.5758781 0.4930330     0.545598 0.03213345 0.02787694
## vocation    1.945363 0.6690861 0.5571202     0.812881 0.03591701 0.02864929
##               socst
## general  0.02344856
## vocation 0.02586717
##
## Residual Deviance: 315.5511
## AIC: 343.5511
```

The coefficients indicate the effect of each predictor variable on the probability of students belonging to a particular program type, relative to the baseline. For example, a positive coefficient for seslow suggests that students with lower SES are more likely to be in the general or vocational program types than in the academic program type.

The standard errors reflect the precision of the coefficient estimates, and the lower they are, the more confident we can be in the results.

The residual deviance measures how well the model fits the data, with smaller values indicating a better fit. The AIC (Akaike information criterion) is a measure of the model's goodness of fit, taking into account both the model's complexity and how well it fits the data.

One category is always considered as baseline. We need to account for the fact that a positive regression coefficient β_j does not necessary imply an increasing probability for category j as x_i increases since it means that the odds for category j increases relative to the reference category.

Predictions

`predict(,type="probs")`

```
head(predict(mod.S, type = "probs"))
```

```
##    academic   general  vocation
## 1 0.3083104 0.4917649 0.1999247
## 2 0.3968192 0.3850708 0.2181100
## 3 0.2481586 0.2668721 0.4849693
## 4 0.5123112 0.2812957 0.2063931
## 5 0.6769583 0.1779816 0.1450601
## 6 0.3378965 0.4178550 0.2442486
```

Considering the predicted probabilities for the first six observations, we observe that none of them shows particularly close to one for any of the program levels. The most certain classification is for sample unit corresponding to the fifth row; it presents a probability equal to 0.68 for the academic program and lower probabilities for the remaining two levels. On the contrary, the second observation shows similar probabilities for both the academic and the general program.

predict()

```
xtabs( ~ predict(mod.S) + hsb$ses)
```

```
##                hsb$ses
## predict(mod.S) high low middle
##       academic   52  18     56
##       general     1  16      4
##       vocation    5  13     35
```

We observe that the majority (52) of subjects with a high socio-economic status are classi- fied in the academic program: only five of these observations are allocated in the vocation program, and one in the general program. Even among subjects with intermediate socio- economic status, the majority are classified in the academic program (56); in this case, however, there is also a high number of subjects allocated in the vocation program (35). Once again, the general program is the least frequent, with only 4 observations. Finally, subjects with low socio-economic status are fairly evenly classified among the three different program types.

(for prediction of probability or for prediction of a new unit see solution 26.6)

#Gaussian mixture model ## Model selection

mclustBIC()

```
require(mclust)
```

```
## Loading required package: mclust

## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.

##
## Attaching package: 'mclust'

## The following object is masked from 'package:bootstrap':
##
##     diabetes

## The following object is masked from 'package:mvtnorm':
##
##     dmvnorm

## The following object is masked from 'package:faraway':
##
##     diabetes
```

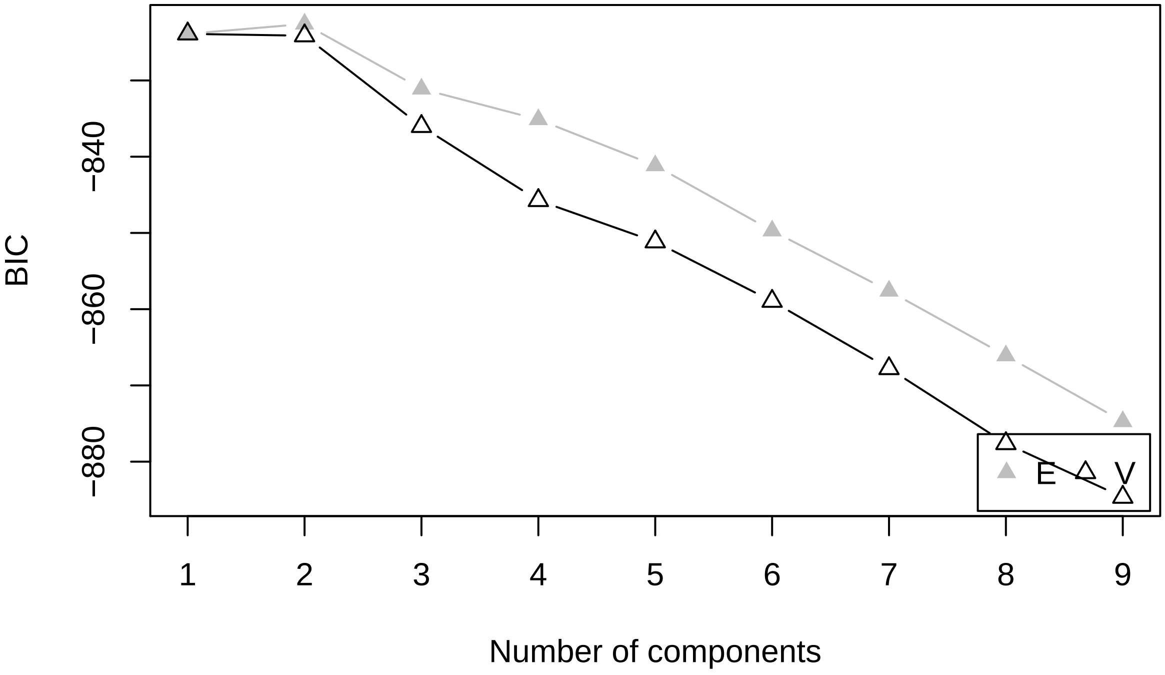
```
mod_sel <- mclustBIC(haemoglobin)
mod_sel
```

```
## Bayesian Information Criterion (BIC):
##           E         V
## 1 -823.8906 -823.8906
## 2 -822.5878 -824.1333
## 3 -831.1105 -836.0230
## 4 -835.1245 -845.7390
## 5 -841.1645 -851.1584
## 6 -849.7154 -858.9562
## 7 -857.6157 -867.7858
## 8 -866.1007 -877.6227
## 9 -874.7309 -884.6570
##
## Top 3 models based on the BIC criterion:
##       E,2       E,1       V,1
## -822.5878 -823.8906 -823.8906
```

The output of the used function shows a matrix with the BIC values for all the estimated models; the rows correspond to the number of components used, while the columns corre- spond to the two variance specifications. In addition, the output presents the list of the top three models in terms of BIC; in this case, the optimal choice is to select 2 groups and assume that the variance is common among all the components of the model. The second and third best models assume only one component; in this case, the model assumes no heterogeneity among the observations.

plot()

```
plot(mod_sel)
```



It is also possible to graphically represent the results obtained for the BIC of the different models by simply using the plot() function. The graph shows the two series of values (one for models with specific variance for each component, one for models with common variance) and their trend with respect to the number of components. The best value for the BIC index implemented in the mclust package is the larger one; the result obviously confirms what was already observed: the optimal model using the BIC index for selection is the one with two components and common variance.

Estimation of the model

```
Mclust()
summary()
```

```
modM <- Mclust(haemoglobin,
               G = 2, #number of cluster
               modelNames = "E") #E=homoschedastic model, V=heteroschedastic model
summary(modM)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -402.7969 70  4 -822.5878 -832.2396
##
## Clustering table:
##  1  2
## 43 27
```

The summary of the model only shows information related to the value of the log-likelihood at convergence (equal to -402.8), BIC and ICL indices, number of observations and degrees of freedom; the latter are computed as the number of free parameters estimated by the model. The outoput also provides the number of observations classified in each of the two groups. We observe that the first group is larger than the second, with 43 subjects compared to 27.

```
summary(mod, parameters = TRUE)
```

```
summary(modM, parameters = TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust E (univariate, equal variance) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -402.7969 70  4 -822.5878 -832.2396
##
## Clustering table:
##  1  2
## 43 27
##
## Mixing probabilities:
##         1         2
## 0.6132646 0.3867354
##
## Means:
##         1         2
##  97.43502 237.36860
##
## Variances:
##        1        2
## 2060.818 2060.818
```

The parameters of the model are (i) the miximg probabilities or weights of the two mixture components, (ii) the means, and (iii) the variances for the two mixture components.

- The two components have weights of 0.61 and 0.39, respectively; note that these values do not represent the proportions of subjects allocated to the two groups.
- The mean value for the haemoglobin for subjects in the first group is 97, much lower than the corresponding value for observations allocated to the second group (237). We can therefore characterize the two subpopulations based on the value of the haemoglobin: higher values on average for the second subpopulation, lower values on average for the first.
- The variance values for the two components are equal, having selected the most parsimonious model. The corresponding standard deviation is approximately 45; this value represents the average variability around the specific mean value for each component. (equal variance by assumption)

Classification

```
cbind()
```

```
round(head(cbind(modM$z, Cluster = modM$classification)), 4)
```

```
##                        Cluster
## [1,] 0.0009 0.9991       2
## [2,] 0.0945 0.9055       2
## [3,] 0.9999 0.0001       1
## [4,] 0.0872 0.9128       2
## [5,] 0.9997 0.0003       1
## [6,] 0.9999 0.0001       1
```

We observe that these six units are equally allocated among the two clusters; the first obser- vation is, for instance, assigned to the second cluster with a very high probability (0.999). The second observation is allocated to the same cluster, although the posterior probability is slightly lower (0.906), with a consequent higher uncertainty.

```
round(head(cbind(Haemoglobin = haemoglobin, Cluster = modM$classification)), 4)
```

```
##      Haemoglobin Cluster
## [1,]    277.9355       2
## [2,]    207.1880       2
## [3,]     30.6607       1
## [4,]    208.4985       2
## [5,]     54.9764       1
## [6,]     39.1063       1
```

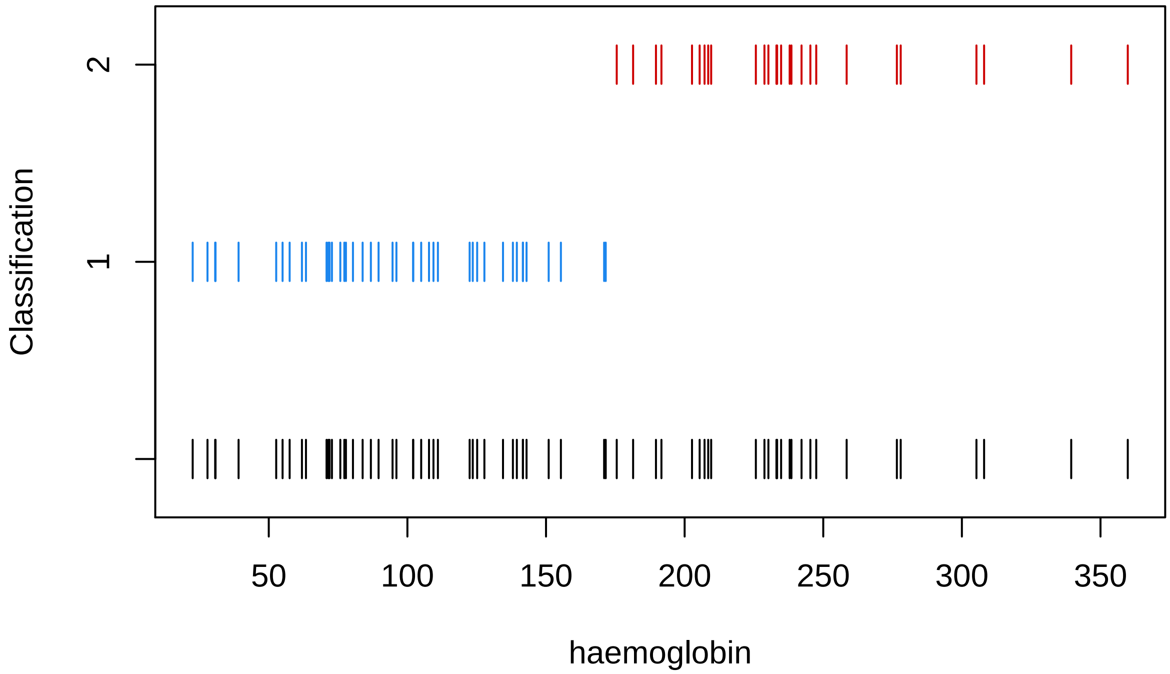
Considering these first six observations it is clear that (as already assessed analyzing the means) cluster 2 conatins individuals with the highest value of haemoglobin.

Plots

Clastering partiton graph

```
plot()
```

```
plot(modM, what = "classification")
```

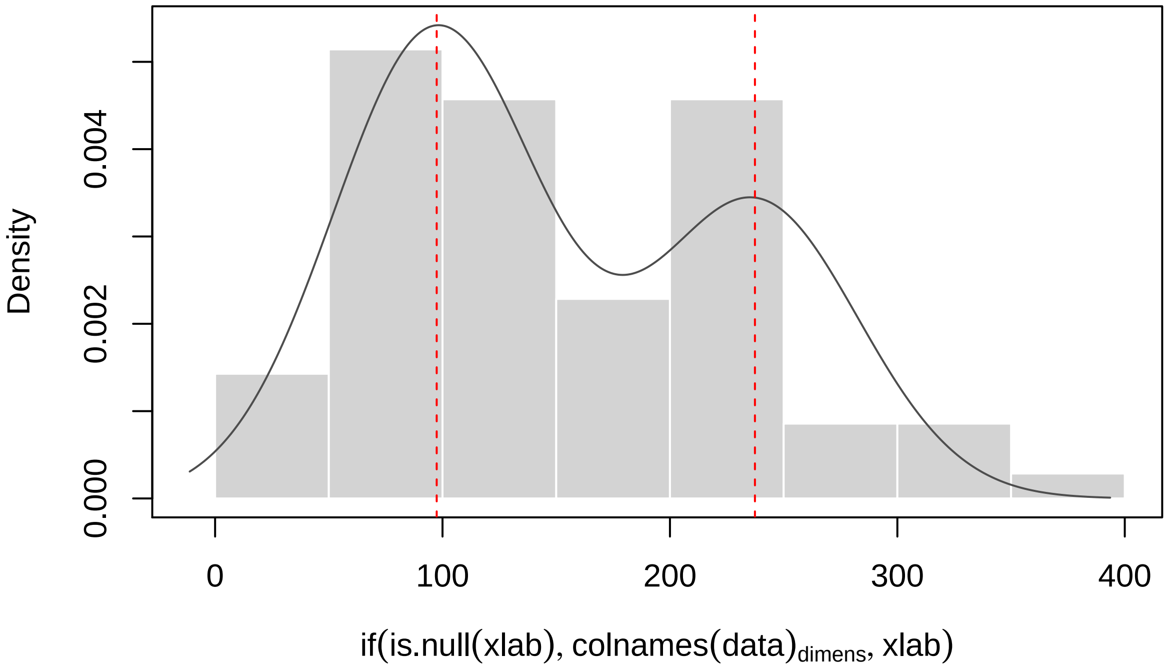


The plot represents the allocation class for each unit, based on the maximum a posteriori probability rule. We observe that the haemoglobin value that discriminates between belong- ing to one class or the other is around 170. The characterization of the first class as the one with lower haemoglobin values is confirmed. It is also observed that both classes have a reasonable number of people (we have already observed that there are 43 observations in the first class and 27 in the second).

Estimated density plot

```
plot()
```

```
plot(modM, what = 'density', xlab = "Haemoglobin", data=haemoglobin)
abline(v = modM$parameters$mean, col = rep("red", 2), lty = c(2, 2))
```



The density plot of the mixture clearly shows the two components with two peaks located at the means of the two components (as highlighted by the vertical lines at the means). The second component has a higher mean. It should be noted that the peak of the first component is higher than the other one due to the higher weight of that component (if the weights were equal, having also equal variance, the maximum of the two components would have been at the same level).

Clustering multivariate

Data desctiption

Skim

`skim_without_charts()`

```
skimr::skim_without_charts(tyr)
```

Table 3: Data summary

Name	tyr
Number of rows	122
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

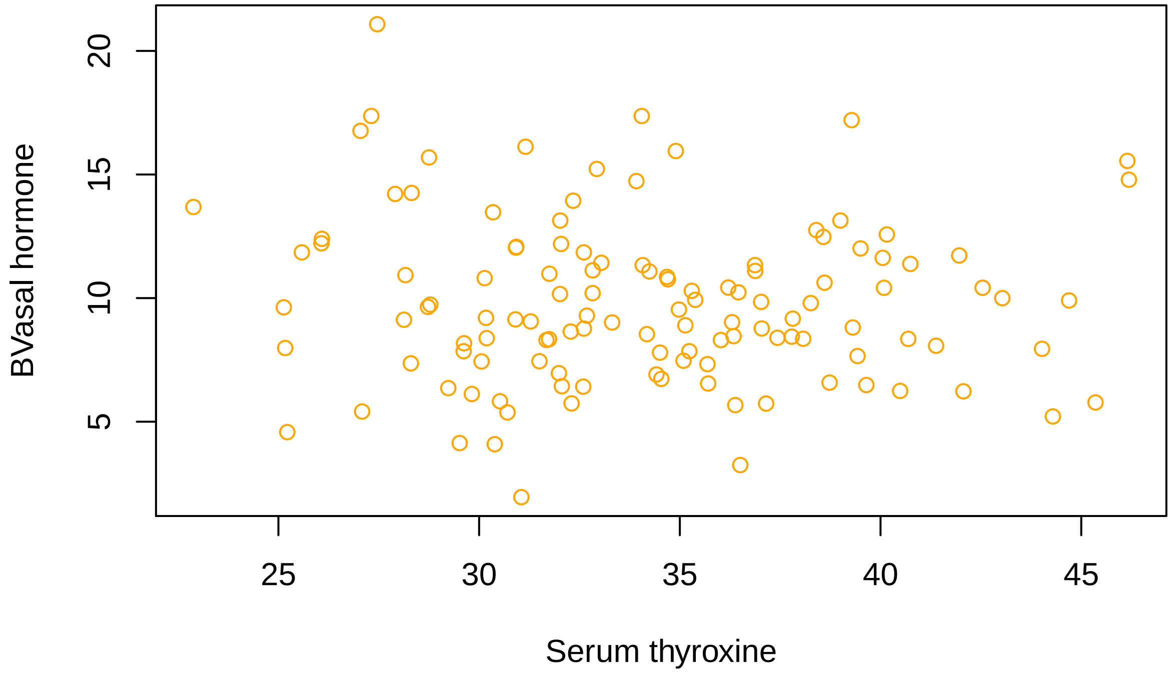
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
totser	0	1	34.19	5.04	22.88	30.57	34.06	37.7	46.19
basal	0	1	9.82	3.31	1.95	7.69	9.41	11.7	21.08

The data on serum thyroxine show a fairly wide range, with a maximum of 46 and a minimum of 23. The mean and median are approximately the same: the average level of serum thyroxine is about 34, and half of the subjects have a level below this value. The variability is quite low: the average deviation from the mean is about 5 (the coefficient of variation is equal to 0.15, indicating that the variability is about 15% of the mean). The range of variation for the level of basal hormone is slightly more limited (going approximately from 2 to 21). Again, the mean and median have the same value, just above 9. Three quarters of the subjects have a basal hormone level below 11.7.

Scatter plot

plot()

```
plot(tyr$totser, tyr$basal,
     xlab = "Serum thyroxine", ylab = "BVasal hormone",
     col = "orange")
```



Points are randomly arranged in the Cartesian plane; there is no evidence of any particular pattern among the data. It should be noted that there are some points that are isolated from the distribution; for example, there are two points with a very high value for serum thyroxine level (and also on average high for basal hormone), or a single point with a very low serum thyroxine level.

Model selection

mclustBIC()

```
require(mclust)
sel2 <- mclustBIC(tyr, modelNames = c("VII", "EII"))
sel2
```

```
## Bayesian Information Criterion (BIC):
##         VII       EII
## 1 -1412.321 -1412.321
## 2 -1406.779 -1403.102
## 3 -1418.899 -1409.324
## 4 -1429.384 -1420.405
## 5 -1445.198 -1434.345
## 6 -1457.886 -1448.502
## 7 -1475.247 -1461.857
## 8 -1491.522 -1476.074
## 9 -1492.960 -1472.579
##
## Top 3 models based on the BIC criterion:
##     EII,2     VII,2     EII,3
## -1403.102 -1406.779 -1409.324
```

The output shows the results in matrix form, with the number of components on the rows and the model specification on the columns. Considering the top three models from the perspective of the BIC index, it is found that the best choice is to select two components and assume common variability between them. This result is preferable (slightly higher BIC value) compared to the model with 2 components and specific variability.

Estimated parameters

mclustBIC()

```
modM2 <- Mclust(tyr, G = 2, modelNames = "EII")
summary(modM2, parameters = TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##       -687.1387 122  6 -1403.102 -1441.016
##
## Clustering table:
##  1  2
## 78 44
##
## Mixing probabilities:
##         1         2
## 0.6283473 0.3716527
##
## Means:
##             [,1]      [,2]
## totser 31.399569 38.916185
## basal   9.933287  9.615058
```

```
##
## Variances:
## [,,1]
##          totser    basal
## totser 11.40707  0.00000
## basal   0.00000 11.40707
## [,,2]
##          totser    basal
## totser 11.40707  0.00000
## basal   0.00000 11.40707
```

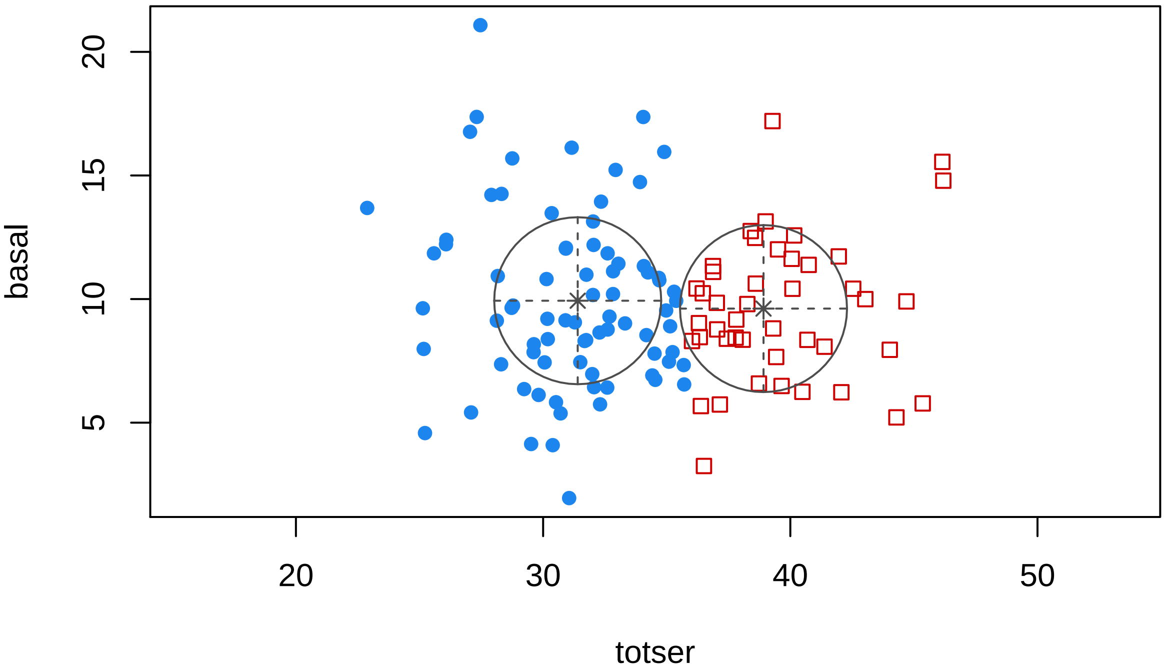
- We observe that the number of free parameters of the model is 6: this includes one of the two weights, the 4 mean values, and the single variance. The corresponding value for the log-likelihood function is -687. Additionally, BIC (used for model selection) and ICL indices based on entropy are reported.
- Regarding the classification of units into two groups, the first subpopulation is significantly larger, containing 78 out of the total 122 observations.
- The analysis of the mixing probabilities shows that the first component has a much higher weight compared to the second: note that these values do not represent the proportion of units allocated to the different classes.
- The means allow us to characterize the two subpopulations from the point of view of the application context; it is observed that the first component has a mean for the serum thyroxine level (equal to 31.4) significantly lower than the corresponding parameter for the units in the second group (38.9). Therefore, the second subpopulation is characterized by those patients with a medium-high serum thyroxine level. Conversely, regarding the basal hormone level, the difference in means between the two components is much more limited: subjects allocated to the first component have a slightly higher value, but the difference is minimal (9.9 versus 9.6). The variable related to the serum thyroxine level appears to be much more relevant for discriminating the belonging of observations to one of the two groups.
- Finally, it is found that the covariance matrix is (as required) the same for both components: in both groups, the average variability with respect to the mean is equal to 3.4 for both variables.

Plots

Clastering classification graph

```
plot(mod)
```

```
plot(modM2, "classification", asp = 1)
```



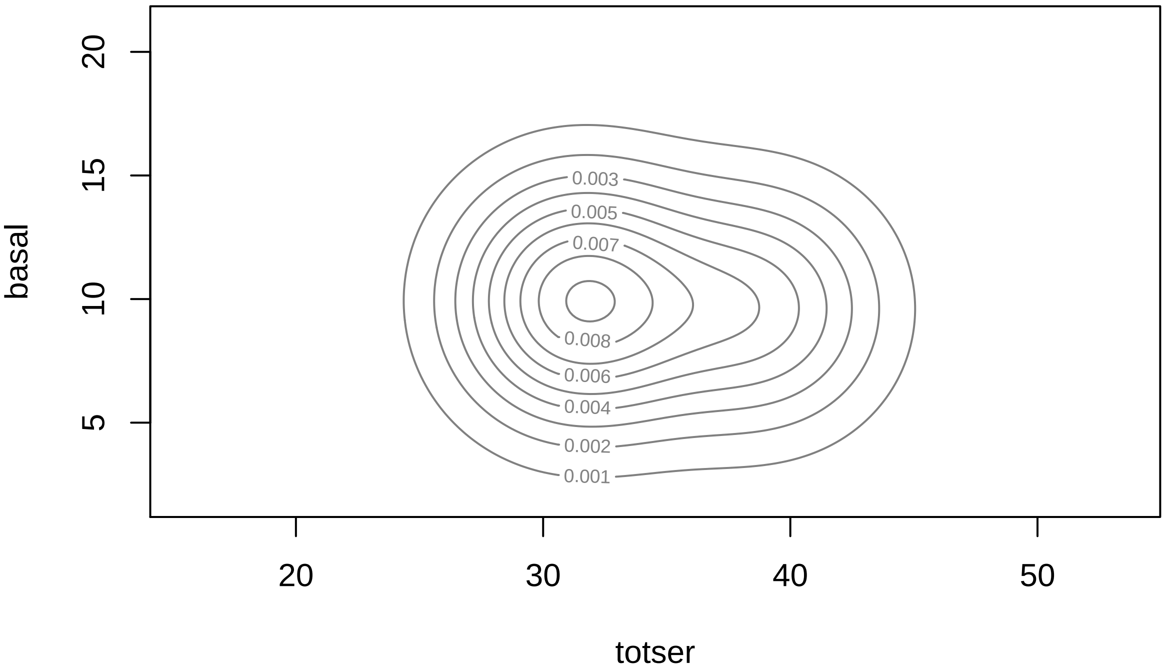
```
#Note that by using the asp = 1 option, it is possible to fix the same scale for both the x-axis and th
```

- The plot shows the arrangement of observations in the Cartesian plane with respect to the two variables (scatter plot). The points corresponding to the units are depicted with different shapes and colors to emphasize their belonging to one subgroup or the other. It is again observed that the variable measuring the white blood cell count (y- axis) is not relevant for the classification of units: the transition from one component to another does not depend on the height of the points, but only on their position along the y-axis (the hyperplane that divides the two regions is an almost vertical line).
- We also observe the presence of circles related to the two components. It is highlighted that the two circles (spherical model) are equal, not only in shape but also in volume (area in this case), indicating that the variability of the two components has been constrained to assume the same value. From the arrangements of the points in the plane, it seems evident a major variability for the points in the first cluster (blue points).
- Finally, we remark the presence of a few points locatred in a central position with respect to the two clusters. This behavior highlights a high uncertainty in the classification.

Estimated density plot

```
plot(mod, "density")
```

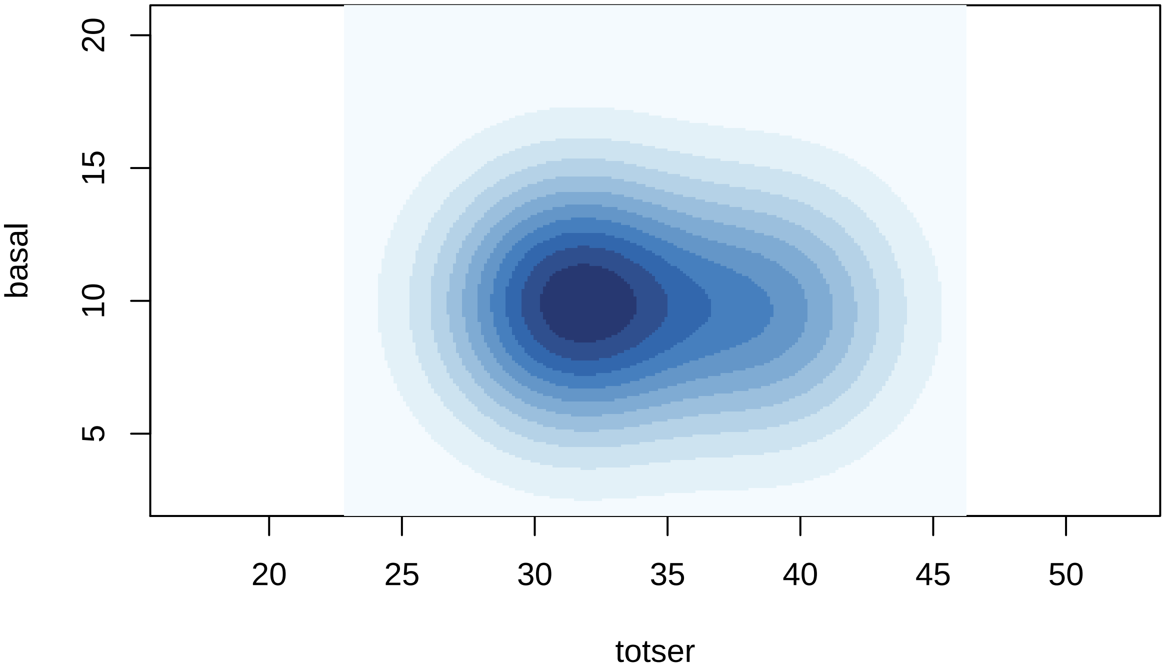
```
plot(modM2, "density", asp = 1)
```



The shape of the density is represented by means of contour lines (projections of the intersections of the 3D surface with planes parallel to the xy-plane and at different heights). It can be observed that the figure is slightly elongated in the horizontal direction, with a more pronounced extension to the right. This is the effect of a first component with a much greater weight (positioned on the left) and a second component, much less heavy, that produces the elongation. Therefore, the presence of two distinct peaks is not observed.

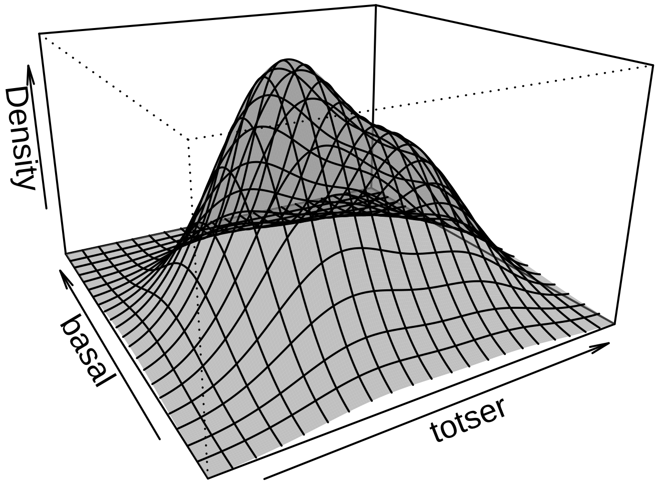
Alternative 1:

```r
plot(modM2, "density", type = "image", asp = 1)
```



Alternative 2:

```r
plot(modM2, "density", type = "persp", asp = 1)
```



Bootstrap standard errors

MclustBootstrap()

```r
set.seed(1234)
bootClust <- MclustBootstrap(modM2)
```

Estimated standard errors

summary(boot, what = "se")

```r
summary(bootClust, what = "se")
```

```
## ----------------------------------------------------------
## Resampling standard errors
## ----------------------------------------------------------
## Model                      = EII
## Num. of mixture components = 2
## Replications               = 999
## Type                       = nonparametric bootstrap
##
## Mixing probabilities:
##          1          2
## 0.09272392 0.09272392
##
## Means:
##                1         2
## totser 0.6988795 1.0842079
## basal  0.5105720 0.5159279
##
## Variances:
## [,,1]
##          totser    basal
## totser 1.296861 0.000000
## basal  0.000000 1.296861
## [,,2]
##          totser    basal
## totser 1.296861 0.000000
## basal  0.000000 1.296861
```

The standard error estimated by the bootstrap procedure are not high for every dimension, this means that the model estimation can be trusted.

Confidence intervals 95%

```
summary(boot, what = "ci")
```

```
summary(bootClust, what ="ci")
```

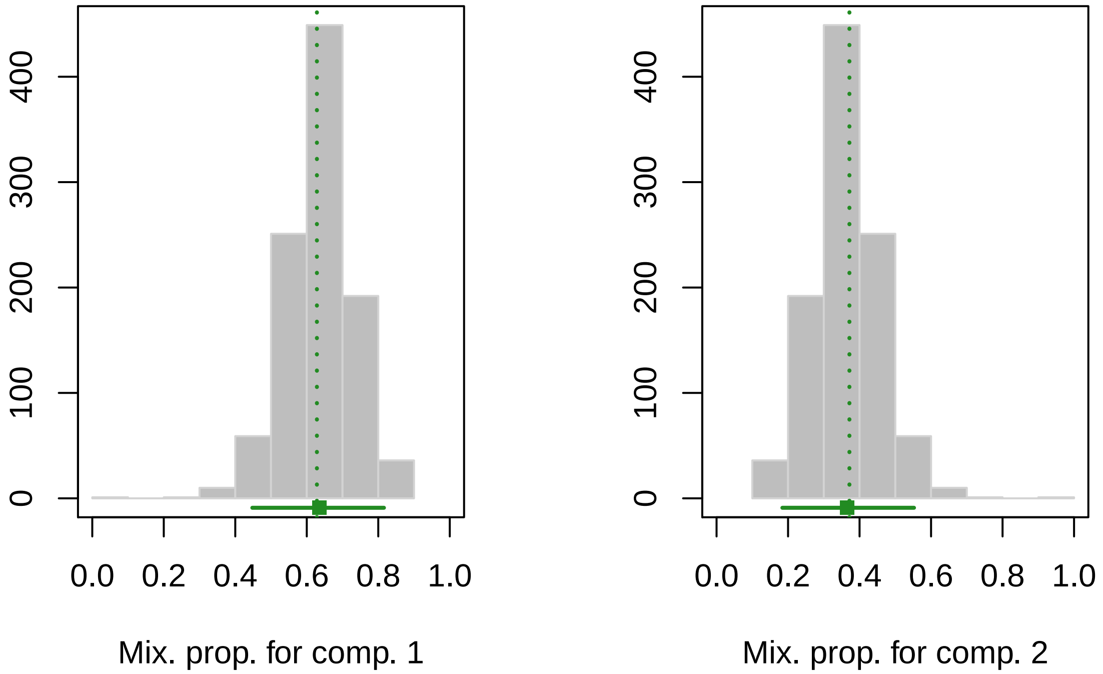
```
## ----------------------------------------------------------
## Resampling confidence intervals
## ----------------------------------------------------------
## Model                      = EII
## Num. of mixture components = 2
## Replications               = 999
## Type                       = nonparametric bootstrap
## Confidence level           = 0.95
##
## Mixing probabilities:
##               1         2
## 2.5%  0.4477164 0.1840785
## 97.5% 0.8159215 0.5522836
##
## Means:
## [,,1]
##         totser     basal
## 2.5%  29.96091  9.100781
## 97.5% 32.67200 10.914595
## [,,2]
##         totser     basal
## 2.5%  37.19429  8.760714
## 97.5% 41.55087 10.850845
##
## Variances:
## [,,1]
##          totser     basal
## 2.5%   8.849027  8.849027
## 97.5% 14.013071 14.013071
## [,,2]
##          totser     basal
## 2.5%   8.849027  8.849027
## 97.5% 14.013071 14.013071
```

This output report the confidence intervals at 95% of every parameter estimated by the models, it is visible how the intervals are quite short this is a positive signal for our model.

Plot

```
plot(boot, what = "pro")
```

```
par(mfrow = c(1,2))
plot(bootClust, what = "pro")
```



This istograms report the probability estimated for every boostap sample, it is visible how the means of the estimated probability are almost hydentical to the ones estimated by our model: 0.62 0.37.