Homeworks

Matteo Suardi

2025-04-28

Contents

1 Week 1 - Descriptive statistics for continuous variables, Observational vs. experimental data, Joint, marginal, and conditional probabilities

1.1 Basic statistical concepts: inference, variability, probabilistic models, exploration, and estimation.

1.1.0.1 1.0 Read the teaching notes and summarize the basic concepts introduced in the first lecture.

Statistics is the **science of learning from experience**, studying collective phenomena through systematic methods. With a history spanning nearly 250 years, it emerged in the 1600s when the inductive empirical method was applied to the social sciences. The first statistical tables appeared in Denmark in 1741, and **Sir Thomas Bayes'** rule was published posthumously in 1763, marking a fundamental advancement in **probability theory**. The development of statistics gained momentum when mathematics was applied to concrete problems rather than abstract concepts. Karl Pearson, a pioneer in the field, contributed to statistical methodology by developing the **chi-squared distribution**. Ronald Fisher introduced **maximum likelihood estimation**, which became a cornerstone of **statistical inference**, as well as discriminant analysis, which laid the foundation for classification methods. Andrey Markov introduced the concept of **Markov chains** in 1906, further advancing probabilistic modeling.

Throughout the 20th century, statistics evolved alongside computational advancements. In the 1950s, Lazarsfeld introduced **model-based clustering**, while Metropolis (and his colleagues) developed an algorithm for Bayesian estimation based on Markov chains. The 1960s saw John Tukey emphasizing the importance of data analysis and computation, advocating for a more application-driven approach to statistics. The **expectation-maximization algorithm**, developed by Baum in 1970, became a key tool for estimating models under the maximum likelihood framework. Around the same time, Cox proposed the **proportional hazards** model for survival analysis, greatly contributing to biostatistics. In 1979, Efron introduced **bootstrap resampling**, an inferential technique based on resampling data, while **Markov Chain Monte Carlo** methods emerged as essential computational tools in Bayesian statistics.

As large and complex datasets became increasingly available, new methodologies emerged to address classification and clustering challenges. **Finite mixture models** were introduced in 1988 to better handle these challenges, while **latent variable models** and causal inference techniques continued to develop. The rise of data science in 2016 signified a shift toward algorithmic processing of large datasets, often with minimal reliance on traditional parametric models. Data science was defined as a disicipline that extracts meaning from raw data using scientific methods, emphasizing predictive algorithms and large-scale computations. Despite that shift, statistical inference remains fundamental in connecting predictive algorithms to well-established methodological frameworks.

1.2 Data exploration: variable types, sample size, context classification, descriptive statistics, and graphical visualization in a financial setting.

Consider the data contained in the profits.Rdata file relating to simulated data on profit margins of different firms of the same sector (y), their net revenue per deposit dollar (x1) and the quota of investments in other relevant assets of the firm (x2) in the sector.

```
load("profits.Rdata")
str(profits)
```

1.2.0.1 2.0 Read the data and define variables types.

```
## 'data.frame':    500 obs. of  3 variables:
##  $ y : num  -0.0323 0.8282 4.0128 4.8241 2.4849 ...
##  $ x1: num  1.45 0.816 1.506 2.259 1.035 ...
##  $ x2: num  2.836 1.315 0.966 2.756 2.088 ...
```

The "profits" data set is a data frame with 500 observations and 3 variables, all of which are of type numeric. Specifically:

- `y` is a continuous variable representing the **profit margins** of different firms operating in the same sector.
- `x1` is a continuous variable indicating the **net revenue per deposit dollar**.
- `x2` is a continuous variable measuring the **quota of investments in other relevant assets** of the firm.

All three variables can assume both positive and negative values, which reflects the heterogeneity in economic conditions and management strategies across firms. Since the variables are quantitative and continuous, the dataset is appropriate for the application of classical statistical methods for description and inference.

```
head(profits)
```

1.2.0.2 2.1 Print and comment on the first and last six rows of data frame. Which is the sample size? Are sample data from an observational context or a randomized experiment? May the observation be assumed as a random sample?

```
##             y        x1       x2
## 1 -0.03229359 1.4498836 2.835682
## 2  0.82817699 0.8164275 1.315232
## 3  4.01284020 1.5060157 0.965941
## 4  4.82411236 2.2585572 2.755820
## 5  2.48486540 1.0347815 2.088001
## 6  1.28942814 1.4005885 2.074641
```

```
tail(profits)
```

```
##               y        x1       x2
## 495  4.27039110 2.1744819 1.211890
## 496  5.11033857 2.0416749 1.233133
## 497  0.02610409 1.2760384 3.386443
## 498  0.13394858 0.8667697 1.260811
## 499 -0.49208862 0.7435544 2.385192
## 500  3.69610848 1.8692255 2.233614
```

The sample size is equal to **500 observations**. Based on the provided description and the structure of the data, there is no indication that a randomized experimental design was adopted. Instead, the data appear to originate from a simulation mimicking firm-level financial indicators, likely inspired by real-world distributions. As such, the observations should be considered the result of an **observational study**, in which firms are not assigned to conditions but observed as they are. The absence of imposed interventions or controlled treatment assignments supports the plausibility of assuming the data to represent a **random sample** from a conceptual population of firms within the same sector.

```r
library(skimr)
skim_without_charts(profits)
```

1.2.0.3 2.2 Calculate, reports, and comments on descriptive statistics.

Table 1: Data summary

Name	profits
Number of rows	500
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
y	0	1	0.94	3.90	-10.87	-1.48	0.96	3.69	12.72
x1	0	1	0.97	1.37	-3.29	0.09	0.95	1.88	5.25
x2	0	1	1.99	0.99	-1.25	1.32	1.99	2.62	4.94

All three variables (`y`, `x1`, and `x2`) have no missing values (n_missing = 0) and complete_rate = 1, indicating that the dataset is complete and suitable for full-case analysis.

y (profit margins): The range is wide, with values spanning from -10.87 to 12.72. The mean (0.94) and median (0.96) are very close, suggesting an approximately symmetric distribution. The standard deviation (3.90) reveals substantial variability across firms, indicating heterogeneity in profit performance.

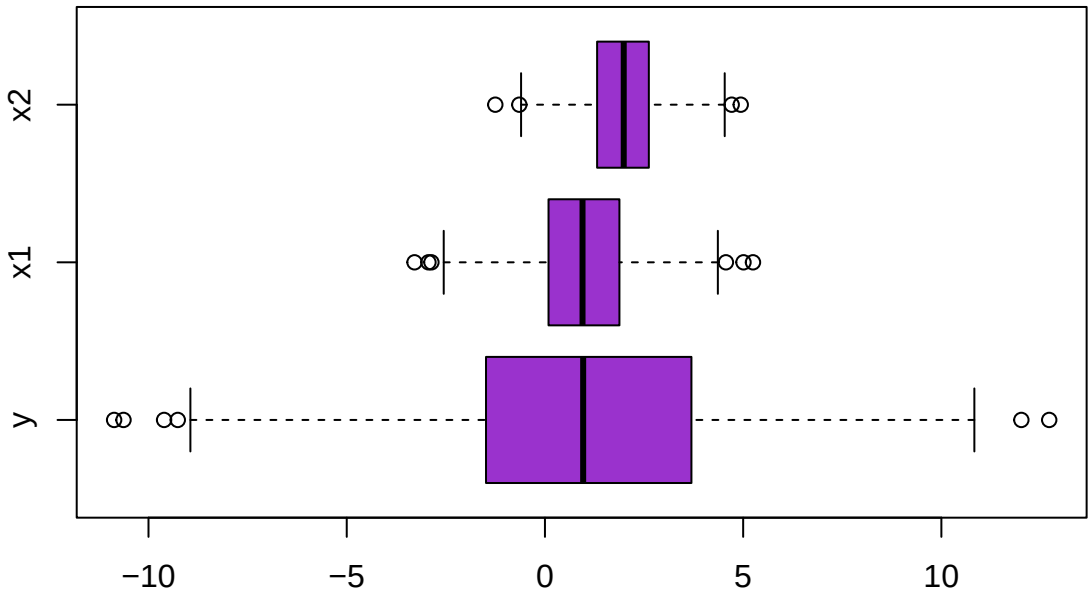
x1 (net revenue per deposit dollar): This variable ranges from -3.29 to 5.25, with a mean of 0.97 and a median of 0.95, again suggesting approximate symmetry. The presence of negative values indicates that some firms report losses per deposit unit. The interquartile range spans from 0.09 to 1.88, showing that most values are concentrated between these limits. The standard deviation (1.37) is lower than that of `y`, but still signals moderate dispersion.

x2 (quota of investments in other relevant assets): The variable ranges from -1.25 to 4.94, with both mean and median equal to 1.99, suggesting a highly symmetric distribution. The interquartile range is from 1.32 to 2.62, indicating a tight concentration around the central value. The standard deviation (0.99) is the smallest among the three variables, reflecting lower dispersion in investment behavior across firms.

Summary: The variable `y` exhibits the greatest variability, both in range and standard deviation, highlighting substantial heterogeneity in profit margins. `x1` shows moderate variability and includes both positive and negative outcomes, while `x2` is the most concentrated and symmetric variable. This suggests that while profit and revenue efficiency vary widely across firms, investment behavior tends to be more consistent. These results underline the importance of further analysis (e.g., correlation, regression) to investigate how `x1` and `x2` relate to `y`.

```r
boxplot(profits,
        horizontal = TRUE,
        col = "darkorchid3")
```

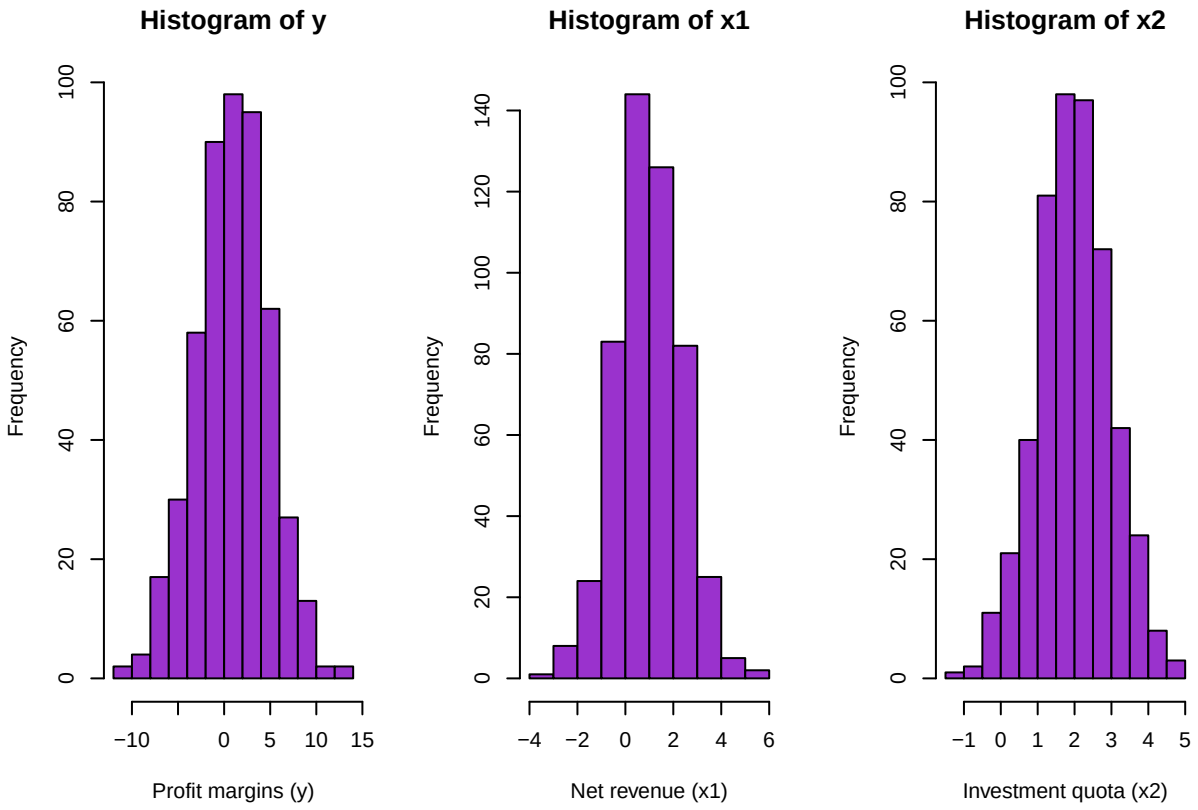
1.2.0.4 2.3 Depict one or more plots you think may be useful to visualize data features. Describe the plots and comment on each one according to the financial context on which data are



collected.

The boxplot displays the distribution of each variable and highlights the presence of outliers. Variable `y` exhibits a wide interquartile range and several extreme values, confirming its high variability. `x1` shows mild asymmetry with some outliers, while `x2` is more concentrated with relatively balanced whiskers, supporting the idea of a symmetric and low-variance distribution.

```r
par(mfrow = c(1,3))
hist(profits$y,
     main = "Histogram of y",
     col = "darkorchid3",
     xlab = "Profit margins (y)")
hist(profits$x1,
     main = "Histogram of x1",
     col = "darkorchid3",
     xlab = "Net revenue (x1)")
hist(profits$x2,
     main = "Histogram of x2",
     col = "darkorchid3",
     xlab = "Investment quota (x2)")
```



The histogram of `y` shows a bell-shaped distribution with heavier tails, suggesting that while most firms cluster around moderate profit margins, there are a few firms with extreme performances. The histogram of `x1` is approximately symmetric with a peak near zero, reflecting a central tendency but allowing for both gains and losses. The histogram of `x2` is concentrated and symmetric, indicating that most firms invest similar proportions in other relevant assets. Overall, the plots confirm the numerical summaries and provide further support for modeling these variables using techniques that account for asymmetry and heterogeneity.

1.2.1 Open questions

- **Which is the conceptual framework of statistical science proposed by Kass 2011. Describe the involved concepts and their associations.**

According to Kass, the data generated in the real world may derive from observational or experimental studies, and what is observed in the data may also be the results of unobservable quantities. The data is characterized by **regularity** and **variability**, and the exploratory analysis of the data, sometimes using algorithms or statistical or machine learning techniques, leads to syntheses that are derived from models that are not probabilistic. On the other hand, in the theoretical world, probabilistic rules are considered through the underlying random variables defining a statistical model along with ran- dom error disturbances. The parameters of the models are known quantities of interest in the population, which are estimated through algorithms, and the parameter estimates are presented along with their measure of uncertainty, thus leading to certain conclusions.

- _____

1.3 Joint, marginal, and conditional probabilities analysis with interpretation and identification of reference distributions.

Consider the following contingency table showing the observed joint probability distribution for high cash flow of a firm (X) and an evaluation of default risk (Y).

```
table <- matrix(
  c(0.0014, 0.0086, 0.8712, 0.1188),
  nrow = 2,
  byrow = TRUE,
  dimnames = list(
    "Default" = c("yes=1", "no=0"),
    "High_CF" = c("yes=1", "no=0")
  )
)

table
```

```
##         High_CF
## Default  yes=1   no=0
##   yes=1 0.0014 0.0086
##   no=0  0.8712 0.1188
```

1.3.0.1 3.1 Define the joint probabilities and interpret the observed values.

- $P(D = 1, C = 1) = 0.0014$
- $P(D = 1, C = 0) = 0.0086$
- $P(D = 0, C = 1) = 0.8712$
- $P(D = 0, C = 0) = 0.1188$

The joint distribution highlights a clear asymmetry in the allocation of probabilities. The probability of a firm being in default is given by the sum of the joint probabilities in the first row:

$$P(D = 1) = P(D = 1, C = 1) + P(D = 1, C = 0) = 0.0014 + 0.0086 = 0.01.$$

Thus, only 1% of the firms are expected to default, indicating that default is a rare event in this population.

Similarly, the probability that a firm has high cash flow is:

$$P(C = 1) = P(D = 1, C = 1) + P(D = 0, C = 1) = 0.0014 + 0.8712 = 0.8726.$$

This means that more than **87%** of firms in the sample show high cash flow levels.

Interpreting these values, we observe that default is rare, and even within the subset of firms that default, only a small fraction (14%) are firms with high cash flow. Conversely, the large mass of probability lies in the cell $(D=0, C=1)$, i.e., firms that are not in default and do have high cash flow (87.12%). This aligns with economic theory: firms with abundant liquidity and earnings are less likely to face financial distress or default.

1.3.0.2 3.2 Calculate the marginal probabilities of the two variables and comment on both.

From the joint table, the marginal probabilities are obtained by summing rows and columns.

Marginal probabilities for D (Default):

- $P(D = 1) = 0.0014 + 0.0086 = 0.01$

- $P(D = 0) = 1 - P(D = 1) = 0.99$

This confirms that default occurs in only **1%** of cases, which is consistent with real-world credit-risk datasets where defaults are rare.

Marginal probabilities for C (High Cash Flow):

- $P(C = 1) = 0.0014 + 0.8712 = 0.8726$
- $P(C = 0) = 1 - P(C = 1) = 0.1274$

The high frequency of firms with $C=1$ (87%) suggests a population predominantly composed of healthy or liquid companies. Such a distribution could stem from sample selection (e.g., excluding distressed firms), sector-specific dynamics, or simulated assumptions.

1.3.0.3 3.3 Which is the reference distribution of each single variable? Which the value of the parameter?

Both `D` (default) and `C` (cash flow) are binary variables. When considered independently (i.e., marginally), each follows a **Bernoulli distribution** with a parameter equal to the probability of observing a success (i.e., the value equal to 1).

- For Default:
  - $D \sim \text{Bernoulli}(p = 0.01)$
- For High Cash Flow:
  - $C \sim \text{Bernoulli}(p = 0.8726)$

Thus, the distribution of each variable is completely characterized by a single parameter corresponding to the marginal probability that the variable equals 1.

1.3.0.4 3.4 Calculate the conditional probability of $\mathbf{Y = 1|X = 1}$.

We now compute the conditional probability that a firm defaults ($Y=1$), given that it has high cash flow ($X=1$), using the classical formula: $P(Y = 1|X = 1) = \frac{P(Y=1,X=1)}{P(X=1)}$.

Given:

- $P(Y = 1, X = 1) = 0.0014$
- $P(X = 1) = 0.8726$

```
round(0.0014 / 0.8726, 4)
```

```
## [1] 0.0016
```

$P(Y = 1|X = 1) \approx 0.0016$

This result implies that, **conditional on having high cash flow**, the probability of default is only **0.16%**, which is **extremely low**. This reinforces the economic intuition: liquidity acts as a protective factor against default, and firms with strong cash flow are far less likely to experience financial distress. In practical terms, this kind of result may be crucial for credit scoring models or for early-warning systems in finance.

1.4 Correlation structure analysis: raw and partial correlations, univariate distributions, and scatterplot matrix interpretation.

With reference to the data of Exercise 2:

1.4.1 4.1 Calculate the sample correlation matrix and comment on the results.

We compute the correlation matrix between the three quantitative variables in the dataset `profits`. Each coefficient ranges from -1 to 1 and quantifies the strength and direction of the linear relationship between a pair of variables.

```
round(cor(profits), 3)
```

```
##        y    x1    x2
## y  1.000 0.963 0.617
## x1 0.963 1.000 0.717
## x2 0.617 0.717 1.000
```

The output shows a **very strong positive correlation** (equal to 0.976) between **profit margins** (`y`) and **net revenue per deposit dollar** (`x1`). This result is consistent with the financial intuition: the more a company manages to generate net revenues for each dollar of deposit, the more its profit margins increase.

There is a **moderately high positive correlation** (equal to 0.741) between **profit margins** and the **share of investments in other assets** (`x2`). A higher investment in alternative assets is associated with higher profits, but the relationship is less strong than the one observed with `x1`.

There is also a **strong positive correlation** (equal to 0.725) between **net revenue per deposit dollar** and **investment in other assets**. Companies that generate high revenues from deposits also tend to invest more in other assets.

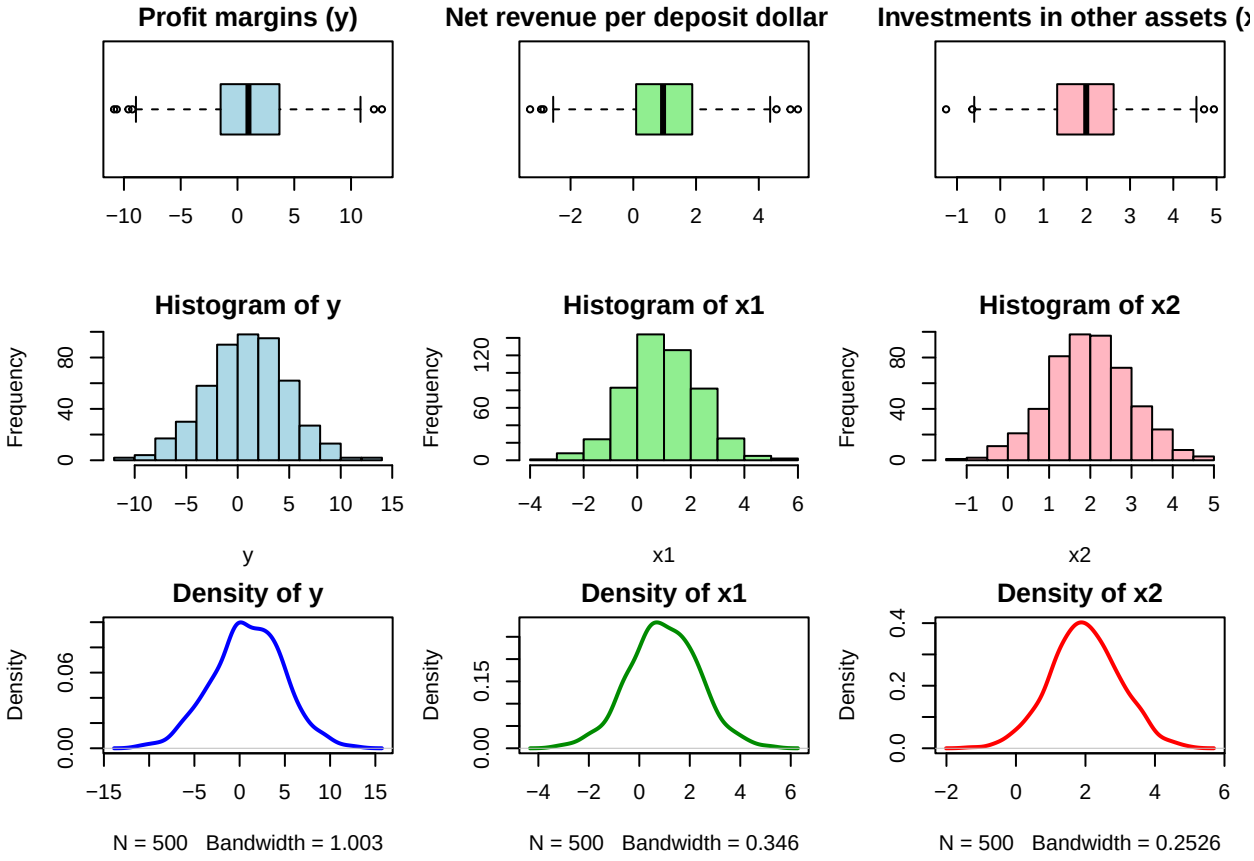
These strong associations suggest that multicollinearity could be an issue if we were to include all covariates in a regression model, which might affect the precision of the estimated coefficients.

1.4.2 4.2 Depict all the univariate plots such as boxplots, histograms and density plots for each variable. Comment on each one and on the observed differences among them.

We now visually inspect the marginal distribution of each variable using boxplots, histograms, and density curves. These plots help assess the central tendency, dispersion, presence of outliers, and the shape of the distribution (e.g. symmetry, modality).

```
par(mfrow = c(3, 3), mar = c(4, 4, 2, 1))

# Boxplots
boxplot(profits$y, main = "Profit margins (y)", col = "lightblue", horizontal = TRUE)
boxplot(profits$x1, main = "Net revenue per deposit dollar (x1)", col = "lightgreen", horizontal = TRUE)
boxplot(profits$x2, main = "Investments in other assets (x2)", col = "lightpink", horizontal = TRUE)

# Histograms
hist(profits$y, main = "Histogram of y", col = "lightblue", xlab = "y")
hist(profits$x1, main = "Histogram of x1", col = "lightgreen", xlab = "x1")
hist(profits$x2, main = "Histogram of x2", col = "lightpink", xlab = "x2")

# Density plots
```

```r
plot(density(profits$y), main = "Density of y", col = "blue", lwd = 2)
plot(density(profits$x1), main = "Density of x1", col = "green4", lwd = 2)
plot(density(profits$x2), main = "Density of x2", col = "red", lwd = 2)
```



- The **boxplot of `y` (profit margins)** shows a symmetric distribution with no evident outliers. The median is well-centered within the interquartile range, suggesting balanced variation.
- The **boxplot of `x1` (net revenue per deposit dollar)** is also symmetric and tightly concentrated, with a narrower IQR compared to `y`. There are no extreme values and the distribution appears quite compact.
- The **boxplot of `x2` (investments in other assets)** is slightly more spread, but still symmetric, showing greater variability and a wider interquartile range. One mild outlier may be present in the upper tail.
- The **histograms** and **density plots** confirm these findings. All three variables show **unimodal and approximately symmetric distributions**, though `x2` exhibits **more dispersion** and slightly **heavier tails**. No strong skewness or multimodality is observed.

In conclusion, all three variables are fairly well-behaved from a distributional standpoint. The different spreads suggest that standardization may be useful in multivariate modeling to ensure that each variable contributes comparably.

1.4.3 4.3 Provide the matrix of the sample partial correlations and comment on the values.

To evaluate the strength of association between variables after removing the effect of the others, we compute the **partial correlation matrix**. This allows us to isolate the direct relationships between each pair of variables, net of confounding effects.

We use the `parcor()` function from the `ggm` package on the sample covariance matrix:

```
library(ggm)

S4 <- cov(profits)
round(parcor(S4), 3)
```

```
##         y   x1     x2
## y   1.000 0.95 -0.394
## x1  0.950 1.00  0.580
## x2 -0.394 0.58  1.000
```

The **partial correlation between `y` and `x1`** is extremely high (0.950), confirming a strong and direct association between **profit margins** and **net revenue per deposit dollar**, even after adjusting for the effect of `x2`.

The **partial correlation between `y` and `x2`** is negative (-0.394), indicating that once the effect of `x1` is removed, the relationship between **profit** and **investment in other assets** becomes moderately negative. This suggests a suppressor effect: while the raw correlation was positive, this direct relationship is in the opposite direction when `x1` is controlled for.

Finally, the **partial correlation between `x1` and `x2`** remains moderately high (0.580), implying that even after adjusting for `y`, **net revenue per deposit dollar** and **investments in other assets** share a non-negligible linear association. This may reflect a common strategic or structural pattern in firm management.

In conclusion, partial correlations offer a clearer view of the unique contributions of each variable. In particular, `x1` retains a central explanatory role for `y`, while the role of `x2` is more complex and possibly affected by its interaction with `x1`.

1.4.3.1 4.4 Depict the `corrplot` of the raw and partial correlations and comment on the figure

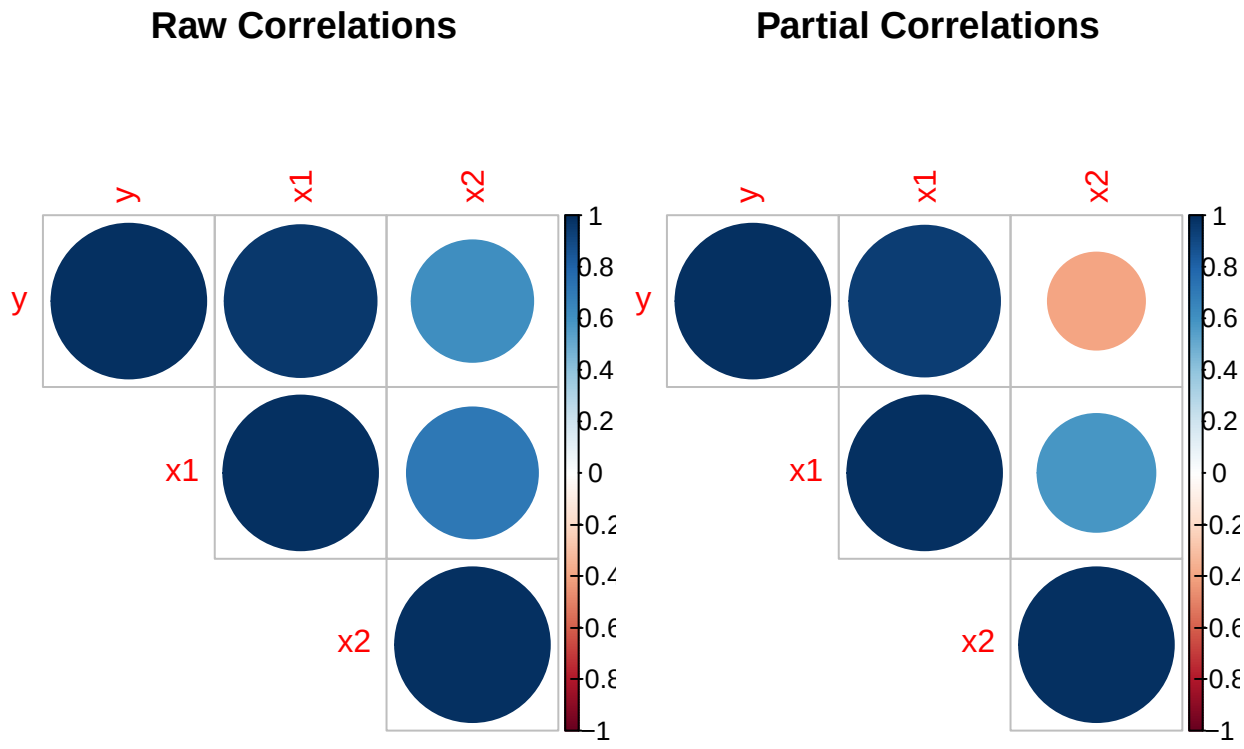
To provide a visual comparison between the raw and partial correlations among the variables, we use the `corrplot()` function. This enables us to highlight how the strength and direction of linear relationships change when controlling for the effect of the remaining variable.

We compute the two matrices (raw and partial) and plot them side by side.

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(ggm)

C_raw4 <- cor(profits) # compute raw correlation matrix

S4 <- cov(profits) # compute partial correlation matrix
C_partial4 <- parcor(S4)

par(mfrow = c(1, 2))
corrplot(C_raw4, type = "upper", title = "Raw Correlations", mar = c(0, 0, 2, 0))
corrplot(C_partial4, type = "upper", title = "Partial Correlations", mar = c(0, 0, 2, 0))
```



The **raw correlation plot** confirms the presence of strong positive associations among all three variables, especially between `y` and `x1`, and to a lesser extent between `y` and `x2`.

On the other hand, in the **partial correlation plot**, the relationship between `y` and `x1` remains very strong and positive, but the association between `y` and `x2` turns negative. This change highlights the fact that the observed positive correlation between `y` and `x2` in the raw matrix was **indirect**, driven by their mutual association with `x1`. Once the influence of `x1` is removed, the true nature of the relationship is revealed.

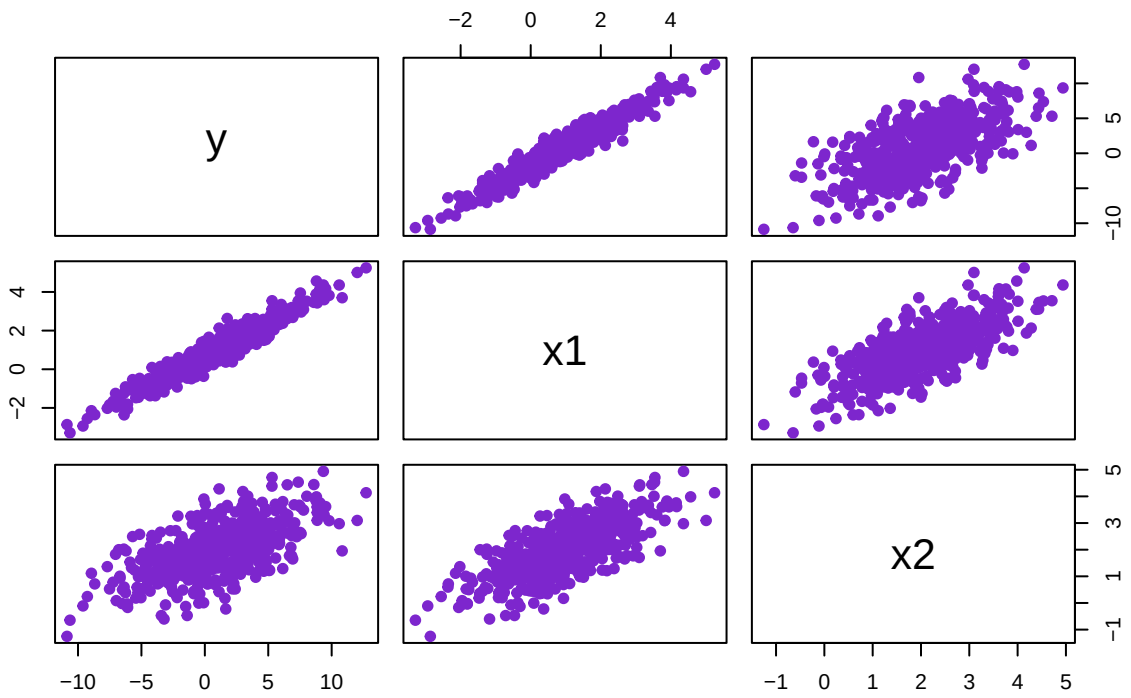
The plot also shows that `x1` and `x2` maintain a moderately strong direct association, which suggests that these two predictors share explanatory information, but are not collinear.

This visual analysis supports the interpretation from the previous exercise and emphasizes the importance of partial correlation analysis when interpreting multivariate relationships.

1.4.4 4.5 Depict the scatterplot matrix and comment on each plot

We now represent the scatterplot matrix to visually assess the relationships between all pairs of variables in the dataset. This matrix provides an immediate sense of linearity, direction, and spread of the points, supporting the interpretation of correlations already computed.

```
pairs(profits,
      main = "Scatterplot Matrix of Profits Dataset",
      pch = 19,
      col = "purple3")
```



Each subplot within the scatterplot matrix displays the bivariate relationship between two variables:

- `y` **vs** `x1`: A very strong **positive linear relationship** is clearly visible, confirming the high raw and partial correlation observed previously. The points lie closely along an upward-sloping line, indicating that higher values of net revenue per deposit (`x1`) are associated with higher profit margins (`y`).
- `y` **vs** `x2`: The points are more **dispersed** and show a **slight positive association** overall. However, this relationship weakens when controlling for `x1`, as observed in the partial correlation matrix. The visual spread supports this, suggesting that the marginal positive trend is not robust.
- `x1` **vs** `x2`: A **moderate positive relationship** can be observed, with points forming a loose upward pattern. This suggests that companies with higher net revenue per deposit tend to also invest more in other assets, although the variability is higher.

In summary, the scatterplot matrix confirms and illustrates the key associations identified through correlation analysis. The strongest and most stable relationship is the one between `y` and `x1`, which is both statistically and visually evident. The plots also highlight the necessity of considering **conditional relationships**—as those revealed through partial correlations—when interpreting associations in multivariate contexts.

1.5 Descriptive and graphical analysis: variable types, summary statistics, ECDFs, and Q-Q plots with interpretation in context.

Consider the file `school.Rdata` containing data referred to 388 school districts in California. The following variables are measured for each school district: total number of enrolled students (`students`), percentage of students qualified for income assistance (`calworks`), expenditure per student (`expenditure`), district average income (`income`), and general score (`score`) averaging reading and mathematical skills.

1.5.0.1 5.0 Print and describe the first six row of the dataframe. Which variable types are observed? Data are from an observational study?

We load the dataset and use the `head()` function to display the values of all the columns for the first 6 rows (note that there exists a similar function, `tail()`, to obtain the last 6 observations instead).

```
load("school.Rdata")
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`.

1.5.0.2 5.1 Report the main descriptive statistics: mean, median, quartiles, range and standard deviation of each variable. Comment on the results according to the applicative context. Which is the variable with the highest variability. Can you provide an explanation for this?

We use the `skimr` package to summarize the main descriptive statistics of the five observed variables:

```
library(skimr)
skim_without_charts(school)
```

Table 3: Data summary

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

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
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

The variable `students`, representing the number of pupils in each district, exhibits an extremely high range, going from a minimum of 81 to a maximum of over 12,000, with a standard deviation of approximately 3801. This confirms a very large variability in the size of school districts.

The variable `calworks` reports the percentage of students eligible for income support. The average value is approximately 13.4%, with values ranging from 0% to over 55%. This indicates that while some districts have no students eligible for assistance, others face high levels of socio-economic vulnerability. The coefficient of variation is very high (standard deviation = 10.79), confirming the strong heterogeneity across districts.

`expenditure`, which measures per-student educational spending, has a mean of around $5267 and a relatively small standard deviation (602), indicating that spending per student tends to be relatively homogeneous across districts.

`income`, which measures the average income per student's family (in thousands of dollars), presents a wide range of variation from less than $500 to over $32,000. The standard deviation (6.26) is relatively high with respect to the mean (14.08), indicating substantial economic disparity between districts.

Finally, `score` refers to the average test score of students in the district. It shows relatively low dispersion (sd 1.15), with most values between 3.3 and 6.6. The distribution appears symmetric, with mean and median values close to 5.89 and 5.78 respectively.

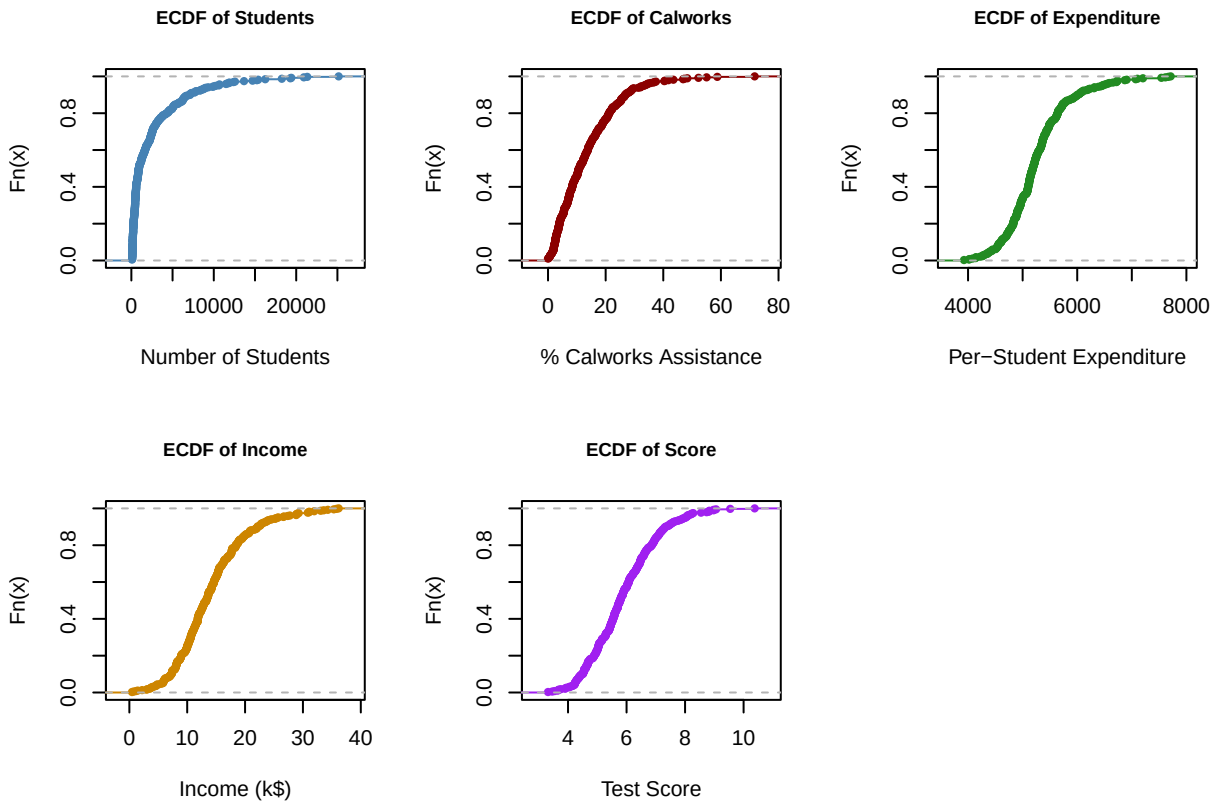
In conclusion, the variables with the highest variability are `students` and `income`. The first reflects structural differences in district size, while the second captures important socio-economic disparities. Both aspects are likely to impact educational outcomes and thus deserve further investigation.

1.5.0.3 5.2 Plot the empirical cumulative distribution function for each variable separately. What can be observed in each plot?

To visualize the distributional characteristics of each variable, we plot the empirical cumulative distribution function (ECDF) for all five variables in the dataset. This graphical tool allows us to observe the shape of each distribution and detect asymmetries, plateaus, or abrupt changes.

```
par(mfrow = c(2, 3))  # 2 rows, 3 columns

plot(ecdf(school$students),
     main = "ECDF of Students",
     xlab = "Number of Students",
     col = "steelblue",
     cex.main = 0.9)

plot(ecdf(school$calworks),
     main = "ECDF of Calworks",
     xlab = "% Calworks Assistance",
     col = "darkred",
     cex.main = 0.9)

plot(ecdf(school$expenditure),
     main = "ECDF of Expenditure",
     xlab = "Per-Student Expenditure",
     col = "forestgreen",
     cex.main = 0.9)
```

```r
plot(ecdf(school$income),
     main = "ECDF of Income",
     xlab = "Income (k$)",
     col = "orange3",
     cex.main = 0.9)

plot(ecdf(school$score),
     main = "ECDF of Score",
     xlab = "Test Score",
     col = "purple",
     cex.main = 0.9)
```



The empirical cumulative distribution functions (ECDFs) displayed in the figure provide a visual summary of how the values of each variable are distributed across the school districts:

- **Students**: The ECDF for the number of students is steep at the beginning and levels off quickly, suggesting a right-skewed distribution. Most school districts have a relatively small student population, while a few have very large enrollments, acting as outliers.
- **Calworks**: The ECDF of the percentage of students eligible for income assistance (`calworks`) increases steadily, indicating moderate skewness. The curve shows that most districts have a low-to-moderate percentage of students receiving aid, but a few exceed 60%.
- **Expenditure**: The ECDF for per-student expenditure suggests a more symmetric distribution, though the steep increase between approximately 5000 and 6500 indicates a strong concentration of values within this range.

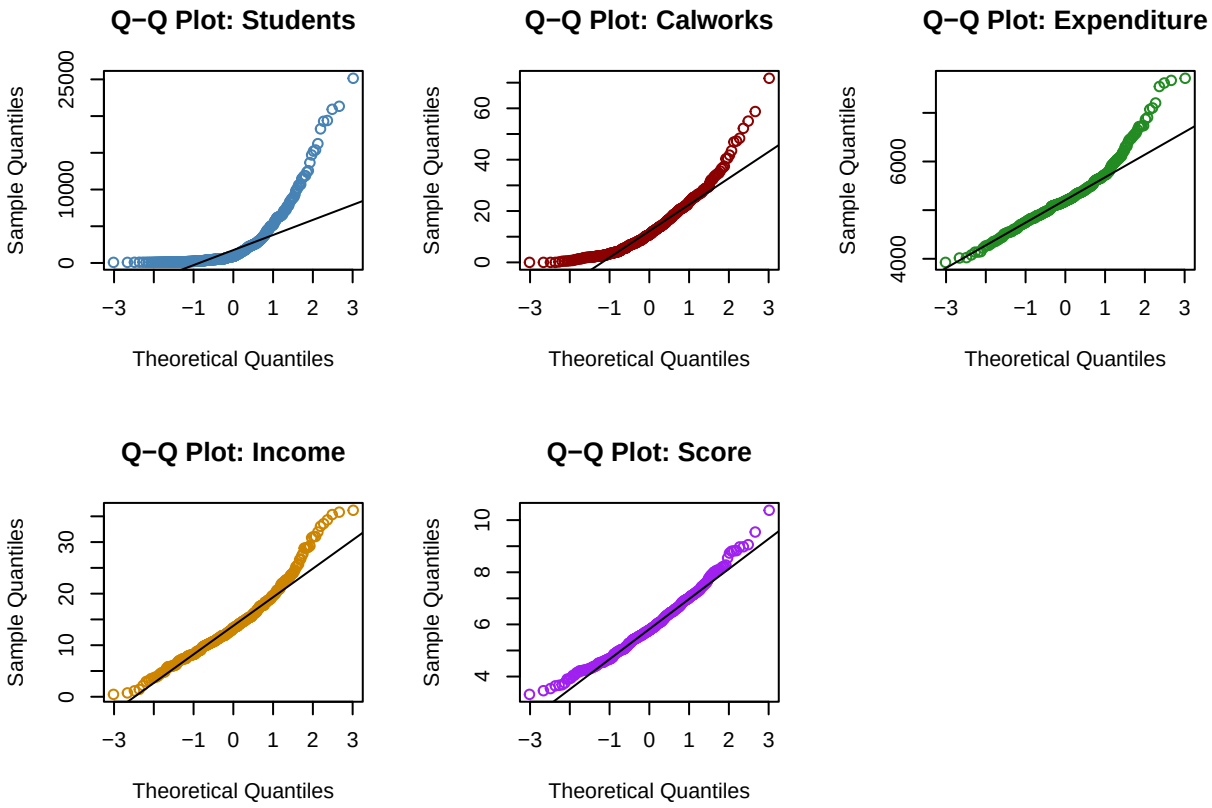
- **Income**: The ECDF of income grows roughly linearly through the central portion, which is typical of a fairly symmetric distribution. The range is wide, but no extreme discontinuities are observed.
- **Score**: The ECDF for test scores also appears symmetric and smooth, suggesting a well-distributed variable with no extreme skewness. Most values lie between 5 and 8, and the curve steepens near the mean.

These ECDF plots help detect asymmetries, concentration zones, and potential outliers in the data, complementing the earlier descriptive statistics. Notably, the variable `students` exhibits the highest dispersion, while `score` and `income` are much more concentrated.

1.5.0.4 5.3 Depict the Q-Q plots in the same graphical window. Comment on the comparison among the figures.

To assess the normality assumption for each variable, we produce the Q-Q plots for all variables in a single graphical window. These plots compare the empirical quantiles of each variable to the theoretical quantiles of a standard normal distribution. If the points lie approximately along the diagonal line, the variable is approximately normally distributed.

```
par(mfrow = c(2, 3))  # set layout for multiple plots
qqnorm(school$students, main = "Q-Q Plot: Students", col = "steelblue")
qqline(school$students)

qqnorm(school$calworks, main = "Q-Q Plot: Calworks", col = "darkred")
qqline(school$calworks)

qqnorm(school$expenditure, main = "Q-Q Plot: Expenditure", col = "forestgreen")
qqline(school$expenditure)

qqnorm(school$income, main = "Q-Q Plot: Income", col = "orange3")
qqline(school$income)

qqnorm(school$score, main = "Q-Q Plot: Score", col = "purple")
qqline(school$score)
```



- **Students**: The Q-Q plot for the number of students shows significant deviations from the diagonal line, particularly in the upper tail. This confirms the presence of high outliers and a strong right-skewness in the distribution.
- **Calworks**: The plot suggests moderate skewness and deviation from normality in the upper tail. Most points lie along the line, but some curvature and extreme values are present.
- **Expenditure**: This variable exhibits the most linear pattern among the variables. There are only mild deviations from normality in the tails, suggesting that expenditure is approximately normally distributed.
- **Income**: The distribution of income is fairly symmetric, but there are slight deviations in both tails, which may be attributed to mild outliers or non-normal features.
- **Score**: The Q-Q plot for the score variable is fairly linear, with minor deviations in the extreme values. Overall, this variable seems to be approximately normal.

Among all the variables, `expenditure`, `score`, and `income` appear to be reasonably close to normality. The variables `students` and `calworks` exhibit more pronounced skewness and heavy tails, indicating that a normality assumption for these would be questionable without transformation.

2 Week 2 - Simulation and graphical analysis of Gaussian and Student-t distributions, Probability calculations with Gaussian and Student-t distributions, Properties of the bivariate Gaussian distribution, Covariance matrix, correlation, and spectral decomposition

2.1 Simulation and comparison of Gaussian and Student's t distributions using histograms, ECDFs, and Q-Q plots.

2.1.0.1 6.1 Consider the following univariate Gaussian distribution X ~ N(7,2); fix the value of the seed to 130 and generate 1000 realizations from X. Comment on the generated data, depict the histogram of the obtained values, the cumulative empirical distribution function and the q-qplot. Comment on each figure.

We set the seed to 130 using the `set.seed()` function to ensure reproducibility of results, and generate 1000 realizations from the Gaussian distribution $X\ N(7,2)$ using the `rnorm()` function.

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

We describe the generated values using univariate summaries, using the `summary()` function.

```
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 generation 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 2.4, 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, using the `var()` and `IQR()` functions.

```
var(x)
```

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

```
IQR(x)
```

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

The variance takes a value of 2.13, which is also very close to the theoretical value of the assumed random variable. 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.

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

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

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

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

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

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

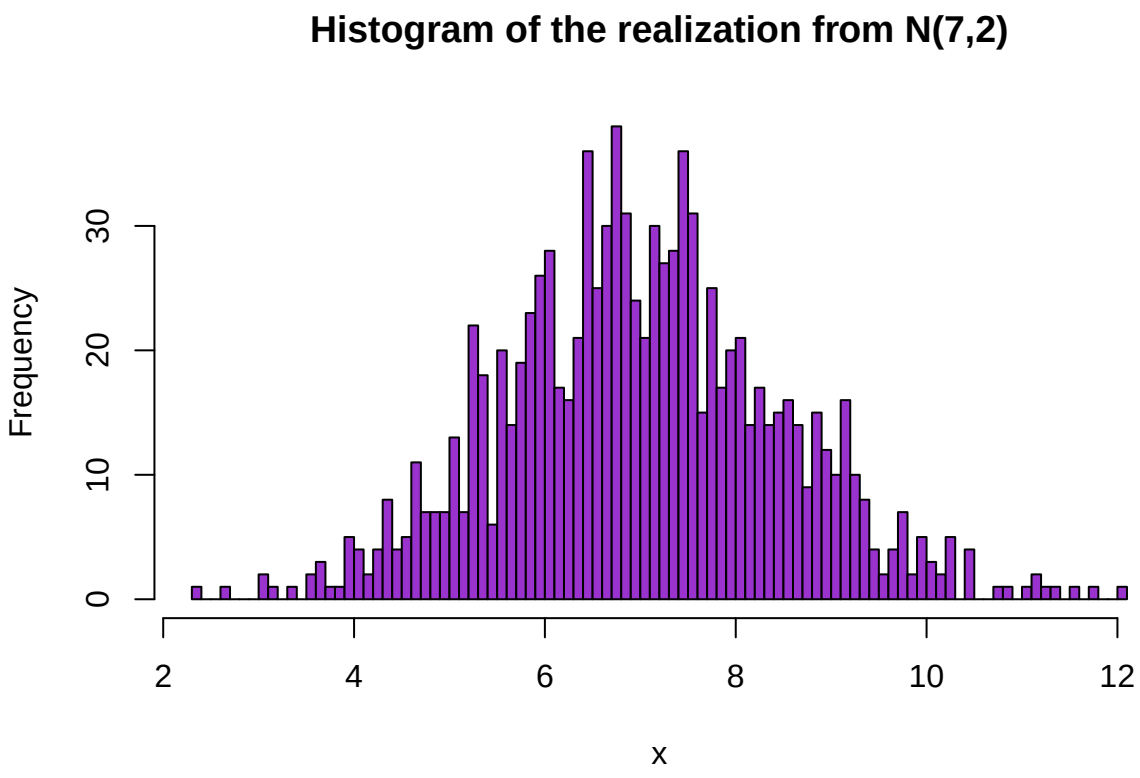
```r
length(x[x > x_Max])
```

```
## [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.

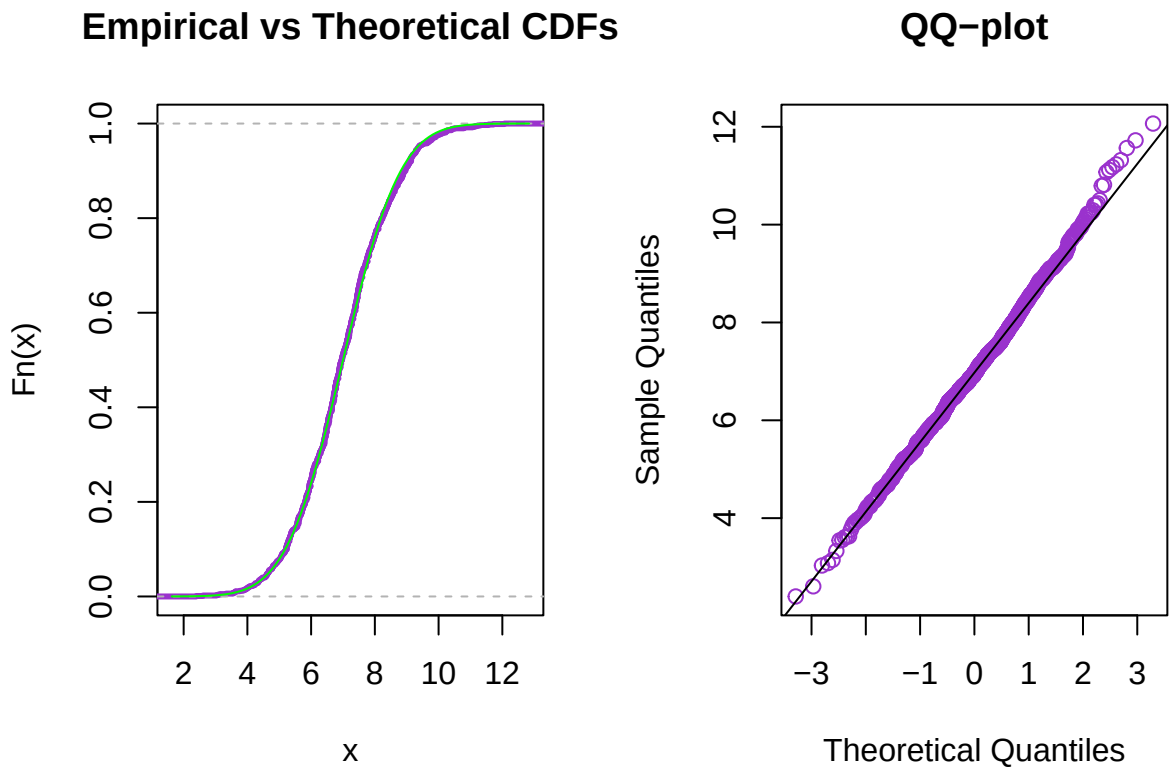
We draw a histogram of the obtained values using the `hist()` function.

```r
hist(x,
     main = "Histogram of the realization from N(7,2)",
     breaks = 100,
     col = "darkorchid3")
```



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. Finally, to assess the normality of the generated values, we depict the cumulative empirical distribution function and the QQ-plot.

```
par(mfrow = c(1,2))

plot(ecdf(x),
     col = "darkorchid3",
     lwd = 3,
     do.points = FALSE,
     main = "Empirical vs Theoretical CDFs")
curve(pnorm(x, mean = 7, sd = sqrt(2)),
      col = "green",
      lwd = 1,
      add = TRUE)

qqnorm(x,
       col = "darkorchid3",
       main = "QQ-plot")
qqline(x)
```



Both plots confirm the normality of the data, since: - the empirical cumulative distribution function perfectly

follows the theoretical one, and - the points in the QQ-plot are distributed along the straight line (bisector of first-third quadrants).

2.1.0.2 6.2 Consider the following Student's t distribution X ~ Tν for two different choices for the number of degree of freedom: $\nu = 1$ and $\nu = 10$. For both cases, fix the value of the seed to 130 and generate 1000 realization from the corresponding distribution. Compare the results by drawing the histograms of the obtained data, the empirical cumulative empirical distribution function and the q-qplot. Comment on each figure.

We generate 1000 realizations from a random variable having a Student-t distribution with $v = 1$ and $v = 10$, using a fixed seed of 130 for both values of the parameter.

```
set.seed(130)
t1 <- rt(1000, df=1)
set.seed(130)
t2 <- rt(1000, df=10)
```

We briefly inspect the generated data using the `summary()` function.

```
summary(t1)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -272.2772   -0.8988   -0.0183    2.2518    0.8977 1887.7612
```

```
summary(t2)
```

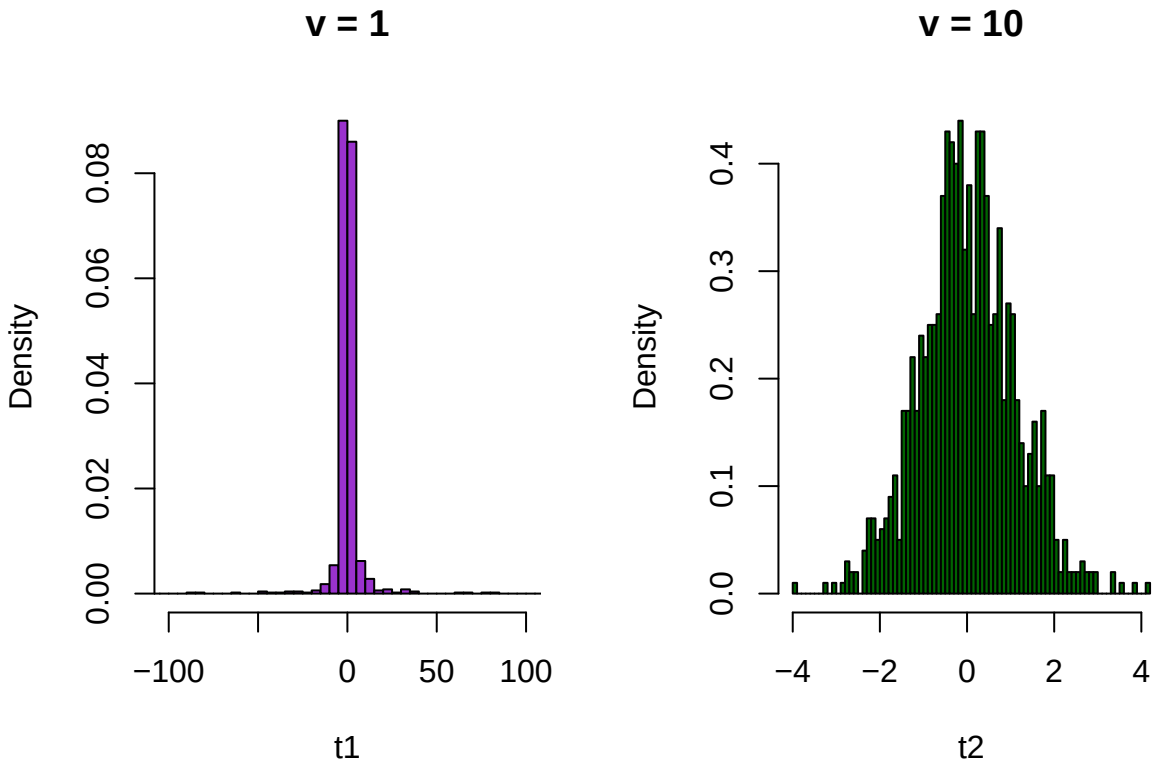
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.90982 -0.66071 -0.02284  0.02171  0.71332  4.16327
```

The first set of data (generated with $v = 1$) presents a very wide range of variation, with maximum and minimum values of -272.27 and 1887.76, 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 ($v = 10$), , 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.

We now represent, in the same graphical window, the histograms of the generated data. Regarding the first set of data ($v = 1$), 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). The `freq = FALSE` argument specifies that the y-axis of the resulting histogram should display the density of the data instead of the frequency count. The `breaks` argument determines the number of intervals (or bins) in which the range of the data is divided.

```
par(mfrow = c(1,2))
hist(t1,
     main = "v = 1",
     col = "darkorchid3",
     breaks = 500,
     freq = FALSE,
```

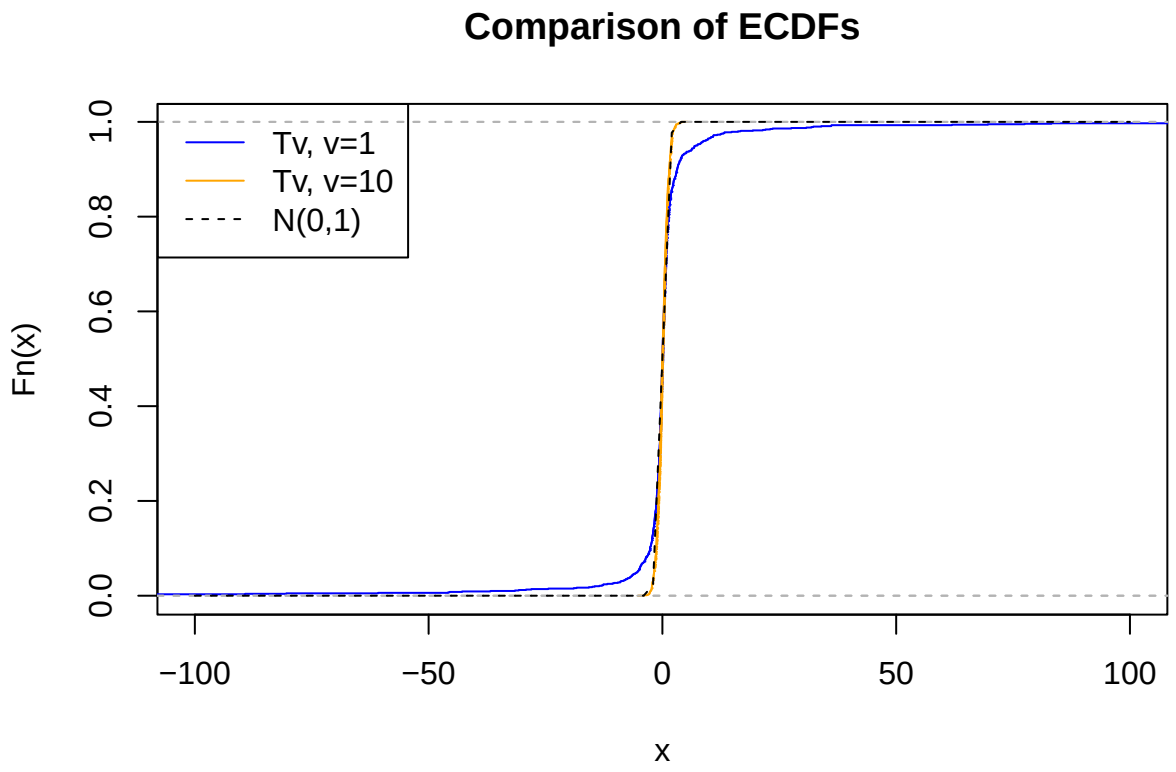
```
     xlim = c(-100,100))
hist(t2,
     main = "v = 10",
     freq = FALSE,
     breaks = 100,
     col = "darkgreen")
```



The comparison between the histograms of the data generated with the two Student-t distributions confirms the results obtained before. In particular, the first distribution with $v = 1$ has a much wider range of variation than the second one with $v = 10$, 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. Next, we also represent the empirical cumulative distribution function (ECDF) of both sets of data. 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.

```
plot(ecdf(t1),
     do.points = FALSE,
     col = "blue",
     xlim = c(-100,100),
     main = "Comparison of ECDFs")
plot(ecdf(t2),
     do.points = FALSE,
     col = "orange",
     add = TRUE)
```

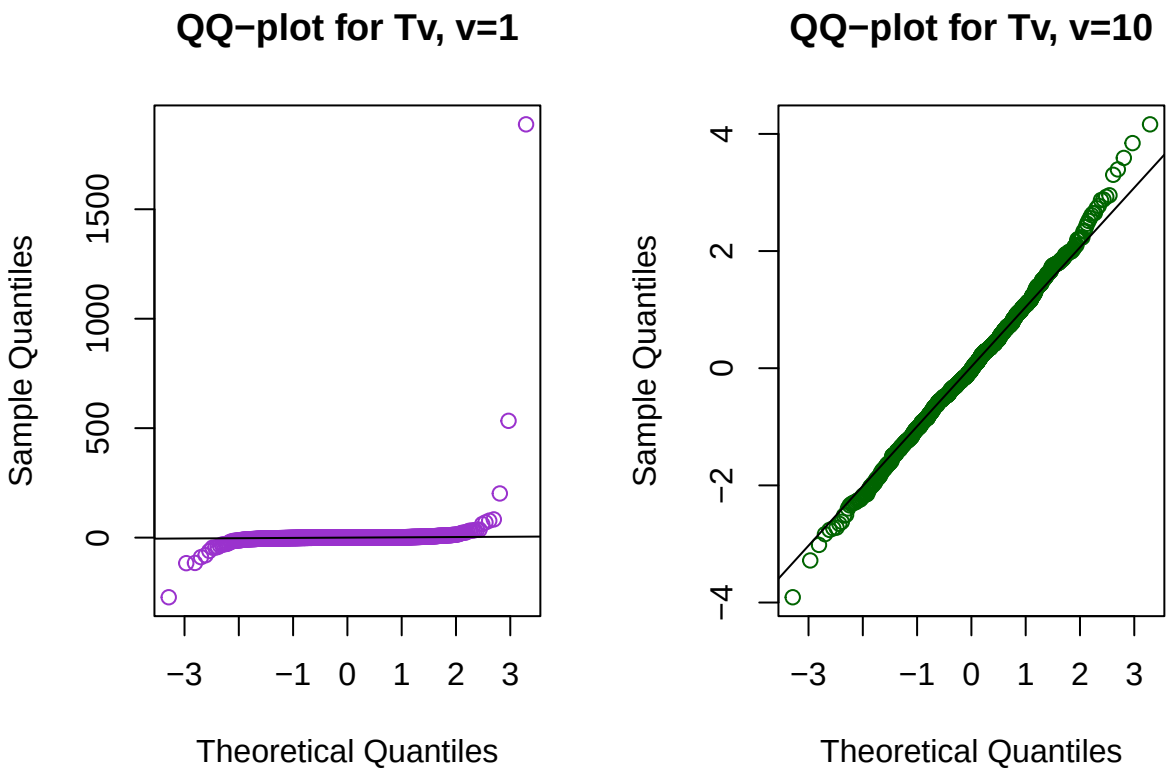
```
curve(pnorm(x), lty = 2, add = TRUE)
legend("topleft",
       c("Tv, v=1", "Tv, v=10", "N(0,1)"),
       lty = c(1,1,2),
       col = c("blue", "orange", "black"))
```



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.

Finally, we depict the QQ-plot of the two distributions.

```
par(mfrow = c(1,2))
qqnorm(t1, col = "darkorchid3", main = "QQ-plot for Tv, v=1")
qqline(t1)

qqnorm(t2, col = "darkgreen", main = "QQ-plot for Tv, v=10")
qqline(t2)
```



The first QQ-plot ($v = 1$) shows that the corresponding distribution is not normal, having heavier tails (both left and right) than a standard Gaussian distribution (points in the right tail above the straight line and points in the left tail below the straight line). The second QQ-plot ($v = 10$), instead, shows that the corresponding distribution is close to normality (points on both the left and the right line close to the straight line); indeed it is known that for $v \to +\infty$, the Student-t distribution converges to a standard Gaussian one.

2.2 Probability calculations and comparison across Gaussian and Student-t distributions using cumulative distribution functions.

We compute the require probabilities through the `pnorm()` (for the Gaussian distribution) and `pt()` (for the Student-t distribution) functions.

- $P(X < 7)$ and $P(3 < X < 6)$ when X $N(5, 4)$

```
pnorm(7, mean = 5, sd = sqrt(4))
```

```
## [1] 0.8413447
```

```
pnorm(6, mean = 5, sd = sqrt(4)) - pnorm(3, mean = 5, sd = sqrt(4))
```

```
## [1] 0.5328072
```

- $P(-1 < X < 1)$, $P(-2 < X < 2)$ and $P(-3 < X < 3)$ when X $N(0, 1)$

```r
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

```r
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

```r
pnorm(3) - pnorm(-3)
```

```
## [1] 0.9973002
```

The last result, in particular, empirically confirms the fact that over 99% of realizations from a univariate Gaussian random variable fall within the interval $[\mu - 3\sigma, \mu + 3\sigma]$.

- $P(-1 < X < 1)$ when X T_1.

```r
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.

- $P(-1 < X < 1)$ when X T_{10}

```r
pt(1, df=10) - pt(-1, df=10)
```

```
## [1] 0.6591069
```

Note that this value is instead quite similar to the corresponding one obtained using the standard Gaussian distribution.

2.3 Bivariate Gaussian simulation with zero correlation: scatter plots, Q-Q plots, 3D and contour density plots, and spectral decomposition.

Consider the bivariate Gaussian distribution of (X_1, X_2).

2.3.0.1 8.1 Generate 3000 realizations from such distribution with the following parameter values: $\mu_{x1} = \mu x_2 = 0$, $\sigma_{x1} = \sigma_{x2} = 3$, $\sigma_{x1x2} = 0$. Set the seed to the value of 625. Comment on the generated data data.

We use the `rmvnorm()` function to generate realizations from the bivariate Gaussian distribution. This function requires the number of desired realizations (`n`), the mean vector (`mean`) and the variance covariance marix (`sigma`). Note that we set the seed value to ensure reproducibility.

```r
require(mvtnorm)
```

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

```r
set.seed(625)
mu <- c(0,0)
sigma <- matrix(c(9,0,0,9), ncol = 2)
X <- rmvnorm(n = 3000, mean = mu, sigma = sigma)

summary(X)
```

```
##        V1                   V2
##  Min.   :-9.440901   Min.   :-11.23894
##  1st Qu.:-1.937981   1st Qu.: -1.96090
##  Median :-0.007103   Median : -0.01647
##  Mean   : 0.032120   Mean   :  0.02925
##  3rd Qu.: 1.993745   3rd Qu.:  1.98664
##  Max.   :10.665299   Max.   : 12.91324
```

The object returned in the output by the 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, since the covariance is 0, both components have the same distribution: (X, Y) $N(0, 9)$. 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.

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

```
## [1] 2.927782
```

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

```
## [1] 2.978838
```

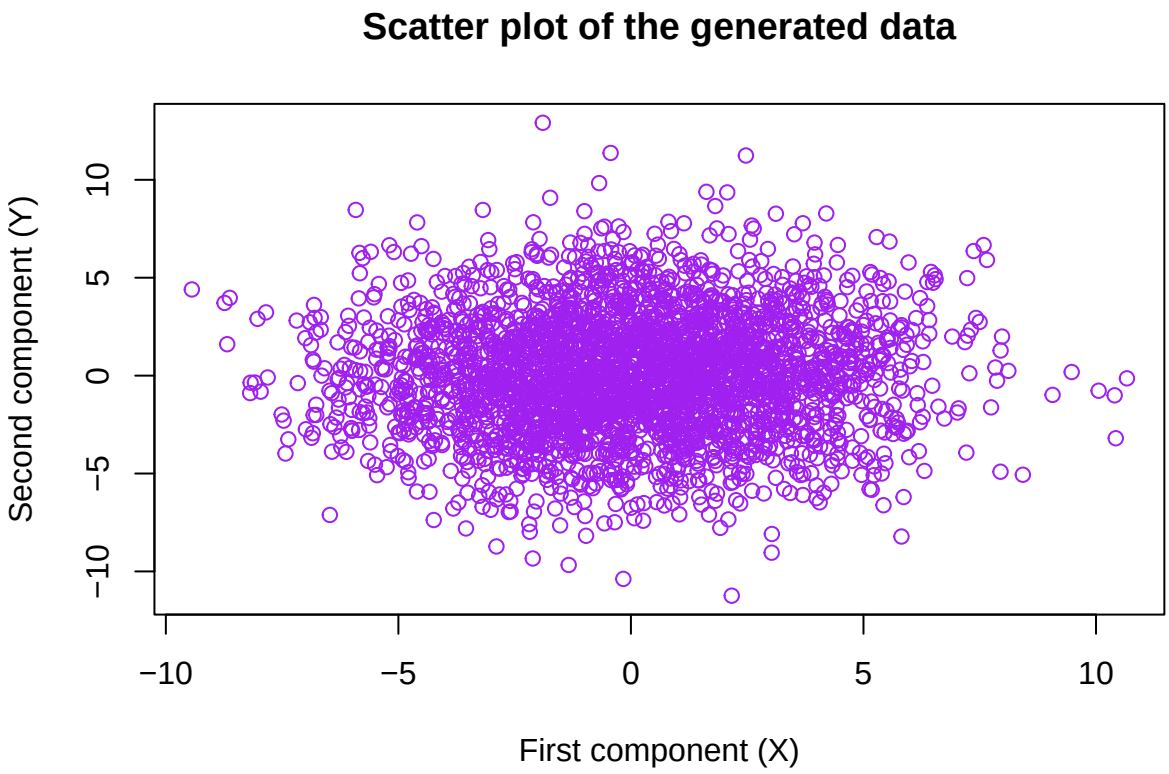
```r
cov(X[,1], X[,2])
```

```
## [1] 0.3353544
```

The standard deviations of the two components are both very close to the assigned theoretical value of 3; 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 0.34) is slightly greater than the theoretical one. The two components are therefore (very weakly) linearly and negatively associated with each other.

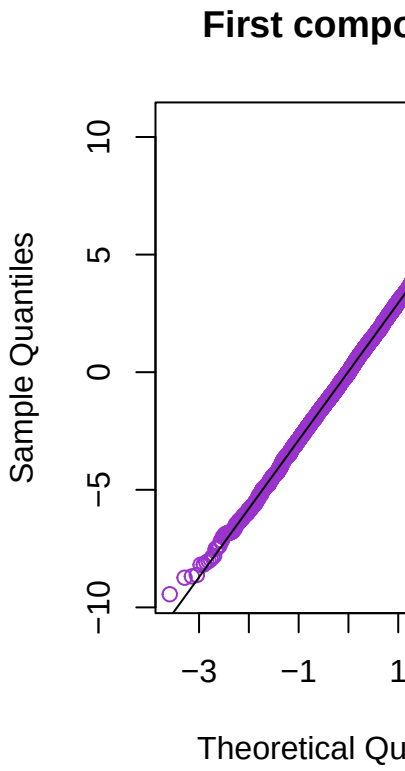
2.3.0.2 8.2 Draw the scatter plot of the realizations and comment on it. We draw the scatter plot of the generated data, representing the two univariate components on the two axes.

```r
plot(X[,1], X[,2],
     col = "purple",
     main = "Scatter plot of the generated data",
     xlab = "First component (X)",
     ylab = "Second component (Y)")
```



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 (also circular, in this specific case) without showing any well-defined orientation. This behavior indicates that the covariance is equal to zero.

```r
par(mfrow = c(1,2))
qqnorm(X[,1], main = "First component", col = "darkorchid3")
qqline(X[,1])

qqnorm(X[,2], main = "Second component", col = "darkgreen")
qqline(X[,2])
```



2.3.0.3 8.3 Depict the quantile-quantile plot for both and comment on them.

In both cases, the curve almost perfectly overlaps with the bisector of the first and third quadrants, indicating that the empirical quantiles and the theoretical quantiles (from a standard Gaussian distribution) are almost identical. Only for extreme values do they deviate slightly; for example, the first component has a slightly heavier right tail and a slightly lighter left tail.

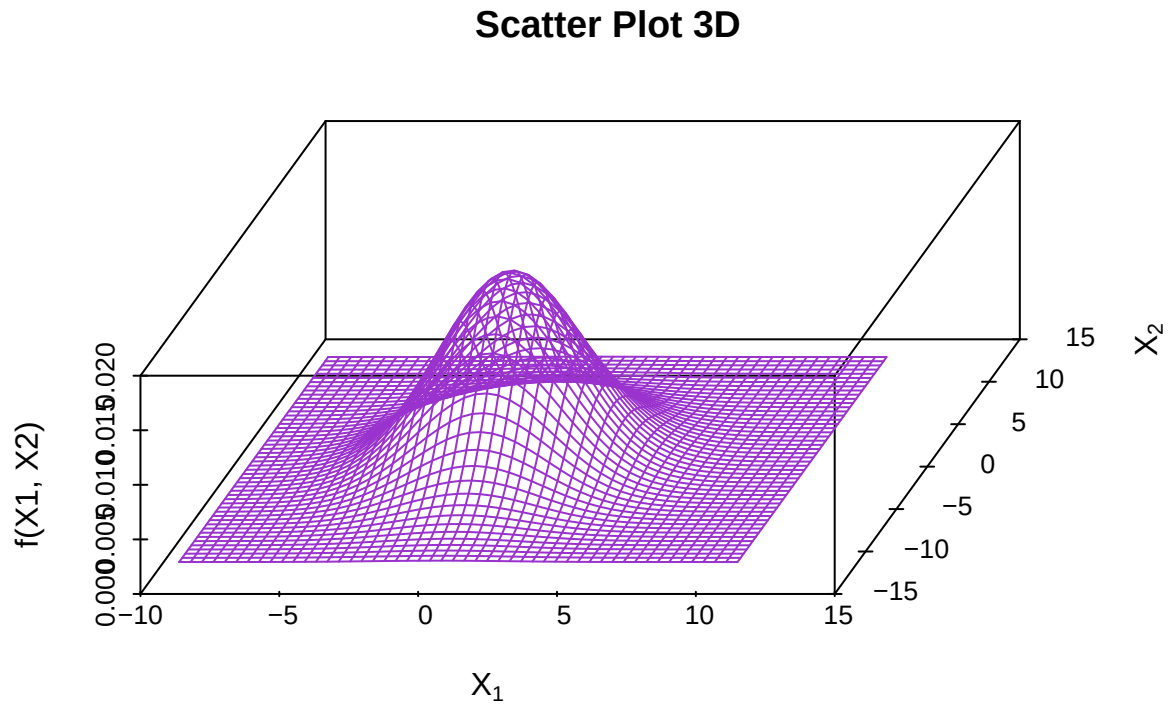
```
require(scatterplot3d)
```

2.3.0.4 8.4 Calculate the theoretical density for some values of X_1 and X_2. Depict the density plot in three dimensions. Comment on the shape of the distribution. What can we observe?

```
## Loading required package: scatterplot3d
```

```
z1 <- seq(min(X[, 1]), max(X[, 1]), length.out = 50)
z2 <- seq(min(X[, 2]), max(X[, 2]), length.out = 50)
grid <- expand.grid(z1, z2)
dens <- matrix(dmvnorm(grid, mean = mu, sigma = sigma),
               ncol = length(z1), byrow = TRUE)

plot3d <- scatterplot3d(z1, z2, seq(min(dens), max(dens),
                                    length.out = length(z1)),
                        xlim = c(min(X[, 1]), max(X[, 1])),
                        ylim = c(min(X[, 2]), max(X[, 2])),
                        type = "n", angle = 70, grid = FALSE,
```

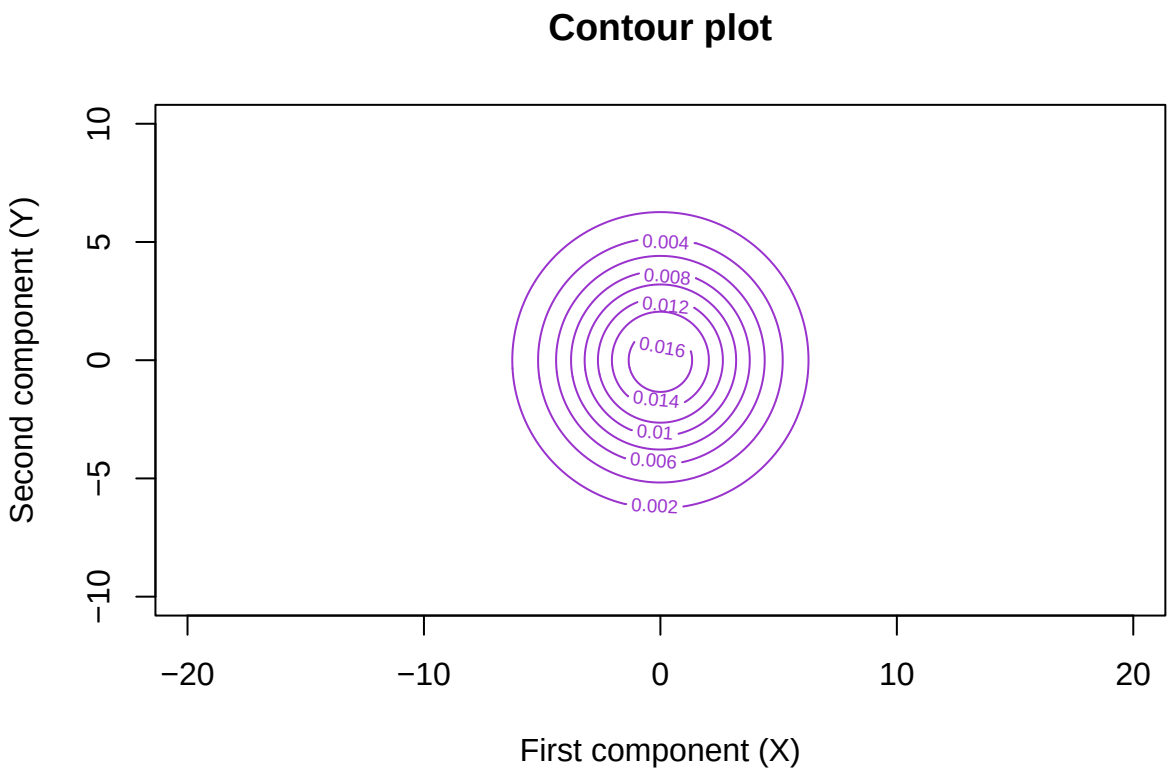
```
                        main = "Scatter Plot 3D",
                        xlab = expression(X[1]),
                        ylab = expression(X[2]),
                        zlab = "f(X1, X2)")

for (i in length(z1):1) {
plot3d$points3d(rep(z1[i], length(z2)), z2, dens[i, ],
type = "l", col = "darkorchid3")
}
for (i in length(z2):1) {
plot3d$points3d(z1, rep(z2[i], length(z1)), dens[, i],
type = "l", col = "darkorchid3")
}
```



2.3.0.5 8.5 Draw and interpret the contour plots obtained from the theoretical density calculated at the previous point. The contour plot shows the shape of the density function of the distribution; the contour lines are obtained by intersecting the three-dimensional surface of the bivariate Gaussian distribution with planes parallel to the base and placed at different heights.

```
z1 <- seq(-10, 10, length.out = 300)
z2 <- seq(-10, 10, length.out = 300)
griglia <- expand.grid(z1, z2)
dens <- matrix(dmvnorm(griglia, mean = mu, sigma = sigma), ncol = length(z1))
contour(z1, z2, dens, col = "darkorchid3",
```

```
        main = "Contour plot",
        xlab = "First component (X)",
        ylab = "Second component (Y)",
        asp = 1)
```



We observe in this plot the perfectly circular 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 and - the value equal to zero of the correlation between the two components. The center of the circle (intended as the intersection point of the two axes) is located at the mean of the distribution.

2.3.0.6 8.6 Extract the eigenvalues and eigenvectors of the variance-covariance matrix. Show the spectral decomposition. We first compute the eigenvalues and eigenvectors of the variance-covariance matrix.

```
eig <- eigen(sigma)
Eig.Val <- eig$val; Eig.Val
```

```
## [1] 9 9
```

```
Eig.Vec <- eig$vec; Eig.Vec
```

```
##      [,1] [,2]
## [1,]    0   -1
## [2,]    1    0
```

Eigenvalues and eigenvectors represent the direction and magnitude of the "rescaled" (principal) components of the bivariate variable. The spectral decomposition is obtained through the following matrix product.

```
Eig.Vec %*% diag(Eig.Val) %*% t(Eig.Vec)
```

```
##      [,1] [,2]
## [1,]    9    0
## [2,]    0    9
```

2.4 Bivariate Gaussian simulation with weak negative correlation: data comparison with previous case using plots, summary statistics, and covariance structure.

Consider a bivariate Gaussian distribution for (x_1, x_2) with the following parameter values: $\mu_{x_1} = \mu_{x_2} = 0$, $\sigma_{x_1} = \sigma_{x_2} = 10$, and $\sigma_{x_1 x_2} = -2$. Repeat the points of the previous exercise and compare the results obtained here with those previously observed.

We simulate a bivariate normal sample with the specified parameters.

```
set.seed(625)
mu9 <- c(0, 0)
sigma9 <- matrix(c(100, -2, -2, 100), ncol = 2)
Y9 <- rmvnorm(n = 3000, mean = mu9, sigma = sigma9)
```

We begin by summarizing the simulated data.

```
summary(Y9)
```

```
##        V1                  V2
##  Min.   :-31.61481   Min.   :-37.53351
##  1st Qu.: -6.45309   1st Qu.: -6.48500
##  Median : -0.05282   Median : -0.05588
##  Mean   :  0.10609   Mean   :  0.09642
##  3rd Qu.:  6.64777   3rd Qu.:  6.56492
##  Max.   : 35.55386   Max.   : 43.10512
```

We compute the standard deviations of the two components and their covariance.

```
sd(Y9[, 1])  # Standard deviation of x1
```

```
## [1] 9.755473
```

```
sd(Y9[, 2])  # Standard deviation of x2
```

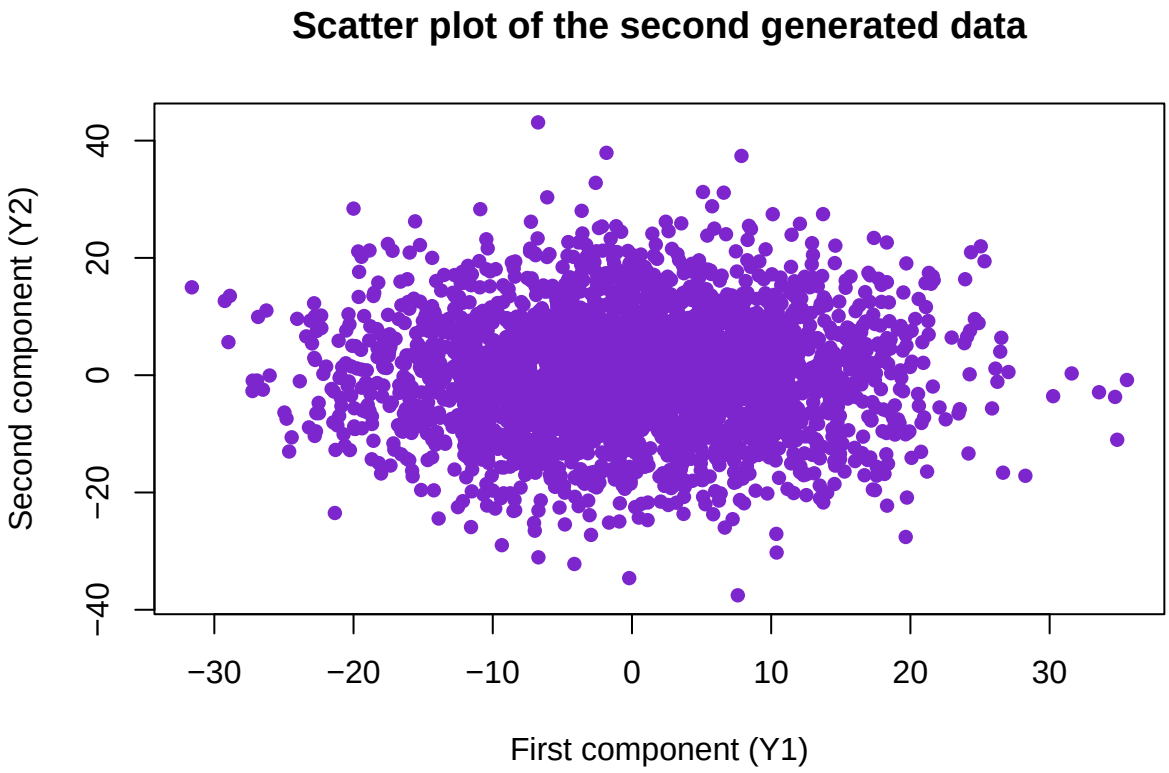
```
## [1] 9.925689
```

```
cov(Y9[, 1], Y9[, 2])  # Covariance between x1 and x2
```

```
## [1] 1.787784
```

We visualize the relationship between the two components using a scatter plot.

```r
plot(Y9[, 1], Y9[, 2],
     col = "purple3",
     pch = 16,
     main = "Scatter plot of the second generated data",
     xlab = "First component (Y1)",
     ylab = "Second component (Y2)")
```



- As expected, both variables have a sample standard deviation close to the theoretical value of 10, given the variance components $\sigma^2 = 100$.
- The covariance is approximately -2, matching the imposed structure of the covariance matrix. Since the covariance is small and negative, the correlation between the two variables is also negative but weak. The correlation can be computed as:

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{-2}{10 \cdot 10} = -0.02$$

- This weak negative correlation is reflected in the **scatter plot**, which appears nearly circular and symmetric, with no evident linear trend.
- Compared to the previous exercise (likely Exercise 8), where the correlation was higher in magnitude (either positive or negative), the current distribution shows less directional association between the variables. The points are more diffusely spread around the origin, indicating **near-independence**.
- Overall, the shape of the cloud confirms that the variables are **almost uncorrelated**, and the dependence structure is minimal. This contrasts with previous settings where strong correlation induced elongated elliptical patterns.

2.5 Open questions

- **Describe the bivariate Gaussian distribution and list its properties.**

The **bivariate Gaussian distribution** is the joint distribution of two continuous random variables X and Y that are normally distributed and may be correlated.

Its **probability density function (PDF)** is:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}Q)$$

with

$$Q = (\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2})$$

Its **main properties** are: - Marginals X $N(\mu_X, \sigma_X^2)$, Y $N(\mu_Y, \sigma_Y^2)$ are univariate normal. - Joint distribution is fully defined by the means, variances and correlation ρ. - If $\rho = 0$, the variables are **uncorrelated** (but not necessarily independent). - If $\rho = 0$ and the distribution is **bivariate normal**, then X and Y are **independent**.

- **Describe the variance-covariance and the correlation matrix.**

The **variance-covariance matrix** (Σ) of a vector of random variables $X = (X_1, X_2, \dots, X_p)$ is:

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

The **correlation matrix** contains the Pearson correlation coefficients:

$$R_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

The main differences consist in the fact that covariance is in units of the variables, while **correlation is dimensionless** and bounded between -1 and 1. Also, covariance expresses **joint variability**; correlation expresses **strength and direction** of linear relationship.

- **Explain the difference between the raw and partial correlation coefficients.**

The **raw (Pearson) correlation** between X and Y measures their direct linear relationship, ignoring other variables.

The **partial correlation** between X and Y controls for the effect of other variables (e.g., Z):

$$\rho_{XY \cdot Z} = \text{correlation between residuals of X and Y after regressing on Z}$$

In practice: partial correlation answers *"Is there still a relationship between X and Y once we remove the effect of Z"?*

- **What we mean when we speak of a spurious association between two random variables?** A **spurious association** occurs when two variables appear to be correlated, but the relationship is actually due to:

- A **confounding variable** (common cause),
- **Coincidence**, or
- **Indirect dependence** (e.g., throygh a third variable).

For example, ice cream sales and drowning deaths may be correlated due to temperature (a confounder), not because one causes the other.

- **If two random variables having a bivariate Gaussian distribution have zero correlation they are also independent? If yes, why?**

Yes, if two random variables X and Y have **bivariate Gaussian distribution** and **zero correlation** ($\rho = 0$), they are also **independent**. This is true because, in the multivariate normal distribution, **zero covariance implies independence**. This is not true in general, but it is **true under the Gaussian assumption**.

- **What is a contour plot? Describe why we observe ellipse in the case of bivariate random variables with a Gaussian distribution.**

A contour plot is a 2D plot showing **level curves** of a 3D surface — for bivariate distributions, it shows lines of **equal probability density**. In the case of a **bivariate normal**, the contours are **ellipses** because: - The density function has a quadratic form in the exponent. - The ellipses represent regions of equal Mahalanobis distance from the mean.

If the variables are uncorrelated, the ellipses are axis-aligned circles (in standardized form); if correlated, they are tilted ellipses.

- **Explain the spectral decomposition.**

Spectral decomposition (also called **eigendecomposition**) is a way to express a **symmetric matrix** A (like a covariance matrix) as:

$$A = Q\Lambda Q^\top$$

where: - Q is an orthogonal matrix of **eigenvectors**, - Λ is a diagonal matrix of **eigenvalues**.

In statistics, the spectral decomposition is used in **Principal Component Analysis (PCA)**, helps understand **directions of maximal variance** and allows **dimensionality reduction**.

In the case of the covariance matrix, the eigenvectors define the directions of the ellipses (principal components), and eigenvalues define their "stretch".

3 Week 3 - Nonparametric bootstrap for estimating standard error and confidence intervals, Bootstrap inference on mean, median, and standard deviation, Graphical interpretation of bootstrap distributions, Properties of estimators and introduction to maximum likelihood estimation (MLE)

3.1 Nonparametric bootstrap to assess mean estimate accuracy: descriptive stats, ECDF, bootstrap distribution, and interpretation.

Consider the following simulated data from an athlete's training session for a 200-meter sprint (see the file `times.RData`). The observations are related to the time required by each one to complete the sprint (in seconds).

```
load("times.RData")
skim_without_charts(cranio1)
```

3.1.0.1 10.1 Illustrate the data through descriptive statistics. Depict the empirical cumulative distribution function.

Table 5: Data summary

Name	cranio1
Number of rows	98
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

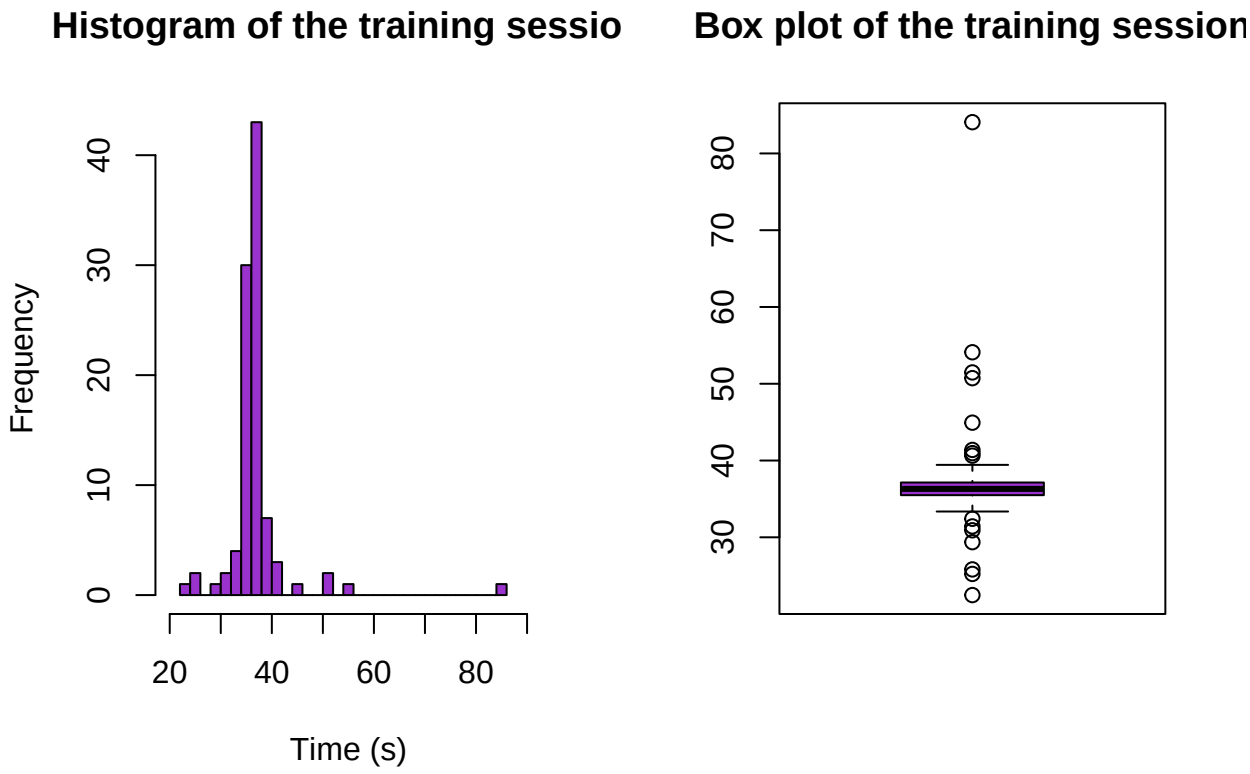
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
data	0	1	36.94	6.26	22.48	35.5	36.31	37.13	84.07

The dataset `cranio1` contains sprint times in seconds recorded during a training session. From the numerical summary, we observe that the **mean and median are nearly equal**, indicating **approximate symmetry** of the central part of the distribution. The **interquartile range (IQR)** is relatively small, showing that most sprint times are concentrated within a narrow window, approximately between 35.5 and 37.1 seconds.

However, the **maximum value** of 84 seconds is far from the upper quartile and substantially increases the overall range (61.6 seconds). This suggests the presence of at least one **extreme observation**, possibly corresponding to an unsuccessful sprint or an anomalous recording.

```r
par(mfrow=c(1,2))
hist(cranio1,
     main = "Histogram of the training session",
     breaks = 30,
     xlim = c(20, 90),
     col = "darkorchid3",
     xlab = "Time (s)")
boxplot(cranio1,
        main = "Box plot of the training session",
        horizontal = F,
        col = "darkorchid3")
```

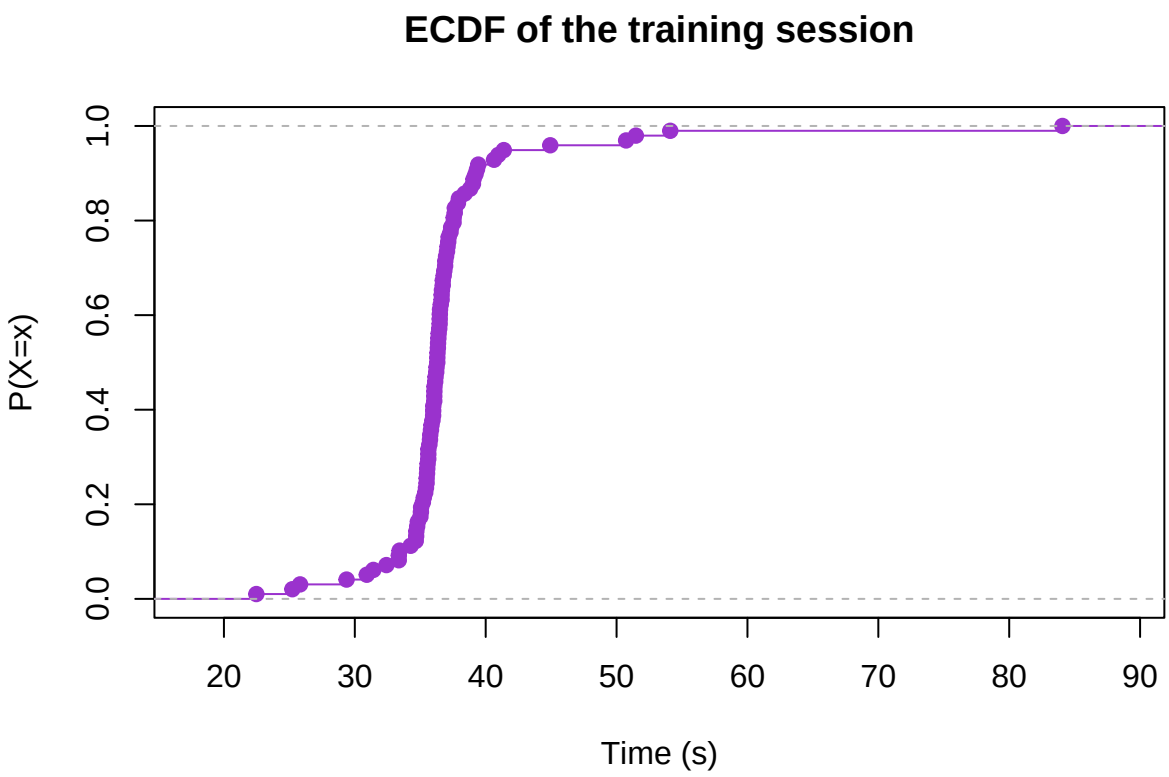


The **histogram** confirms that the data are **highly concentrated around the central mode**, with the bulk of the observations falling between 35 and 38 seconds. The tail of the histogram extends to the right, revealing **positive skewness**. A small number of sprint times are much larger than the rest, which may represent exceptional cases (e.g., injury, fatigue, interruption).

The **boxplot** illustrates the same characteristics: a compact interquartile range and a number of **outliers**, especially on the right. The right whisker is significantly longer than the left, and several points are plotted individually, indicating extreme values. These graphical elements emphasize the skewed structure of the data despite the symmetric appearance of the central mass.

```r
plot(ecdf(cranio1),
     col = "darkorchid3",
     main = "ECDF of the training session",
     xlab = "Time (s)",
```

```
    ylab = "P(X=x)")
```



The **empirical cumulative distribution function (ECDF)** rises sharply between 34 and 38 seconds, reinforcing the earlier observation that most sprint times lie within this narrow interval. The **slow increase in the tails** reveals the presence of a few very fast and very slow observations. Specifically:

- The **left tail** shows a gradual increase, indicating a few **extremely fast** performances (potentially outliers).
- The **right tail** flattens more noticeably, reflecting the presence of **exceptionally slow sprints**, which are infrequent but strongly impact the range.

Overall, the ECDF confirms the key features already observed in the histogram and boxplot: a **central core of consistent sprint times**, coupled with **skewed tails** and some **extreme outliers**, particularly in the upper range. These characteristics may warrant further investigation (e.g., robustness analysis or trimming) to determine whether outliers should be considered measurement errors, exceptional conditions, or genuine part of performance variability.

3.1.0.2 10.2 Consider as parameter of interest the arithmetic mean, use the nonparametric bootstrap to determine its measure of accuracy with B = 1000 bootstrap samples. Report the result and comment on it.

The goal is to estimate the **sampling variability** (i.e., standard error) of the arithmetic mean of the 200-meter sprint times using the **nonparametric bootstrap** technique. This resampling-based method is particularly suitable in this context, as the presence of **outliers** in the data could affect analytical standard error calculations, especially under deviations from normality.

We first implement the bootstrap manually using a `for` loop and the `sample()` function, generating $B = 1000$ bootstrap samples of size equal to the original sample (`n10`), with replacement.

```r
B10 <- 1000
n10 <- length(cranio1)
Tboot10 <- rep(0, B10)
set.seed(16253)
for (i in 1:B10) {
  Xstar10 <- sample(cranio1,
                    n10,
                    replace = TRUE)
  Tboot10[i] <- mean(Xstar10)
}

round(sd(Tboot10), 3)
```

```
## [1] 0.637
```

The standard deviation of the bootstrap estimates (`Tboot10`) serves as the **bootstrap estimate of the standard error** of the sample mean. The result is approximately **0.63 seconds**.

We now replicate the procedure using the function `bootstrap()` from the `bootstrap` package, which streamlines the computation.

```r
require(bootstrap)
```

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

```r
boot10 <- bootstrap(cranio1,
                    nboot = 1000,
                    theta = mean)

round(sd(boot10$thetastar), 3)
```

```
## [1] 0.628
```

The two implementations yield **almost identical results**, as expected. The small numerical differences are attributable to internal algorithmic details or random variation in resampling.

The resulting **bootstrap standard error is around 0.63**, which is **very small relative to the overall variability** of the dataset (standard deviation of the original sprint times 6.26 seconds). This confirms that:

- Despite the presence of **outliers**, the sample mean remains a **stable and reliable estimator** of central tendency in this context.
- The **accuracy of the mean is high**, meaning that even under resampling perturbations, the estimate does not fluctuate substantially.

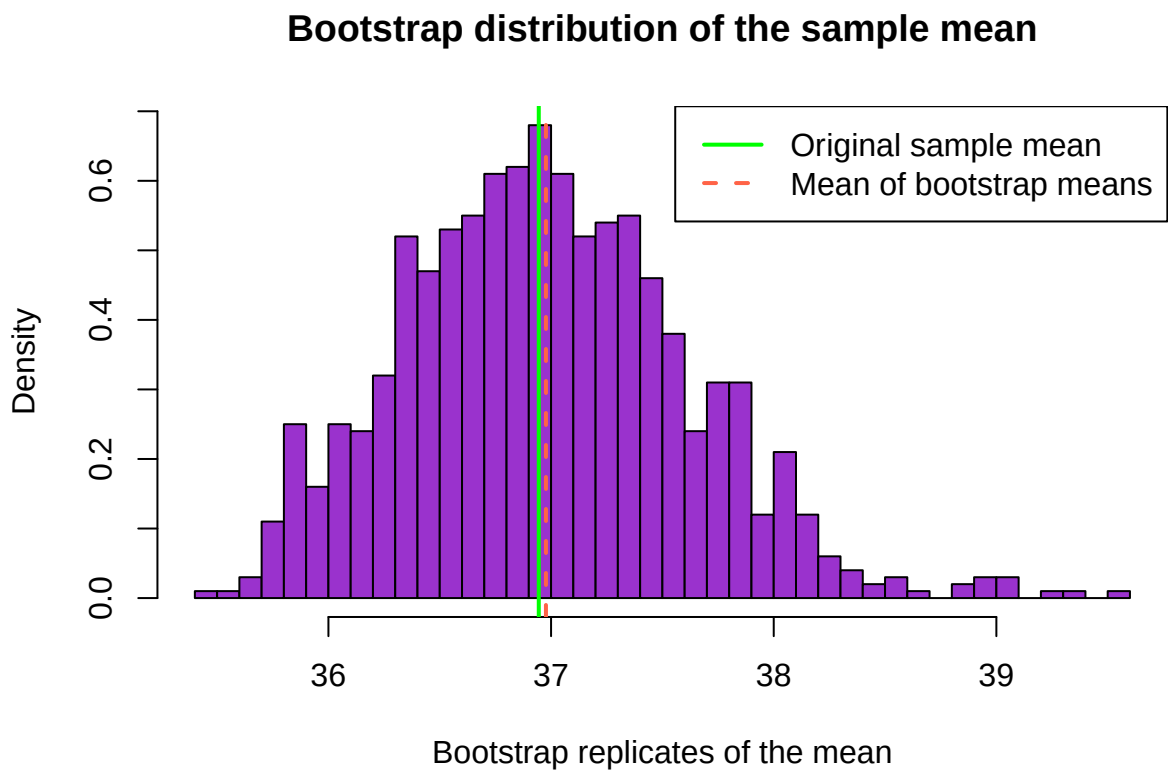
In conclusion, the use of nonparametric bootstrap provides empirical evidence that the average sprint time is estimated with **high precision**, and the effect of occasional extreme values is limited due to the mean's robustness in this sample configuration.

3.1.0.3 10.3 Depict and comment on the bootstrap distribution, include also the point estimate of the parameter calculated from the original sample.

We use the 1000 bootstrap replicates of the arithmetic mean of the 200-meter sprint times to visualize the distribution of the estimator and compare it to the point estimate from the original sample.

```
p_est <- mean(cranio1) # original mean (point estimate)
```

```
hist(
  Tboot10,
  breaks = 30,
  probability = T,
  col = "darkorchid3",
  main = "Bootstrap distribution of the sample mean",
  xlab = "Bootstrap replicates of the mean")

abline(v = p_est, col = "green", lwd = 2) # original point estimate
abline(v = mean(Tboot10), col = "tomato", lwd = 2, lty = 2) # bootstrap mean

legend("topright",
       legend = c("Original sample mean", "Mean of bootstrap means"),
       col = c("green", "tomato"),
       lty = c(1, 2),
       lwd = 2)
```



The histogram shows the bootstrap distribution of the sample mean, based on 1000 resampled datasets. The shape of the distribution is **approximately normal (bell-shaped)** and centered close to the original

sample mean. Also, the distribution is narrow, indicating that the sample mean is a precise estimator.

The green vertical line represents the point estimate (original mean) computed from the observed data. The dashed red line shows the mean of the bootstrap replicates, which closely aligns with the red line, confirming that the estimator is unbiased.

The bootstrap distribution confirms that the sample mean is a stable and reliable estimate of the true average sprint time. Its small spread visually supports the bootstrap standard error previously computed (≈ 0.63), and the alignment of the original and bootstrap means suggests low bias.

3.2 Bootstrap analysis for comparing two groups: difference of means and medians, standard errors, confidence intervals, and distribution visualizations.

Consider the following data, measuring the processing times (in milliseconds) for 8 files uploaded to two different servers:

- Processing time (in ms) with Server A: 176, 125, 152, 180, 159, 168, 160, 151;
- Processing time (in ms) with Server B: 164, 121, 137, 169, 144, 145, 156, 139.

3.2.0.1 11.1 Illustrate the data using appropriate descriptive statistics and graphical representations.

We briefly analyze the data through the main descriptive statistics.

```
ServerA <- c(176, 125, 152, 180, 159, 168, 160, 151)
ServerB <- c(164, 121, 137, 169, 144, 145, 156, 139)

summary(ServerA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   125.0   151.8   159.5   158.9   170.0   180.0
```

```
summary(ServerB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   121.0   138.5   144.5   146.9   158.0   169.0
```

The numerical summaries reveal that **server B is consistently faster** than server A. The **mean processing time** with Server A is higher than that of Server B by approximately 12 milliseconds, indicating a general performance advantage for Server B.

Moreover, the **median** values are 160 ms for Server A and 145 ms for Server B, confirming that the central tendency of Server B is lower. Observing the quartiles, we note that:

- For Server A, the third quartile (Q3) is 170 ms, meaning that **25% of files take more than 170 ms** to be processed.
- For Server B, the maximum value observed is **169 ms**, so **no file takes longer than 170 ms** with this server.

Thus, Server B consistently achieves better performance across the entire distribution.

```r
sd(ServerA)
```

```
## [1] 17.24146
```

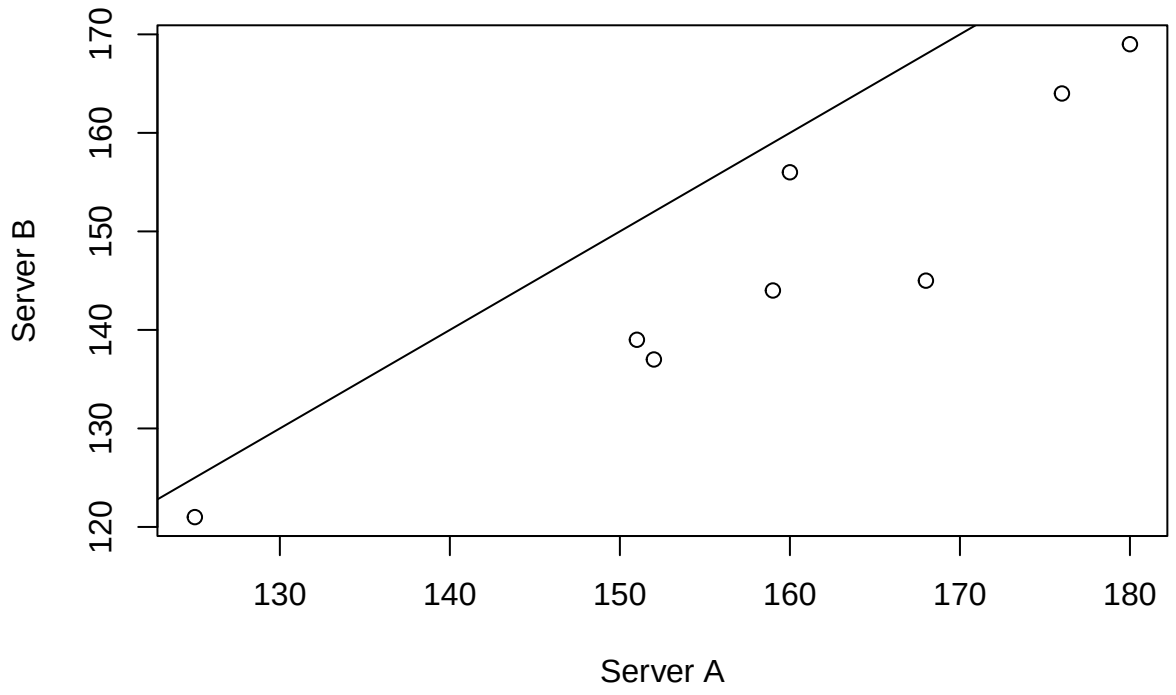
```r
sd(ServerB)
```

```
## [1] 15.61535
```

The **standard deviation** of Server A is slightly higher than that of Server B, indicating **greater variability** in its performance. However, since the sample size is small (only 8 observations per server), caution is needed when generalizing these results to the broader population.

We now examine the relationship between the two servers using a **scatterplot**, where each point represents the processing time of the same file on both servers. The **bisector** (line with slope = 1) is added as a reference: points below it indicate that Server B was faster for that file.

```r
plot(ServerA, ServerB,
     xlab = "Server A",
     ylab = "Server B")

abline(a=0, b=1)
```



All the points lie **strictly below the bisector**, meaning that **for every file**, the processing time on Server B was **lower than or equal to** that on Server A. This **uniform dominance** of Server B across all observations reinforces the numerical findings and suggests a **systematic performance advantage**.

In particular, one file shows a marked reduction in processing time with Server B compared to Server A, appearing as a point far from the bisector. This case may reflect an **outlier in performance gap**, possibly due to system-specific optimizations or transient slowdowns on Server A.

In conclusion, both numerical and graphical evidence support the idea that Server B provides **faster and slightly more consistent** processing times than Server A, at least on the basis of the files analyzed.

3.2.0.2 11.2 Consider as parameter of interest the difference between the two arithmetic means. Apply the bootstrap method using a number of bootstrap replications equal to 1000 and using the bootstrap function to obtain the standard error for this difference. Comment on the result.

As previously observed, the difference between the average processing times with the two servers is 12 ms. The estimate of the standard error (accuracy measure) for this value is obtained using non-parametric bootstrap by using the `bootstrap()` function of the eponymous package. This requires as input the name of the dataset, the number of replications, and the function to be calculated for each replication. In this case, 1000 replications of the original dataset are considered. We set the seed value to 130 to ensure reproducibility.

```
require(bootstrap)
set.seed(130)
n11 <- length(ServerA)
aux_fun <- function(ind) {
  ServerA.boot <- ServerA[ind]
  ServerB.boot <- ServerB[ind]
  mean(ServerA.boot) - mean(ServerB.boot)
}
mean.boot11 <- bootstrap::bootstrap(x = 1:n11,
                                    nboot = 1000,
                                    theta = aux_fun)
summary(mean.boot11$thetastar)
```

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

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.

```
sd(mean.boot11$thetastar)
```

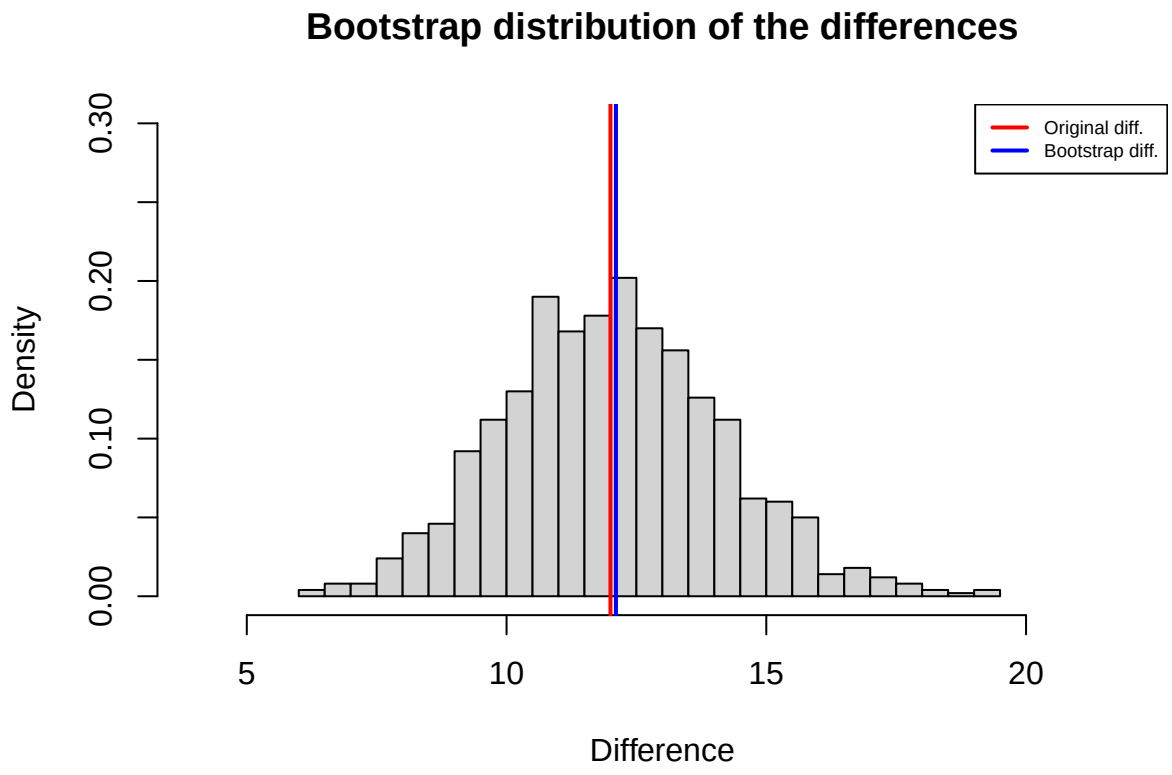
```
## [1] 2.118793
```

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.

3.2.0.3 11.3 Draw the bootstrap distribution and comment on it.

We graphically represent the bootstrap distribution as a histogram. We also add the lines corresponding to the difference calculated on the original dataset and the mean of the differences calculated on the bootstrap datasets.

```r
Mean1 <- mean(ServerA) - mean(ServerB)
Mean2 <- mean(mean.boot11$thetastar)

hist(mean.boot11$thetastar,
     main = "Bootstrap distribution of the differences",
     breaks = 40,
     freq = FALSE,
     ylab = "Density", xlab = "Difference",
     xlim = c(4, 22), ylim = c(0, 0.30))
abline(v = c(Mean1, Mean2),
       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.

3.2.0.4 11.4 Calculate the confidence interval at a confidence level of 95% with the percentile method. Comment on the length of the interval and on the plausible values for the parameter.

The bootstrap confidence interval calculated using the percentile method is based on the empirical distribution of bootstrap replications. The appropriate quantiles of the distribution of bootstrap replications are used as the limits of the confidence interval. Specifically, setting the coverage level to 0.95 determines the endpoints by considering the 2.5th and 97.5th percentiles. The `quantile()` command is used to calculate the specified quantiles.

```
Q11 <- quantile(mean.boot11$thetastar, c(0.025,0.975)); Q11
```

```
##     2.5%    97.5%
##  8.12500 16.25312
```

The 95% confidence interval for the difference is therefore

$$95\%\ \text{CI} = [8.13,\ 16.25]\ \text{milliseconds}$$

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).

3.2.0.5 11.5 Apply the bootstrap method as in point 11.2, considering instead as estimator of interest the difference between the two medians. Depict the bootstrap distribution and comment on the shape.

We repeat the same procedure introduced in point 11.2, but considering instead the difference between the medians.

```
median.orig11 <- median(ServerA) - median(ServerB); median.orig11
```

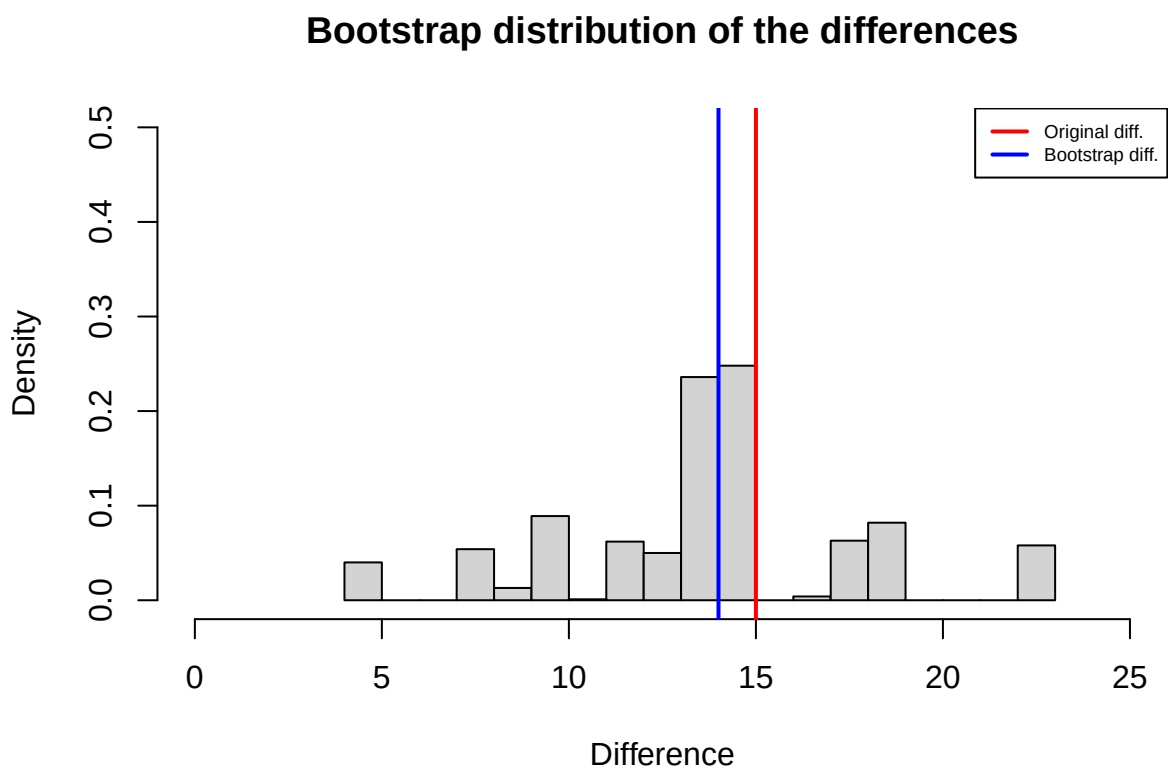
```
## [1] 15
```

```
set.seed(130)
n11 <- length(ServerA)
aux_fun <- function(ind) {
  ServerA.boot <- ServerA[ind]
  ServerB.boot <- ServerB[ind]
  median(ServerA.boot) - median(ServerB.boot)
}
median.boot11 <- bootstrap::bootstrap(x = 1:n11,
                                      nboot = 1000,
                                      theta = aux_fun)

summary(median.boot11$thetastar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   12.00   14.00   13.98   15.00   23.00
```

Analyzing the descriptive statistics, we observe that the difference between the medians of the processing times with servers A and B has a slightly wider range of variation compared to that obtained for the difference of means, ranging from a minimum of 4 to a maximum of 23. Again, the values of mean and median are approximately equal (14 ms), although not coinciding in this case with the value calculated on the original sample (15 ms). The standard deviation is slightly wider than in the previous case, highlighting greater variability for this bootstrap distribution.

```r
Median1 <- median(ServerA) - median(ServerB)
Median2 <- median(median.boot11$thetastar)

hist(median.boot11$thetastar,
     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(Median1, Median2),
       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)
```



In this case, from the histogram of the bootstrap distribution, it is noted that the differences between the medians calculated on the bootstrap samples tend to assume only a small number of values. For example, a high number of differences are observed with a value between 13 and 15 and almost none between 15 and 17. This can be explained by the small number of originally available observations: the bootstrap tends to repeat the observations and therefore also the values of the final result. In this case, with the bootstrap, it is not feasible to estimate some quantiles of the distribution.

3.2.0.6 11.6 Consider the estimator at the previous point (difference between the medians) and provide a confidence interval at a confidence level of 90%. Comment on the result.

We calculate the 90% bootstrap confidence interval calculated using the percentile method.

```
Q11 <- quantile(median.boot11$thetastar, c(0.05,0.95)); Q11
```

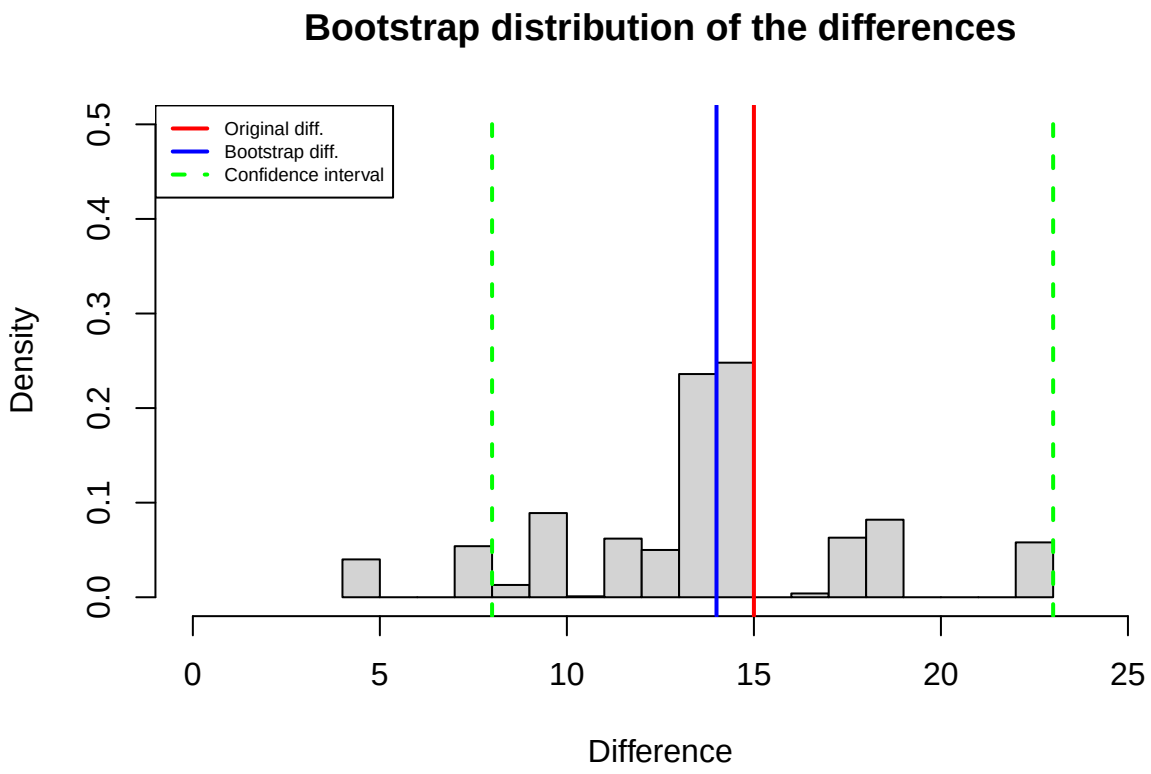
```
##  5% 95%
##   8  23
```

According to the obtained confidence interval, 90% of the bootstrap point estimates lies between 8 and 23. We observe that in this case the confidence interval is quite wide, covering almost completely the range of the 1000 bootstrap values. The difference computed on the original sample lies approximately in the middle of the obtained interval.

3.2.0.7 11.7 Depict the histogram of the bootstrap distribution adding the bars referred to upper and lower bounds of the confidence interval, add also the legend.

We finally add the (dashed) lines representing the confidence interval endpoints to the plot.

```
hist(median.boot11$thetastar,
     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(Median1, Median2, Q11),
       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. 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.

3.3 Bootstrap estimation of standard deviation: descriptive statistics, standard error computation, and annotated histogram of the bootstrap distribution.

Consider the following (simulated) data regarding the daily calorie intake by 20 individuals which are submitted to a special regime:

2700, 2900, 2500, 2800, 2400, 3200, 2000, 2400, 2600, 3000, 3100, 2900, 2700, 2900, 2200, 2800, 2800, 2800, 2700, 2600.

3.3.0.1 12.1 Illustrate the data commenting on the descriptive statistics with respect to the applicative context.

Below are the key descriptive statistics.

```
cals <- c(2700, 2900, 2500, 2800, 2400, 3200, 2000, 2400, 2600, 3000, 3100, 2900, 2700, 2900, 2200, 280

skim_without_charts(cals)
```

Table 7: Data summary

Name	cals

Number of rows	20
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
data	0	1	2700	293.8	2000	2575	2750	2900	3200

The mean daily intake is exactly 2700 kcal, which may represent a target value defined by the special diet plan. The median (2750 kcal) is slightly higher than the mean, suggesting a slight left-skew in the distribution, possibly caused by a few individuals consuming notably lower calories (e.g., the minimum value of 2000 kcal). The interquartile range (IQR = Q3 - Q1 = 2900 - 2575 = 325 kcal) indicates that 50% of individuals consume between 2575 and 2900 kcal, showing moderate variability. The standard deviation ($ $294 kcal) confirms a moderate spread around the mean, which is acceptable in dietary monitoring, considering natural daily variations.

The data suggest good adherence to the regime: the majority of individuals remain within ±300 kcal of the target (2700 kcal). Extreme values (2000 kcal and 3200 kcal) might deserve individual nutritional review, as they represent significant deviations from the average intake and may affect the regimen's effectiveness or safety. From a practical standpoint, this level of variability appears realistic and controlled, but further evaluation (e.g., time trends or metabolic outcomes) would be necessary to assess compliance and effectiveness over time.

```
boot12 <- bootstrap(cals,
                    nboot = 1000,
                    theta = sd)

round(sd(boot12$thetastar), 2)
```

3.3.0.2 12.2 Use the bootstrap method (selecting an appropriate number of replicates) implemented with the function `bootstrap` to obtain the standard error for the estimate of the standard deviation. Comment on the results obtained.

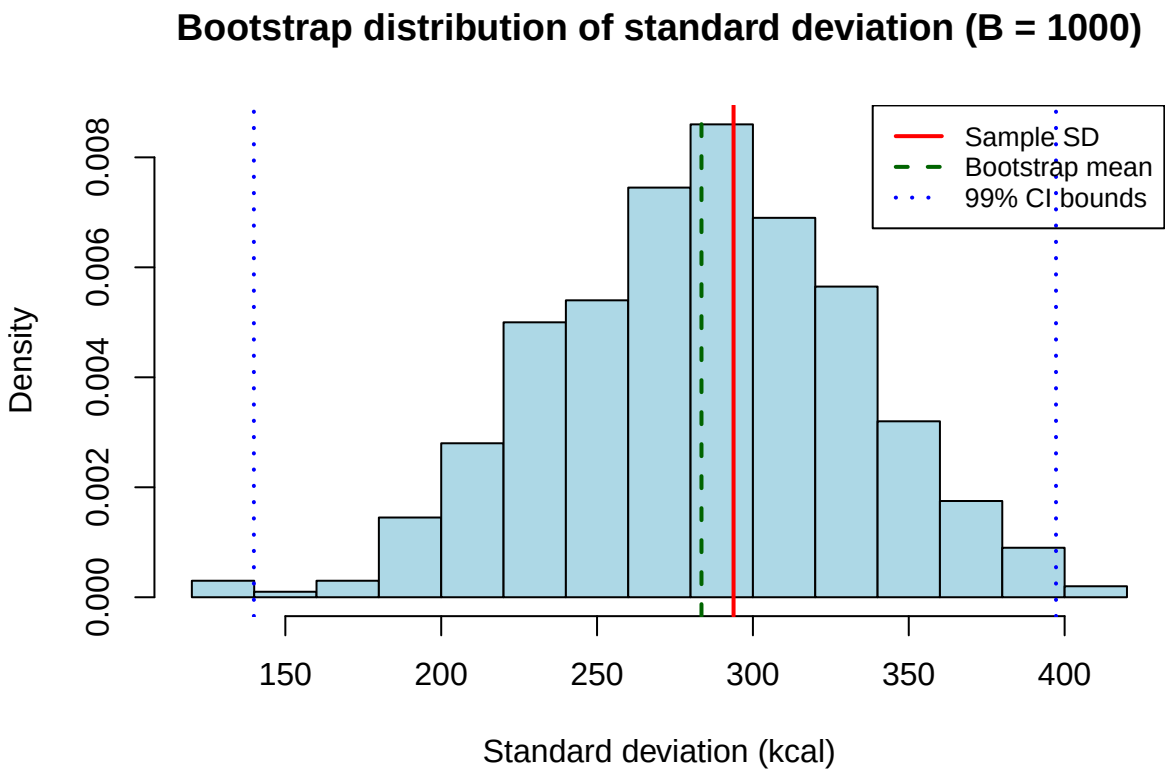
```
## [1] 49
```

The standard deviation of the original sample was approximately 293.8 kcal. The estimated standard error of this value is 50.53 kcal, meaning that if we were to draw 1000 similar samples of 20 individuals, the sample standard deviation would typically vary by about ±50 kcal.

This moderate level of uncertainty reflects the fact that variability (standard deviation) is itself a less stable estimator when the sample size is small (only 20 individuals). However, the estimated SE (50.53 kcal) is much smaller than the value of the standard deviation ($ $294 kcal), which confirms that the estimate is still reasonably precise. In practical terms, this means we can trust that the true variability in daily calorie intake under the special diet is likely close to the observed value, within a margin of error of around 50 kcal.

```r
ci_bounds_99 <- quantile(boot12$thetastar, probs = c(0.005, 0.995))

hist(boot12$thetastar,
     breaks = 20,
     freq = FALSE,
     col = "lightblue",
     main = "Bootstrap distribution of standard deviation (B = 1000)",
     xlab = "Standard deviation (kcal)")

abline(v = sd(cals), col = "red", lwd = 2)

abline(v = mean(boot12$thetastar), col = "darkgreen", lwd = 2, lty = 2)

abline(v = ci_bounds_99[1], col = "blue", lwd = 2, lty = 3)
abline(v = ci_bounds_99[2], col = "blue", lwd = 2, lty = 3)

legend("topright",
       legend = c("Sample SD", "Bootstrap mean", "99% CI bounds"),
       col = c("red", "darkgreen", "blue"),
       lty = c(1, 2, 3),
       lwd = 2,
       cex = 0.8)
```

3.3.0.3 12.3 Report and describe the histogram of the bootstrap distribution, adding the value calculated on the original sample, the mean value on the bootstrap replications, the extremes of the confidence interval obtained with a confidence level of 99% and add the legend.



The histogram shows the bootstrap distribution of the standard deviationvof daily calorie intake, based on 1000 bootstrap samples.

The red vertical line marks the standard deviation computed from the original sample (~293.8 kcal), vwhile the green dashed line represents the mean of the bootstrap replications (approximately 283.5 kcal).

The blue dashed lines indicate the 99% percentile-based confidence interval:

$$99\%\ \text{CI} = [144.72,\ 393.41]\ \text{kcal}$$

The distribution appears unimodal and approximately symmetric,vindicating stability in the bootstrap estimates. The close alignment between the sample SD and the bootstrap mean confirms that the sample estimate is unbiased, while the confidence interval provides a realistic range of plausible values for the true population standard deviation.

3.4 Open questions

- **Lists and describe the properties required for an estimator.** In statistics, an estimator is a rule or function used to estimate an unknown parameter based on sample data. A good estimator should satisfy certain desirable properties that ensure its reliability and interpretability, which are:
  - **Unbiasedness** An estimator θ s unbiased if its expected value equals the true parameter:

    $$\mathrm{E}[\hat{\theta}] = \theta$$

    This means, on average, the estimator does not systematically overestimate or underestimate the parameter.

- **Consistency** An estimator is consistent if it converges in probability to the true value as the sample size increases:
$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \to \infty$$
That is, larger samples yield estimates increasingly close to the real parameter.
- **Efficiency** An estimator is efficient if it has the lowest variance among all unbiased estimators of the same parameter. The most efficient estimator reaches the Cramér-Rao Lower Bound.
- **Sufficiency** An estimator is sufficient if it captures all the information about the parameter present in the data. Formally, a statistic $T(X)$ is sufficient for θ if the conditional distribution of the sample X given $T(X)$ does not depend on θ.
- **Robustness** A robust estimator is less sensitive to outliers or deviations from model assumptions. Example: The median is more robust than the mean under heavy-tailed distributions.

To summarize: a "good" estimator is ideally unbiased, consistent, efficient, and sometimes sufficient and robust, depending on the context. These properties guide the choice of estimators in both theoretical and applied statistical analysis.

- **Describe the likelihood function and the maximum likelihood estimation. Provide an example on how this procedure works**

The likelihood function expresses how likely it is to observe the given data as a function of unknown parameters. In a binomial context, where x successes are observed in n trials, the likelihood function for parameter p is:

$$\mathcal{L}(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Maximum Likelihood Estimation (MLE) is the method of choosing the parameter value that **maximizes the likelihood**.

For example, given $x = 7$ heads in $n = 10$ coin flips:

```
likelihood <- function(p) {
  choose(10, 7) * p^7 * (1-p)^(10-7)
}
```

The MLE is:

$$\hat{p}_{MLE} = \frac{7}{10} = 0.7$$

This means that, based on the data, the best estimate for the probability of heads is 70%. The likelihood is maximized when the parameter p matches the relative frequency of the observed event.

- **What are the properties of the estimator obtained through the maximum likelihood method?** The Maximum Likelihood Estimator (MLE) is a widely used estimation method in statistics due to its **desirable asymptotic properties**. Although MLEs may not always be perfect in small samples, under general regularity conditions, they possess the following key properties:
  - **Consistency** The MLE is consistent, meaning that as the sample size increases, the estimator **converges in probability** to the true value of the parameter.
  - **Asymptotic normality**

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta))$$

The distribution of the MLE approaches a normal distribution centered at the true parameter value as $n \to \infty$, with variance equal to the inverse of the Fisher Information.

- **Asymptotic efficiency** The MLE attains the Cramér-Rao Lower Bound asymptotically, meaning it is the most efficient (lowest variance) among all consistent estimators in large samples.
- **Invariance** If $\hat{\theta}_{MLE}$ is the MLE of θ, and $g(\theta)$ is a function of θ, then:

$$\text{MLE of } g(\theta) = g(\theta_{MLE})$$

The MLE is **invariant under transformations**: the MLE of a function of the parameter is the function applied to the MLE.

- **What we refer to when we deal with the Fisher Information?** The Fisher Information is a fundamental concept in statistical inference that measures the amount of information an observable random variable carries about an unknown parameter of interest. It quantifies how much the likelihood function $\mathcal{L}(\theta)$ tells us about the parameter θ. Formally, for a parametric model with density $f(x|\theta)$, the Fisher Information is defined as:

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]$$

Alternatively, under regularity conditions, it can also be expressed as the **negative expected second derivative** of the log-likelihood:

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\left(\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right)\right]$$

This quantity reflects how sharply peaked the likelihood function is around its maximum: **the more curved the log-likelihood, the more informative the data are about the parameter**.

The Fisher Information plays a key role in the **Cramér-Rao inequality**, which states that for any unbiased estimator θ:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}$$

Thus, Fisher Information sets a theoretical **lower bound for the variance** of an estimator. An estimator that attains this bound is considered efficient.

In the multivariate case, Fisher Information generalizes to a **matrix**, which is central in constructing **asymptotic covariance matrices** for maximum likelihood estimators.

- **Is the Fisher information used to compute standard errors? In which way?** Yes, the Fisher Information is used to compute standard errors, especially in the context of MLE. When estimating a parameter θ via MLE, under regularity conditions, the sampling distribution of the estimator $\widehat{\theta_{MLE}}$ is asymptotically normal:

$$\hat{\theta}_{\text{MLE}} \sim \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

where n is the sample size and $\mathcal{I}(\theta)$ is the Fisher Information.

This implies that the **asymptotic variance** of the MLE is the **reciprocal of the Fisher Information**. Therefore, the **standard error (SE)** of the estimator is given by:

$$\mathrm{SE}(\hat{\theta}) \approx \sqrt{\frac{1}{n \cdot \mathcal{I}(\hat{\theta})}}$$

In practice, since the true value θ is unknown, we compute the observed Fisher Information, often using the second derivative (the Hessian) of the log-likelihood at the MLE:

$$\mathcal{I}_{\text{obs}}(\hat{\theta}) = -\left.\frac{\partial^2}{\partial \theta^2} \log \mathcal{L}(\theta)\right|_{\theta=\hat{\theta}}$$

The standard error is then estimated as: $\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{1}{\mathcal{I}_{\text{obs}}(\hat{\theta})}}$. This approach is widely used in asymptotic inference to construct confidence intervals and perform hypothesis tests for MLEs.

- **Can you provide a brief description of the meaning of Bayesian Inference?** Bayesian inference is a statistical framework that updates the probability of a hypothesis as more evidence or data becomes available. It is based on **Bayes' Theorem**, which combines **prior beliefs** with **new observed data** to form a **posterior distribution**:

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta) \cdot P(\theta)}{P(\text{data})}$$

Where: - $P(\theta)$ is the **prior distribution**: what we believe about the parameter θ before seeing the data; - $P(\text{data} \mid \theta)$ is the **likelihood**: how likely the observed data are for a given θ; - $P(\theta \mid \text{data})$ is the **posterior distribution**: our update belief about θ after seeing the data.

In Bayesian inference, **parameters are treated as random variables**, and all uncertainty is expressed through probability distributions. This contrasts with classical (frequentist) inference, where parameters are fixed and only the data are considered random.

Bayesian methods are especially useful when: prior information is available or required, sample sizes are small and/or full probability distributions for parameters are needed (e.g., for decision making or uncertainty quantification).

- **What is it the nonparametric bootstrap?** The nonparametric bootstrap is a **resampling method** used to estimate the **sampling** distribution of a statistic without making assumptions about the underlying population distribution.

It works by generating many "new" datasets (called **bootstrap samples**) by **sampling with replacement** from the original observed data. Each bootstrap sample has the same size as the original dataset and may contain repeated observations.

The procedure is the following:

1. Given a dataset of size n, generate B bootstrap samples by sampling with replacement.
2. Compute the statistic of interest (e.g., mean, median, standard deviation) on each sample.
3. Use the resulting distribution of bootstrap statistics to estimate: standard errors, bias, confidence intervals.

The nonparametric bootstrap is especially valuable when analytical formulas for variance or confidence intervals are not available or are unreliable due to model assumptions.

- **How it is used to evaluate the accuracy of an estimator?** The nonparametric bootstrap is widely used to evaluate the accuracy of an estimator by **estimating its sampling variability**, typically summarized by the **standard error** or a **confidence interval**.

Since we often have only one sample from the population, the bootstrap simulates what would happen if we could resample the population many times, using the original sample as a stand-in for the population.

The bootstrap provides a data-driven way to assess the stability and precision of an estimator, especially when theoretical formulas are difficult to apply or the sample size is small.

- **Describe the distribution of a certain estimator obtained through this resampling method. Why and how it can be used to construct confidence intervals?** When we apply the nonparametric bootstrap, we repeatedly compute a statistic (such as the mean, median, or standard deviation) on many samples drawn with replacement from the original dataset. This produces an empirical distribution of the estimator, called the **bootstrap distribution**. It approximates how the estimator would vary if we could repeat the sampling process many times, and therefore reflects the variability and uncertainty of the estimate based solely on the observed data, without relying on parametric assumptions.

The bootstrap distribution can be used to construct **confidence intervals**. One common approach is the **percentile method**, which simply takes the appropriate percentiles from the bootstrap estimates. For example, a 95% confidence interval is given by the 2.5th and 97.5th percentiles of the bootstrap distribution. Another method involves computing the **standard error** of the estimator as **the standard deviation of the bootstrap values**, and then applying the normal approximation formula $\widehat{\theta} \pm z_{\alpha/2} \cdot \mathrm{SE}$. More advanced approaches, such as the bias-corrected and accelerated (BCa) method, adjust for bias and skewness in the distribution.

In summary, the bootstrap distribution provides a powerful and flexible way to evaluate the accuracy of an estimator and to construct confidence intervals, especially when the sample size is small or the assumptions of traditional methods are not met.

- **Describe the bootstrap percentile method used to obtain a confidence interval for the parameter estimate.** The bootstrap percentile method is a simple and widely used approach to construct a confidence interval for a parameter estimate based on the **empirical distribution** of bootstrap replications. After generating a large number of bootstrap samples (typically 1000 or more), the statistic of interest (e.g., mean, median, standard deviation) is computed for each sample, resulting in a distribution of estimates.

To construct a **confidence interval at level** $1-\alpha$, we select the **lower** $\alpha/2$ **and upper** $\alpha/2$ **percentiles** of the bootstrap estimates. For example, a 95% confidence interval corresponds to the 2.5th and 97.5th percentiles of the bootstrap distribution. This interval **does not rely on any parametric assumptions** about the shape of the sampling distribution and is especially useful when the statistic has a skewed or non-normal distribution.

The percentile method is intuitive and fully data-driven. It provides an **approximate range of plausible values** for the true parameter based solely on the observed data and its resampled versions. It is particularly effective when the bootstrap distribution is approximately symmetric and centered near the original estimate, but it can still offer useful inference even in small samples or with complex estimators.

4 Week 4-5 - Descriptive analysis and correlation structure, Multiple linear regression and model interpretation, Model selection and residual diagnostics, Out-of-sample prediction and bootstrap confidence intervals, Theoretical foundations of linear regression

4.1 Multiple linear regression on firm scores: model fitting, residual analysis, parameter interpretation, and assessment of model adequacy.

The `ratings.Rdata` file analyzed on Page 30 of the teaching notes (applications) is related to the ratings of the firms received according to their level of profitability, capital structure, and financial flexibility. (See in the teaching notes of the applications the data description and preliminary explanatory statistical analyses).

4.1.0.1 13.1 Write the research questions which can be of interest when scores of the firms are considered as response variable.

When the credit rating score of firms is considered as the response variable, several research questions can be formulated to guide the statistical analysis:

- Is the credit rating of a firm significantly associated with its financial characteristics, such as profitability, capital structure, and financial flexibility?
- Which covariates play the most important role in determining the firm's rating?
- How much does the rating score change, on average, for a one-unit increase in each of the covariates?
- Can we use the available data to build a regression model that allows us to predict the credit rating of a new firm based on its financial indicators?
- How precise are these predictions, and how much uncertainty is associated with them?
- Do the associations observed in the sample allow us to make inferences about the population of firms operating in the same industry?

These questions aim to understand not only the **relationship between the response and the covariates**, but also to **quantify the effects** and to provide **useful predictive tools** for decision-making in financial risk analysis.

```
load("ratings.RData")
model13 <- lm(rating ~ profit + capital + f_flex, data = ratings)

summary(model13)
```

4.1.0.2 13.2 Fit a multiple linear regression model. Report the results printed with the 'summary' function.

```
##
## Call:
## lm(formula = rating ~ profit + capital + f_flex, data = ratings)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## profit        0.3277     4.4598   0.073  0.94198
## capital       3.9118     1.2484   3.133  0.00425 **
## f_flex       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

We fitted a multiple linear regression model with the firm's **rating** as the response variable and **profit**, **capital**, and **financial flexibility (flex)** as explanatory variables. The model was estimated using the `lm()` function in R.

The regression output is summarized as follows:

- The estimated model is:

$$\hat{y}_{\text{rating}} = -28.88 + 0.33 \cdot \text{profit} + 3.91 \cdot \text{capital} + 19.67 \cdot \text{flex}$$

Among the covariates: - **Capital** is **statistically significant** at the 1% level ($p = 0.00425$), indicating a strong positive association with the firm's rating. A one-unit increase in capital is associated with an estimated +3.91 points in the rating. - **Financial flexibility (flex)** is also **significant** at the 5% level ($p = 0.03108$): a one-unit increase in flexibility corresponds to an estimated +19.67 points in the rating. - **Profit** in **not statistically significant** ($p = 0.94198$), suggesting that it does not meaningfully explain variations in the rating within this sample.

- The model explains about 65.2% of the variability in the response ($R^2 = 0.6518$), with an **adjusted** $R^2 = 0.6116$ to account for the number of predictors.
- The overall model is **highly significant**, as indicated by the **F-statistic** of 16.22 and a **p-value** $<$ **0.001**, showing that the covariates collectively provide a statistically significant explanation of the firm's rating.

4.1.0.3 13.3 Write the equation of the estimated model for unit with id 29. Interpret the values.

$$\hat{y}_{29} = -28.88 + 0.33 \cdot 5.802 + 3.91 \cdot 6.685 + 19.67 \cdot 1.08$$

For unit 29, with profit = 5.802, capital = 6.685, and financial flexibility = 1.08, the predicted rating is:

$$\hat{y}_{29} \approx 20.42$$

This value represents the conditional expected rating of firm 29 given its observed financial indicators.

The actual observed value for unit 29 is $y_{29} = 13.4$, and therefore the residual is:

$$z_{29} = y_{29} - \hat{y}_{29} = 13.4 - 20.42 = -7.02$$

This residual indicates that the observed rating is approximately 7 points lower than the fitted value predicted by the model, suggesting that either additional unobserved factors influenced the rating, or that the linear model only partially explains the firm's evaluation.

In accordance with the theoretical framework, the estimated parameters $\hat{\beta}_j$ can be interpreted as **partial regression coefficients**, i.e., the change in the expected rating for a unit change in each explanatory variable, holding the others fixed. For example, for every additional unit of capital structure, the rating is expected to increase by 3.91 points, all else equal.

4.1.0.4 13.4 Comment on the descriptive statistics of the estimated residuals.

The residuals $z_i = y_i - \hat{y}_i$ provide a numerical quantification of the deviation of each observation from the regression hyperplane and are used to assess the adequacy of the model assumptions, particularly regarding linearity, constant variance (homoscedasticity), and absence of outliers or influential observations.

The descriptive statistics of the estimated residuals from the fitted model are:

```
summary(residuals(model13))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -17.391  -6.612  -1.009   0.000   4.908  25.449
```

These values indicate that residuals are approximately centered around zero, as expected under the assumption $\mathbb{E}[\varepsilon_i] = 0$. However, the distribution is **slightly asymmetric**, with a longer right tail (maximum = 25.449 vs minimum = –17.390), suggesting the presence of one or more firms whose ratings are substantially **higher than predicted** by the model.

The **residual standard error**, which estimates the standard deviation of the disturbance term σ, is approximately 10.13. This value reflects the average discrepancy between the observed and predicted ratings. In the context of this analysis, it means that, on average, the actual rating of a firm may deviate by about **10 rating points** from the model prediction, once the effects of `profit`, `capital` and `flex` are accounted for.

In conclusion, the residuals exhibit a distribution compatible with the assumptions of the classical linear model, but further diagnostic checks (e.g., residual plots, QQ-plots) would be needed to fully validate homoscedasticity and normality.

4.1.0.5 13.5 Comment on the residual standard error and on the R^2 and the adjusted R^2.

The residual standard error (RSE) is an estimate of the standard deviation σ of the error terms ε_i in the linear regression model. It represents the typical size of the residuals, that is, the **average deviation of the observed values from the fitted values**. In the fitted model, the RSE is approximately:

$$\hat{\sigma} = 10.13$$

This means that, on average, the observed credit ratings deviate from the values predicted by the model by about 10 points. The residual standard error is measured in the same units as the response variable (here, rating) and gives an indication of the **absolute accuracy** of the prediction.

The **coefficient of determination**, denoted by R^2, is equal to $R^2 = 0.6518$.

This quantity represents the **proportion of the total variability** in the response variable that is explained by the regression model. In this case, approximately **65.2% of the variance in credit ratings** is explained by the covariates `profit`, `capital`, and `flex`.

The **adjusted R^2, which is equal to Adjusted $R^2 = 0.6116$, takes into account the number of explanatory variables relative to the sample size and penalizes the inclusion of variables that do not contribute meaningfully to the model. Since the adjusted R^2 is slightly lower than R^2, it suggests that while the model has reasonable explanatory power, not all covariates may be equally informative.

In summary, the residual standard error indicates that the model provides a moderately accurate fit in absolute terms, while the values of R^2 and adjusted R^2 confirm that the model captures a substantial part of the variability in firm ratings, though there is still unexplained variation likely due to omitted variables or inherent uncertainty.

4.1.0.6 13.6 Describe the estimated values of the regression parameters. What we can conclude about the research question?

The **intercept** $\hat{\beta}_0 = -28.88$ represents the expected value of the response variable (rating) when all explanatory variables are equal to zero. Although this value has no direct interpretation in the context of this application—since it is unlikely that a firm simultaneously has zero profit, capital, and flexibility—it is required for model completeness.

The coefficient for **profitability** is $\hat{\beta}_1 = 0.33$, but it is **not statistically significant** ($p = 0.942$), suggesting that, after adjusting for capital structure and financial flexibility, profit does not show a measurable influence on the firm's rating. Therefore, its contribution to the explanation of the response appears negligible in this model.

The coefficient for **capital structure** is $\hat{\beta}_2 = 3.91$, and it is **significant at the 1% level** ($p = 0.004$). This indicates that, holding other covariates constant, a one-unit increase in capital is associated with an average increase of 3.91 points in the firm's credit rating. This reflects a meaningful positive conditional association.

The coefficient for **financial flexibility** is $\hat{\beta}_3 = 19.67$, and it is **significant at the 5% level** ($p = 0.031$). This suggests that a one-unit increase in financial flexibility corresponds to an average increase of nearly 20 points in the rating, controlling for the other variables. This indicates a strong and positive effect.

With respect to the initial research question—whether the credit rating can be associated with other firm-level financial variables—we can conclude that **capital structure** and **financial flexibility** are relevant explanatory variables. They exhibit statistically significant and positive conditional effects on the rating. On the other hand, **profitability** does not contribute meaningfully in the presence of the other covariates.

Thus, the model supports the idea that **firm ratings are systematically associated with measurable financial indicators**, particularly capital and financial flexibility, which can be used for explanation and prediction within the population of interest.

4.1.0.7 13.7 Which is the residual for unit with id 29?

As previously computed in point 13.3, the residual for unit with id 29 is $z_{29} = -7.02$, meaning that this firm's observed credit rating is **7.02 points lower** than the value predicted by the model based on its financial indicators. This negative residual may reflect unmeasured characteristics, model imperfections, or idiosyncratic effects affecting this particular firm.

4.1.0.8 13.8 Check the properties of the residual on your results.

According to the classical linear model assumptions (Section 4.1 of the Teaching Notes), the residuals $z_i = y_i - \hat{y}_i$ obtained from the least squares estimation should satisfy a set of **key properties**, which can be checked on the estimated model.

1. **Zero mean**: the sum (and hence the mean) of the residuals is always equal to zero in a linear regression model with intercept:

$$\sum_{i=1}^{n} z_i = 0$$

In practice, due to rounding, we expect a numerical mean of the residuals very close to 0. This can be verified with:

```
mean(residuals(model13))
```

```
## [1] -4.736952e-16
```

2. **Uncorrelated with the fitted values**: residuals should be uncorrelated with the predicted values $\hat{y}_i$. That is:

$$\text{Cov}(z_i, \hat{y}_i) = 0$$

This can be assested numerically with:

```
cor(residuals(model13), fitted(model13))
```

```
## [1] 7.396095e-17
```

3. **Uncorrelated with the explanatory variables**: residuals are **orthogonal to the columns of the design matrix** X, meaning they are uncorrelated with each explanatory variable:

$$\sum_{i=1}^{n} z_i x_{ij} = 0, \quad \text{for each } j$$

These conditions hold by construction in OLS and can be checked via:

```
cor(residuals(model13), ratings$profit)
```

```
## [1] 4.138691e-16
```

```
cor(residuals(model13), ratings$capital)
```

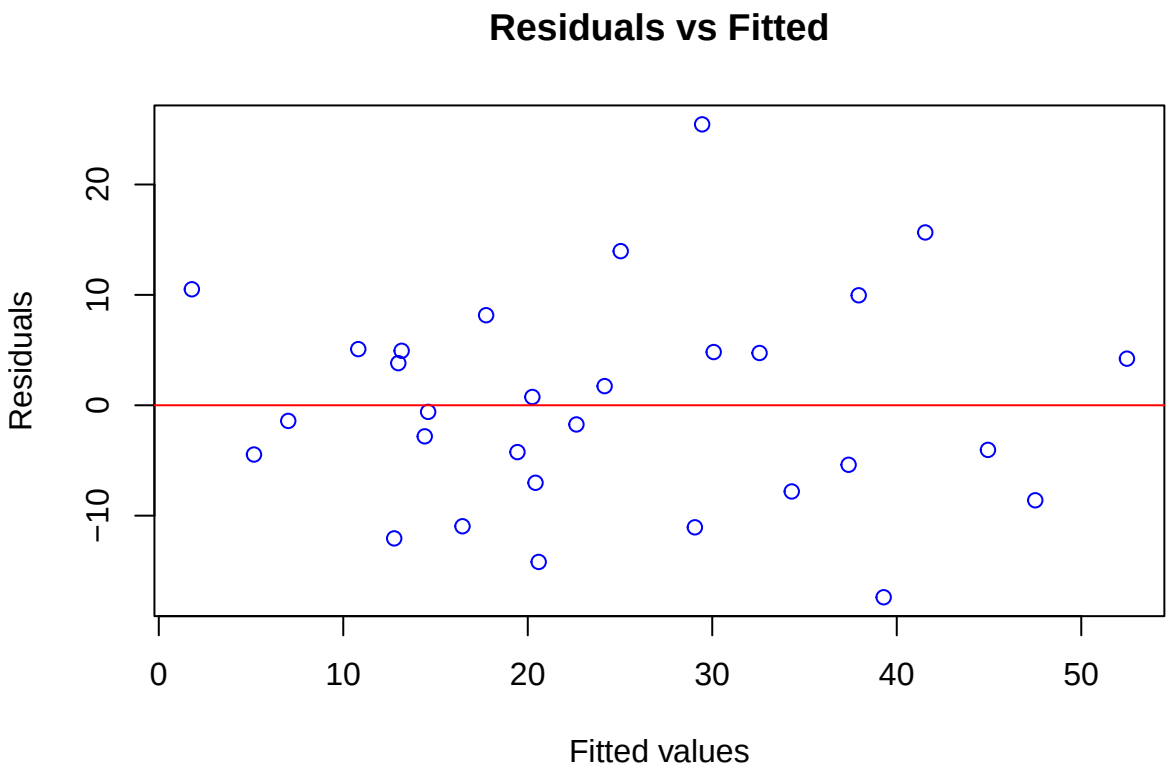
```
## [1] -7.205652e-17
```

```
cor(residuals(model13), ratings$f_flex)
```

```
## [1] -5.260178e-17
```

4. **Homoscedasticity**: under the classical assumptions, the residuals should exhibit **constant variance** across levels of the fitted values. This can be visually assessed by plotting:

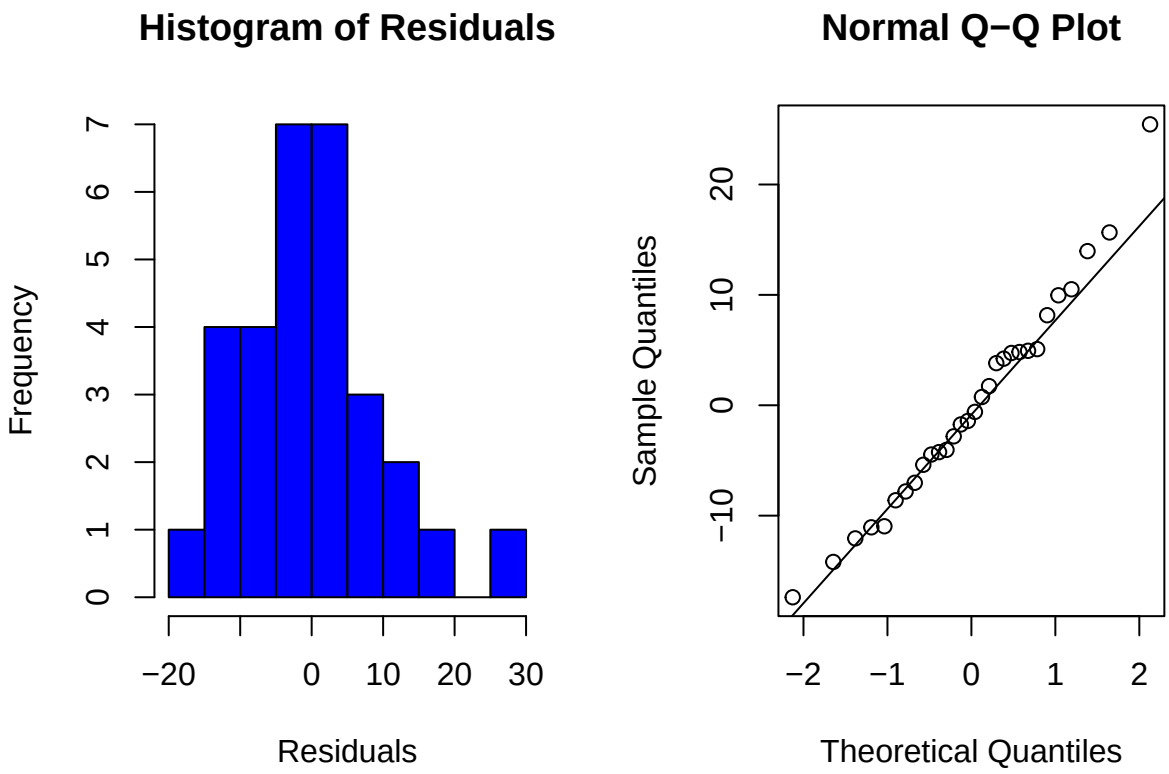
```
plot(fitted(model13), residuals(model13),
     main = "Residuals vs Fitted",
     xlab = "Fitted values",
     ylab = "Residuals",
     col = "blue")
abline(h=0, col = "red")
```



In a well-behaved model, residuals should appear **randomly scattered** around zero with **no clear pattern** or funnel shape.

5. **Approximately Normal distribution**: while not a required assumption for OLS estimation, **normally distributed residuals** are desirable for inference (e.g., confidence intervals, p-values). This can be verified with:

```
par(mfrow = c(1,2))
hist(residuals(model13),
     main = "Histogram of Residuals",
     xlab = "Residuals",
     breaks = 10,
     col = "blue")
qqnorm(residuals(model13)); qqline(residuals(model13))
```



In conclusion: the residuals of the model estimated in Exercise 13.2 are expected to conform to the theoretical assumptions of the classical linear model. Empirical verification (mean ≈ 0, lack of correlation with fitted and explanatory variables, constant variance, approximate normality) is essential for validating model adequacy and reliability of statistical inference. Based on these diagnostics, one can assess whether the model appropriately captures the structure of the data and whether any refinements are needed.

4.2 Exploratory and regression analysis on starting salaries: study design, covariate relationships, correlation structure, and model interpretation.

The data in the file `bank.Rdata` are related to starting salaries of all skills, entry level, clerical workers between 1965 and 1975.

We dispose of the following variables:

- `basl`: beginning salary (annual salary at time of hire)
- `sal77`: annual salary in 1977
- `senior`: months since hired
- `age`: in months
- `edu`: years of education
- `exper`: months of prior work experience

4.2.0.1 14.1 Explain if the data are collected with a randomized experiment or under an observation study. Define the sample size, and identify the response variable and the covariates. Show also the last six rows of the data and comment on these values. The data in the file `bank.Rdata` are collected under an observation study, not a randomized experiment. The sample size,

computed as the number of rows in the dataset, is equal to 93. The response variable is the starting annual salary of the workers (bsal), while the covariates include salary in 1977 (`sal77`), months since hired (`senior`), age in months, years of education (`edu`), and months of prior work experience (`exper`): we aim at predicting the starting salary as a function of the other aforementioned variables (Note that alternatively the `sal77` variable could serve as the response variable, instead of bsal). The `tail()` function displays the last six rows of the data.

```
load("bank.Rdata")
tail(bank)
```

```
##    bsal sal77 senior age educ exper
## 88 4800  9240     84 571   16 214.0
## 89 6000 11940     86 486   15  78.5
## 90 4380 10020     93 313    8   7.5
## 91 5580  7860     69 600   12 132.5
## 92 4620  9420     96 385   12  52.0
## 93 5220  8340     70 468   12 127.0
```

Among the considered observations, the one with ID 90 shows particularly low values for the number of years of education (8) and the number of months of previous work experience (7.5); it is also the youngest among the subjects considered. The starting salary is rather low (the lowest among the six statistical units examined). Nevertheless, since being employed for a particularly long time (93 months), the annual salary in 1977 is quite high compared to that of the other 5 subjects examined.

We also observe the following:

- `bsal` varies between 4380 and 6000, showing some variability in entry-level salaries.
- `educ` ranges from 8 to 15 years, which may reflect differences in high school vs some college education.
- There is wide variation in `exper`, from just 7.5 months to over 132 months, suggesting a heterogeneous background in terms of prior work history.
- Some individuals are relatively older at hire, with `age` over 90 months (i.e., 7.5+ years older) than others.
- All individuals have `senior` > 84, indicating they had been with the bank for several years at the time of observation.

These observations support the idea that starting salary (`bsal`) may be influenced by both formal education and work experience, and possibly moderated by age or tenure with the bank.

4.2.0.2 14.2 Describe the observations reporting and commenting on the descriptive statistics for each variable.

We use the `skim_without_charts()` function to obtain the main univariate descriptive statistics.

```
skim_without_charts(bank)
```

Table 9: Data summary

Name	bank
Number of rows	93
Number of columns	6
Column type frequency:	

numeric	6
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
bsal	0	1	5420.32	709.59	3900	4980.0	5400	6000	8100
sal77	0	1	10392.90	1789.64	7860	9000.0	10020	11220	16320
senior	0	1	82.28	10.25	65	74.0	84	90	98
age	0	1	474.40	140.21	280	349.0	468	590	774
educ	0	1	12.51	2.28	8	12.0	12	15	16
exper	0	1	100.93	90.95	0	35.5	70	144	381

Considering the two variables that measure the amount of salary (`bsal` and `sal77` variables), it is observed that the salary recorded in 1977 is significantly higher than the initial one. On average, the workers considered earned almost twice as much. The range of variation for both salaries is quite wide; for example, that related to the time of hiring ranges from a minimum of $3900 to a maximum of $8100. The value of the standard deviation is consequently high: the workers have an average deviation around the mean of about $709. The mean and median values are approximately equal: half of the workers earned less than $5400 at the time of signing, and half of the workers earned less than $10,020 in 1977. Another characteristic that the two variables related to the amount of salary have in common is the overall shape of the distribution, which is particularly pointed (in both cases, the first and third quartiles are very close to the mean and median): most of the workers receive a salary not far from the average, and only a small minority earn significantly more or less.

The variable `exper` measures the number of months of previous employment; the minimum value is 0, indicating that for some workers (at least one), this is their first job. Conversely, workers with more previous experience can have up to 381 months (more than 31 years) of previous work. Again, the range of variation is quite wide. The variable is strongly skewed, with a long tail to the right (skewness index greater than 1: `e1071::skewness(bank$exper)` = 1.13): there are many workers with few months of previous employment and viceversa. The number of years of education (`educ` variable) is much less variable: no worker has more than 16 years or less than 8 years. The mean and median values are approximately equal to 12 years.

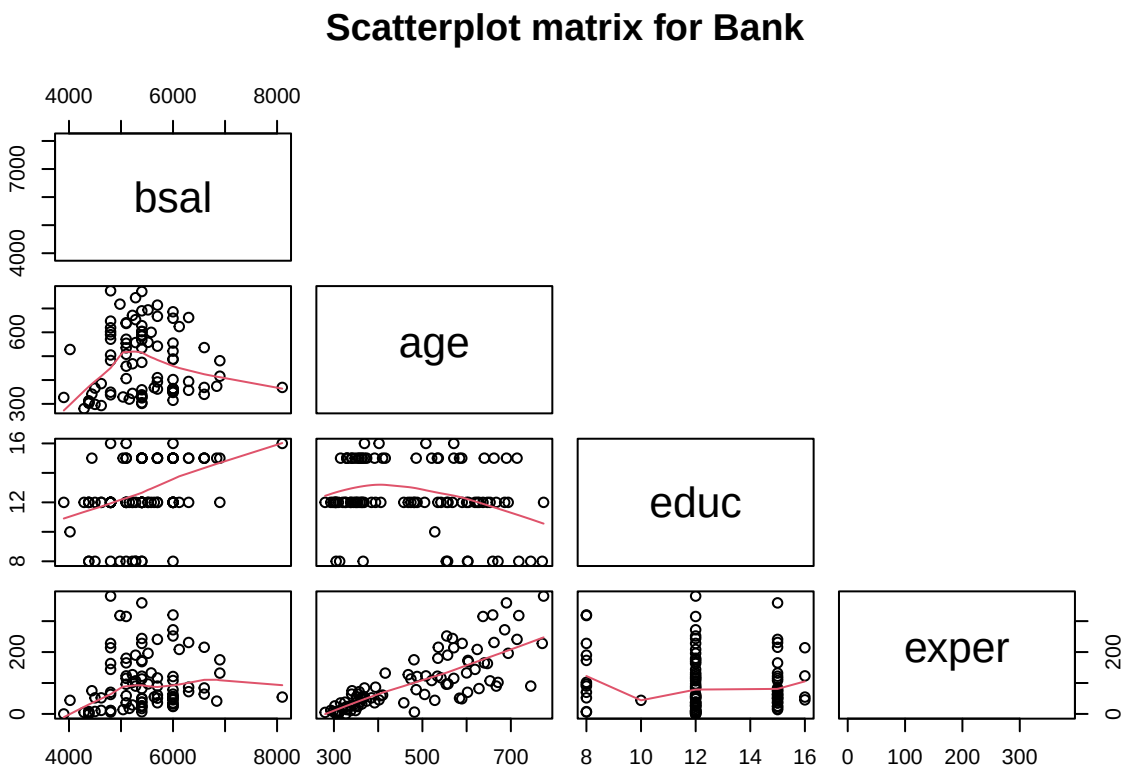
Regarding the number of months since hiring (`senior` variable), no worker has been employed for less than 6 months (about 5 and a half years), and none for more than 98 months (just over 8 years); it is a very narrow range. The median value is 84 months (7 years), slightly shifted towards the right side of the distribution. Finally, the age of the workers (`age` variable) ranges from a minimum of 23 years to a maximum of 65. Half of the workers are under 39 years old, and a quarter are under 29.

4.2.0.3 14.3 How would you expect age, experience and education be related to starting salaries? Generate appropriate explanatory plots. Are the associations as you expected? What implication this results may have for modelling?

Experience and education are likely to be positively associated to the starting salaries: workers with more work experience and higher levels of education tend to have developed valuable skills and knowledge leading to higher salaries. Age is likely to have a positive association with salary for younger individuals with lower age values, while this association becomes nearly zero or even negative as age increases. One possibility to check this hypothesis is `pairs()` function, which represents the matrix of scatterplots between each pair of variables.

```
pairs(bank[, c(1, 4, 5, 6)],
      panel = panel.smooth,
```

```
      upper.panel = NULL,
      main = "Scatterplot matrix for Bank")
```



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.

4.2.0.4 14.4 Do you think it would be important to control for the number of years with the bank? Why?

Yes, it would be **important to control for the number of years with the bank** (captured by the variable `senior`, measured in months), because it may reflect **institution-specific experience**, **loyalty**, and **internal progression opportunities** that can directly influence salary decisions.

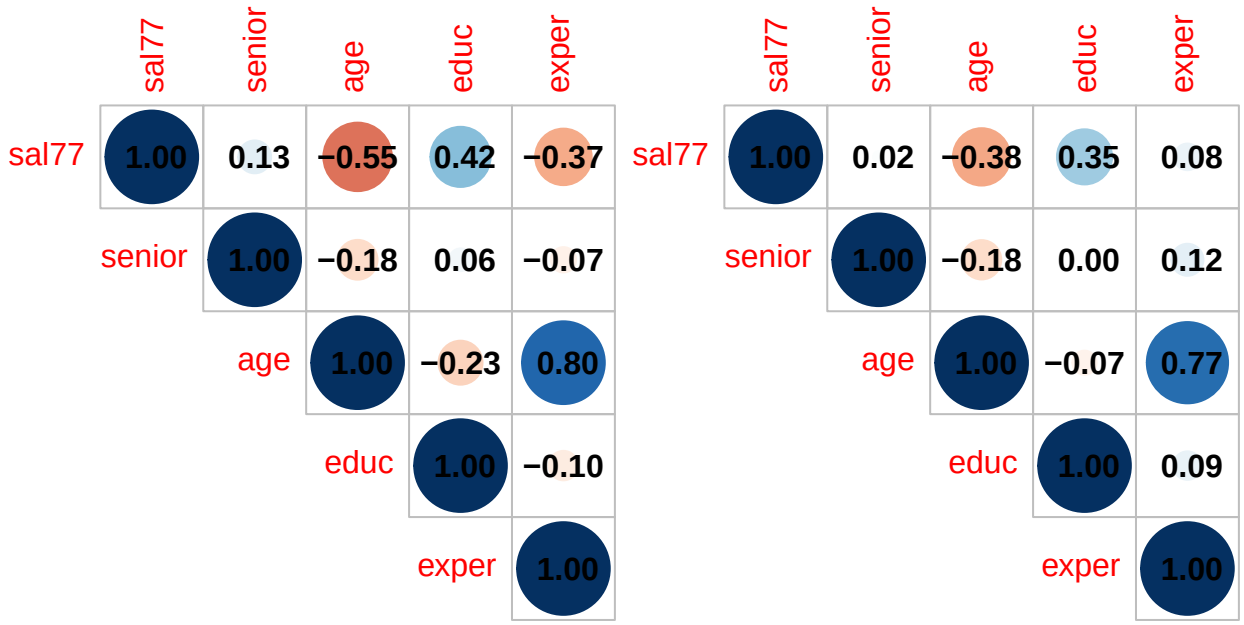
Even if an individual has a similar level of education or prior experience, having spent more time within the same institution could mean: - they received more **promotions or raises** due to internal evaluations; - they may have reached a **higher position** or job grade; - they benefit from **institution-specific knowledge** valued by the employer.

Not including `senior` in the model could lead to **omitted variable bias**, causing us to overestimate or underestimate the effects of other covariates such as age or total experience. Thus, **controlling for seniority helps isolate the effect of other predictors on salary**, allowing for more accurate and meaningful interpretations of the model.

4.2.0.5 14.5 Are the covariates associated? Plot and comment the corrplots for raw and partial correlations. What implications have high correlations for modelling?

To evaluate the associations between the covariates, we compute and visualize both the **raw** correlation matrix and the **partial correlation**.

```
library(ggm)
library(corrplot)

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.

4.2.0.6 14.6 Fit a multiple linear regression model with starting salary as response variable, experience and education as covariates. Report the results using the summary function and comment on the estimated parameters. Is there a difference of the covariates on the salary and how is it?

We fit the multiple linear regression using the `lm()` function, specifying the name of the dataset and the formulation of the model.

```
model14 <- lm(bsal ~ exper + educ, data = bank)
summary(model14)
```

```
##
## Call:
## lm(formula = bsal ~ exper + educ, 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 ***
## exper          1.6430     0.7331   2.241   0.0275 *
## educ         134.7096    29.2112   4.612 1.32e-05 ***
## ---
## 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 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^{-4}: 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 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 β_1 is zero.
  - The same interpretation 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.

4.3 Linear regression on agricultural output: exploratory analysis, model fitting, coefficient interpretation, residual diagnostics, and model assumptions.

Consider the data in the file `agriculture.Rdata` (simulated data) that pertain to an agricultural company's production during the last year.

The available variables are the following:

- number of agricultural products (`quantity` variable, measured in tonnes)
- number of cultivated hectares (`extension` variable, measured in hectares)
- number of workers employed for cultivation (`workers` variable, measured in a number of workers).

```
load("agriculture.Rdata")
skim_without_charts(agriculture)
```

4.3.0.1 15.2 Perform data description and preliminary explanatory statistical analyses. Comment on the reported quantities and add the scatterplot matrix. Comment on the graphs, they suggest linear associations?

Table 11: Data summary

Name	agriculture
Number of rows	729
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
quantity	0	1	72.37	12.38	24.0	64.0	72.0	80.0	122.0
extension	0	1	32.47	6.89	18.2	27.5	32.4	36.6	67.1
workers	0	1	33.32	11.75	21.0	24.0	29.0	41.0	81.0

Based on the descriptive statistics, all three variables (`quantity`, `extension`, and `workers`) appear to be complete, with no missing values and reasonably wide ranges, indicating diversity in the agricultural settings observed.

- **Quantity (tonnes)** has a mean of approximately **72.37** with a standard deviation of **12.38**, ranging from **24 to 122**. This shows a moderate to high variability in production levels among observations.
- **Extension (hectares)** has a mean of **32.47** and standard deviation of **6.89**, with values ranging from **18.2 to 67.1**, suggesting that most farms are within a moderate size, but some large outliers exist.
- **Workers** show a mean of **33.32** with a standard deviation of **11.75**, and a minimum of **21** and maximum of **81**, indicating a relatively broad spread in labor input.

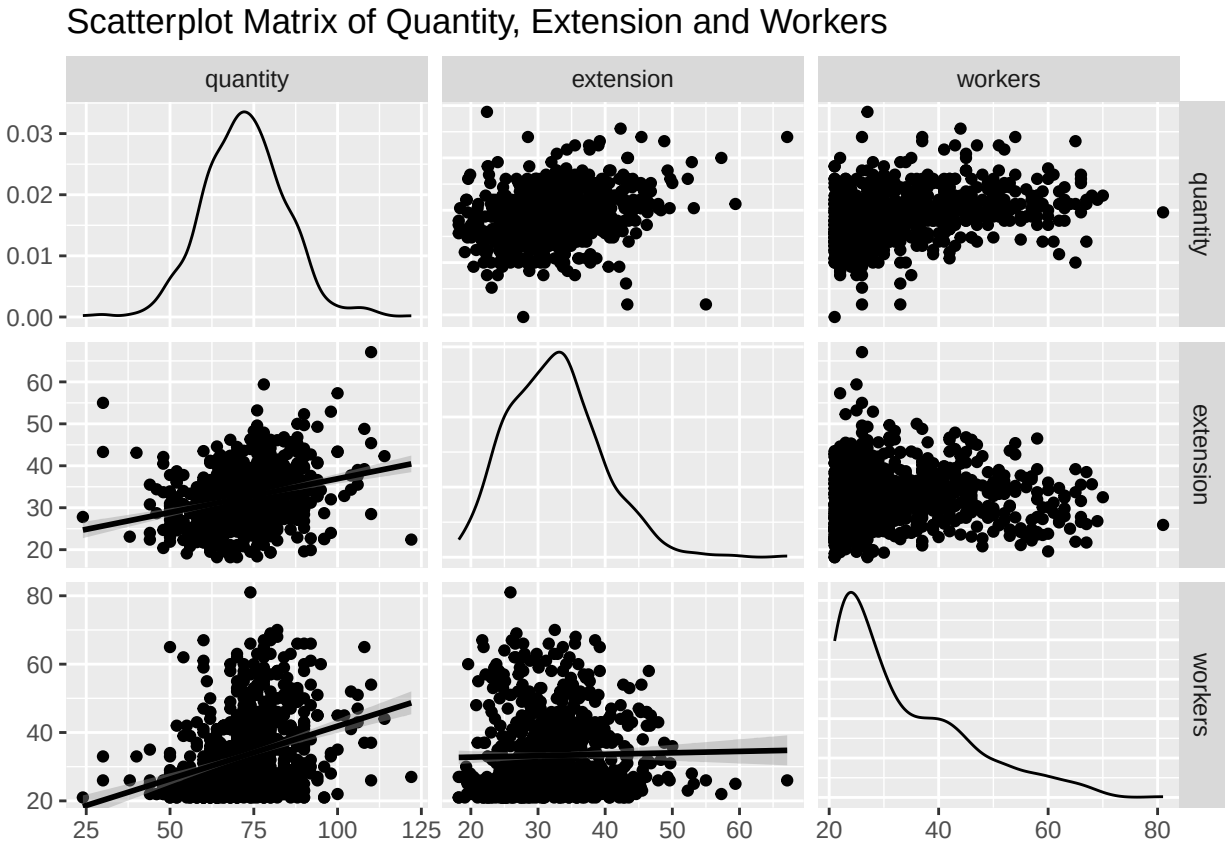
To explore potential **linear relationships**, a scatterplot matrix is helpful.

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)

ggpairs(agriculture,
        title = "Scatterplot Matrix of Quantity, Extension and Workers",
        upper = list(continuous = "points"),
        diag = list(continuous = "densityDiag"),
        lower = list(continuous = "smooth"))
```



The scatterplot matrix provides a detailed view of the pairwise relationships among the three variables. The diagonal panels show the estimated **density distribution** of each variable, which highlights their spread and modality. For example, **quantity** appears approximately symmetric, while **workers** is skewed with a concentration around 20.

The **lower panels** display smoothed regression lines with confidence bands. These suggest that both **extension** and **workers** are positively associated with **quantity**, though with different strengths. The relationship between **extension and quantity** is clearer and more linear, while the association between **workers and quantity** appears weaker and more dispersed.

The **upper panels**, with raw scatterplots, confirm these impressions. Most points are clustered within specific ranges (e.g., quantity between 60 and 90), and some outliers are visible in all pairings. The absence of strong curvature or heteroscedasticity supports the assumption of approximate **linear relationships**, at least between **quantity and extension**.

In summary, the plots support a **moderate linear association between extension and quantity**, and a **weaker relationship with workers**. These patterns suggest that in a regression model explaining quantity, **extension is likely to be a more relevant predictor than workers**.

```
model15 <- lm(quantity ~ extension + workers, data = agriculture)

summary(model15)
```

4.3.0.2 15.3 Fit a linear model to explain the amount of agricultural products. Report the general equation of the estimated model.

```
##
## Call:
## lm(formula = quantity ~ extension + workers, data = agriculture)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.304  -6.798  -0.802   6.683  56.847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.75491    2.29450  19.505  < 2e-16 ***
## extension    0.50573    0.06023   8.397  2.4e-16 ***
## workers      0.33590    0.03528   9.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 726 degrees of freedom
## Multiple R-squared:  0.1854, Adjusted R-squared:  0.1831
## F-statistic: 82.59 on 2 and 726 DF,  p-value: < 2.2e-16
```

We fit a linear regression model where the response variable is `quantity` (tonnes of agricultural products), and the predictors are `extension` (hectares of cultivated land) and `workers` (number of employees). The general form of the estimated model is:

$$\hat{y}_p = 44.75 + 0.51 \cdot \text{extension} + 0.34 \cdot \text{workers}$$

The **intercept** (44.75) represents the expected output when both `extension` and `workers` are equal to zero. While this situation is not practically interpretable (since farms with no land and no workers cannot exist), the intercept is necessary to anchor the linear model and center the prediction space.

The **coefficient for `extension` (0.51)** indicates that, **on average**, each additional hectare of land is associated with an increase of **0.51 tonnes** in production, holding the number of workers constant. This highlights the contribution of land area to productivity.

The **coefficient for `workers` (0.34)** suggests that **each additional worker** contributes an estimated **increase of 0.34 tonnes** in production, assuming the extension is held constant. The effect is weaker than that of land, but still positive.

The model's **coefficient of determination** is:

$R^2 = 0.1854$

This means that approximately **18.5% of the variability** in agricultural production is explained by the linear combination of the predictors. Although the explained variance is modest, the model still captures meaningful associations between inputs and output.

The **residual standard error (11.19)** quantifies the typical deviation between observed and predicted values, indicating that predictions deviate from actual production values by roughly **11.2 tonnes** on average.

Finally, the **F-statistic is highly significant** ($p < 2.2 \times 10^{-16}$), indicating that at least one predictor variable contributes significantly to explaining variability in the response. This supports the overall adequacy of the linear model and justifies the inclusion of the covariates in the regression.

4.3.0.3 15.4 Comment on the values of estimated regression coefficients.

The estimated regression coefficients provide insight into the contribution of each explanatory variable to the quantity of agricultural products produced:

- The **intercept** is **44.75**, which represents the **expected quantity (in tonnes)** when both *extension* (hectares) and *workers* are zero. While not interpretable in a practical sense (as both predictors being zero is unrealistic), it serves as a baseline value in the model.
- The **coefficient for extension** is **0.51**, meaning that for **each additional hectare cultivated**, the expected quantity increases by approximately **0.51 tonnes**, *holding the number of workers constant*. This coefficient is **highly significant** (*p-value < 2e-16*), indicating strong evidence of a positive linear relationship between cultivated land and production quantity.
- The **coefficient for workers** is **0.34**, suggesting that for **each additional worker**, the expected quantity increases by about **0.34 tonnes**, *controlling for the size of cultivated land*. This effect is also **highly significant** (*p-value < 2e-16*), confirming a meaningful association between labor input and output.

Both explanatory variables — **extension** and **workers** — have a **positive and statistically significant effect** on agricultural quantity. The **small standard errors** and **large t-values** confirm that the estimates are precise and meaningful in the context of this model. The interpretation aligns well with agronomic expectations: more land and more workers both contribute to greater output.

```
data.frame(
  unit = c(23, 56, 103),
  observed = agriculture$quantity[c(23, 56, 103)],
  fitted = fitted(model15)[c(23, 56, 103)],
  residual = residuals(model15)[c(23, 56, 103)]
)
```

4.3.0.4 15.5 Calculate and show the residuals for units 23, 56 and 103. The fitted values are similar to the observed values?

```
##     unit observed   fitted  residual
## 26    23       70 74.25512 -4.255124
## 60    56       64 73.13255 -9.132551
## 110  103       85 71.73087 13.269133
```

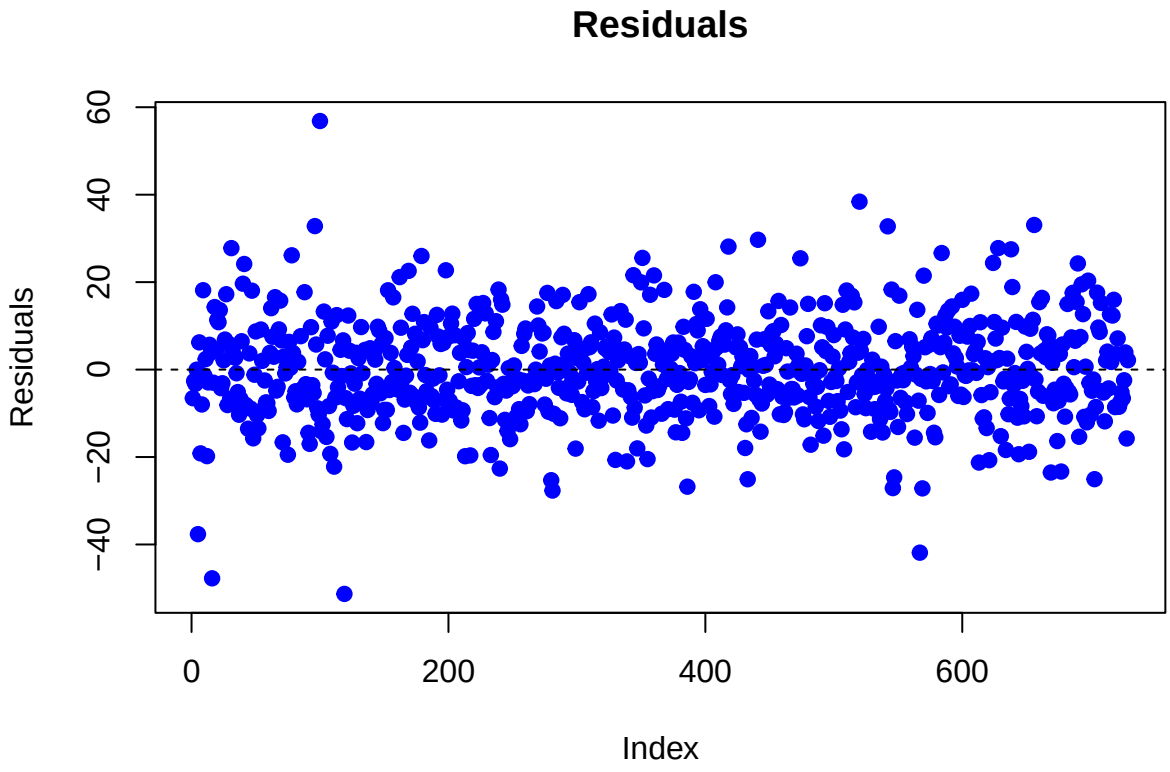
The residuals for units 23, 56, and 103 indicate how far the predicted values from the linear model are from the actual observed values:

- **Unit 23** has an observed quantity of 70, while the model predicts 74.26, resulting in a **residual of -4.26**, meaning the model **slightly overestimates** the actual value.
- **Unit 56** shows an observed quantity of 64, but the predicted value is 73.13, with a **residual of -9.13**. This suggests the model **notably overestimates** the true value for this unit.
- **Unit 103** has an observed quantity of 85, whereas the predicted value is 71.73, leading to a **residual of +13.27**. In this case, the model **underestimates** the actual quantity by a larger margin.

These results indicate that while the fitted values are in a **reasonable range**, there is **some variability** between predicted and observed values. This reflects the **residual variability** not captured by the model, which is consistent with the relatively **low R^2 value** observed in the regression summary.

```r
residuals_model15 <- residuals(model15)
fitted_model15 <- fitted(model15)

plot(residuals_model15,
     main = "Residuals",
     ylab = "Residuals",
     col = "blue", pch = 19)
abline(h = 0, lty = 2)
```

4.3.0.5 15.6 Plot the residuals, first alone, and then against the fitted values and against each covariate. Comment each plot. Can we detect some patterns? The assumption of linearity could be plausible? The constant variance of the error terms can be plausible?

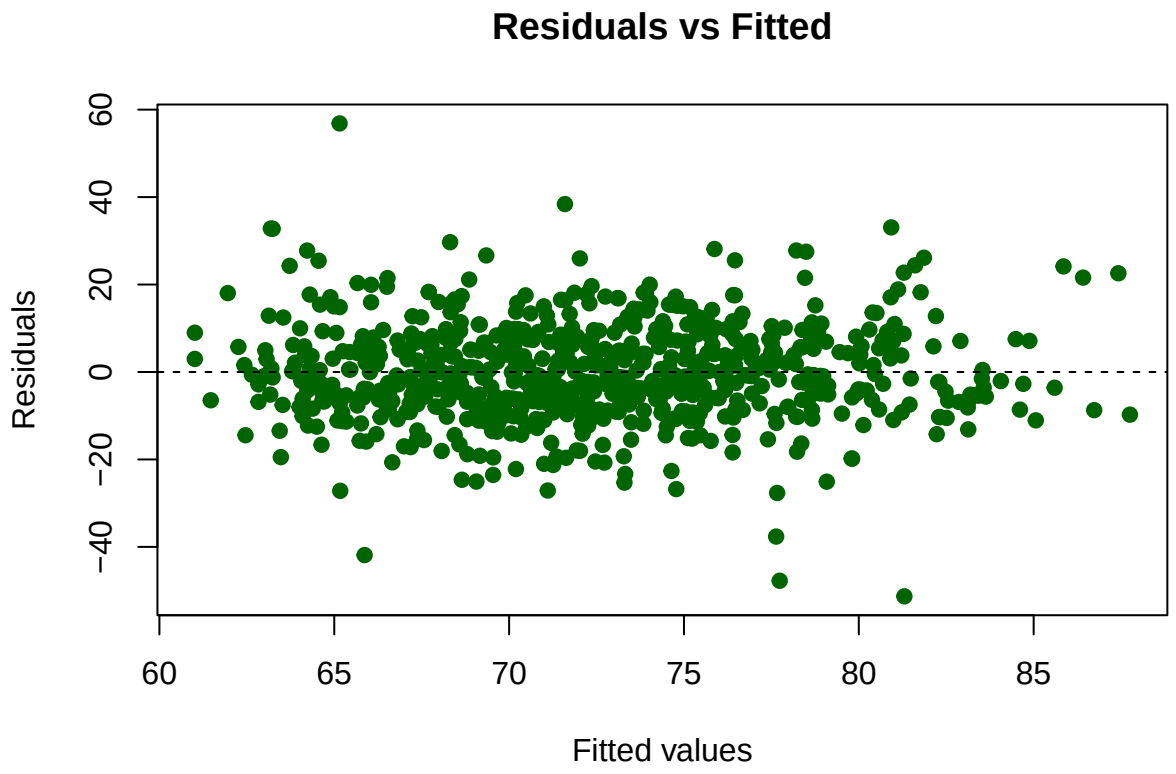


We observe that the **residuals are fairly symmetrically scattered around zero**, without any apparent trend over the index of observations. This supports the assumption that **errors are centered at zero** and **independent** across observations (i.e., there's no visible autocorrelation or systematic pattern over time or ordering).

The spread seems relatively constant across the range of the data, although there are **a few outliers** with large residuals. However, there is **no funnel shape or curve**, which is a good sign in terms of satisfying the **homoscedasticity** and **linearity** assumptions of the regression model.

In summary, this plot supports the basic assumptions of the linear model: **independent, zero-mean errors with constant variance**.

```r
plot(fitted_model15,
     residuals_model15,
     main = "Residuals vs Fitted",
     xlab = "Fitted values",
     ylab = "Residuals",
     col = "darkgreen",
     pch = 19)
abline(h = 0, lty = 2)
```

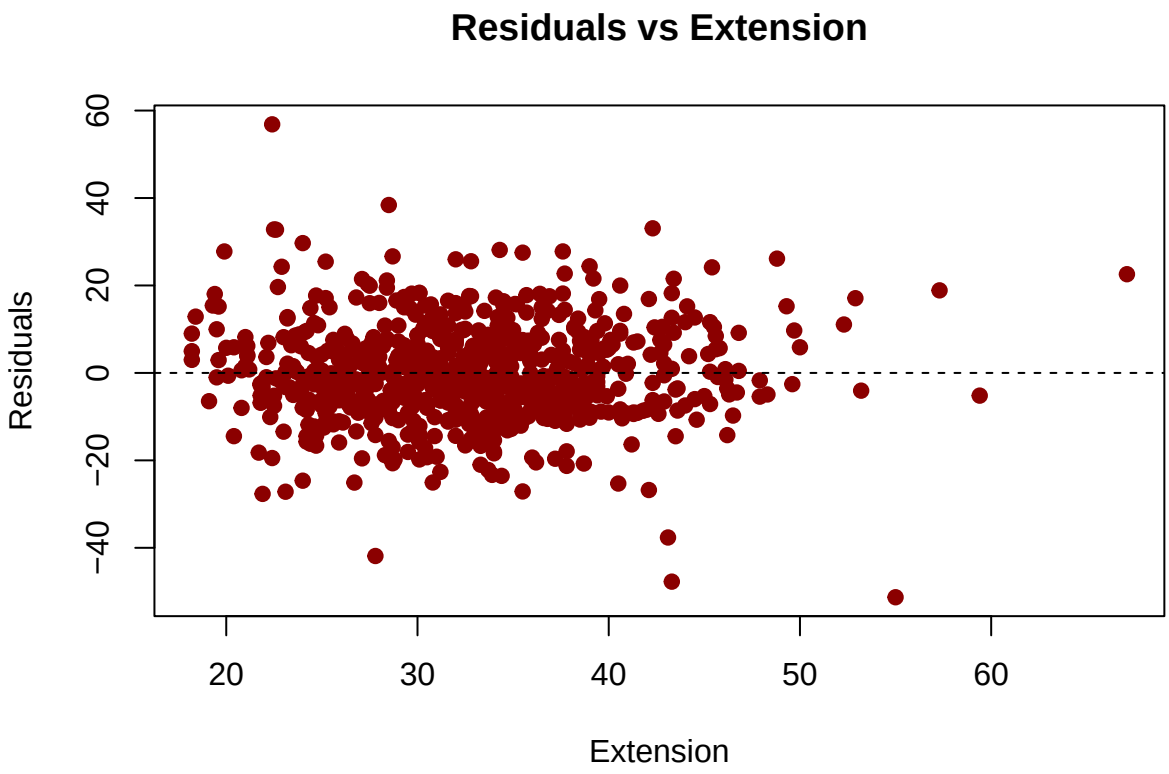


We see that the residuals are mostly spread **evenly around the zero line**, without a clear systematic pattern. This supports the assumption of **linearity**, as no curvature is evident.

However, there may be **slight signs of increasing spread** (i.e., **heteroscedasticity**) for higher fitted values, as the residuals seem to become more dispersed as the fitted values increase. This could suggest that **the variance of the error terms grows with the predicted quantity**, which might violate the constant variance assumption.

Overall, while **linearity seems plausible**, **homoscedasticity could be slightly questionable**, and it may be worth exploring variance-stabilizing transformations or robust regression if this pattern proves influential.

```r
plot(agriculture$extension, residuals_model15, main = "Residuals vs Extension",
     xlab = "Extension", ylab = "Residuals", col = "darkred", pch = 19)
abline(h = 0, lty = 2)
```

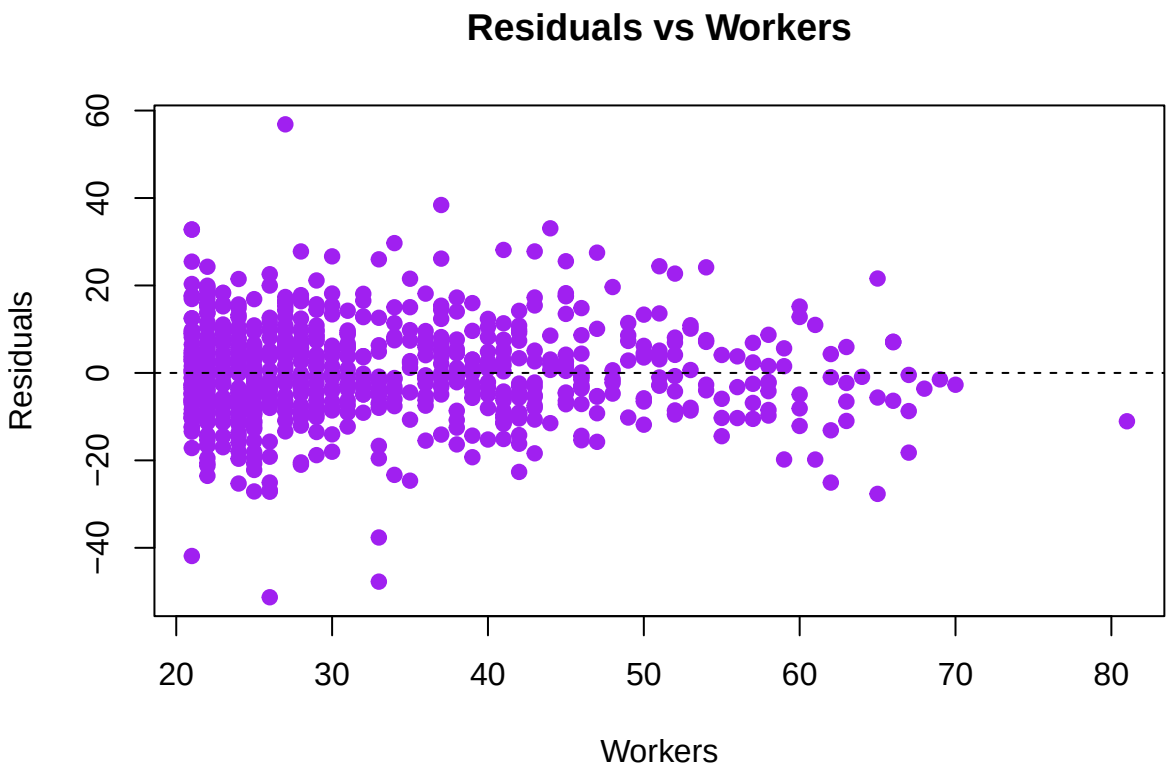


The residuals appear to be **fairly evenly distributed around zero**, with **no strong nonlinear patterns**, which supports the **assumption of linearity** between `extension` and the response variable `quantity`.

However, there is some **visible spread in the residuals for lower values of `extension`**, and a **tighter concentration for higher values**, suggesting a **mild heteroscedasticity** (i.e., non-constant variance). The effect is not dramatic, but it may slightly affect standard error estimates.

In conclusion, the linearity assumption appears **reasonable**, but the **constant variance assumption** may be **slightly violated**, especially for low `extension` values.

```
plot(agriculture$workers, residuals_model15, main = "Residuals vs Workers",
     xlab = "Workers", ylab = "Residuals", col = "purple", pch = 19)
abline(h = 0, lty = 2)
```



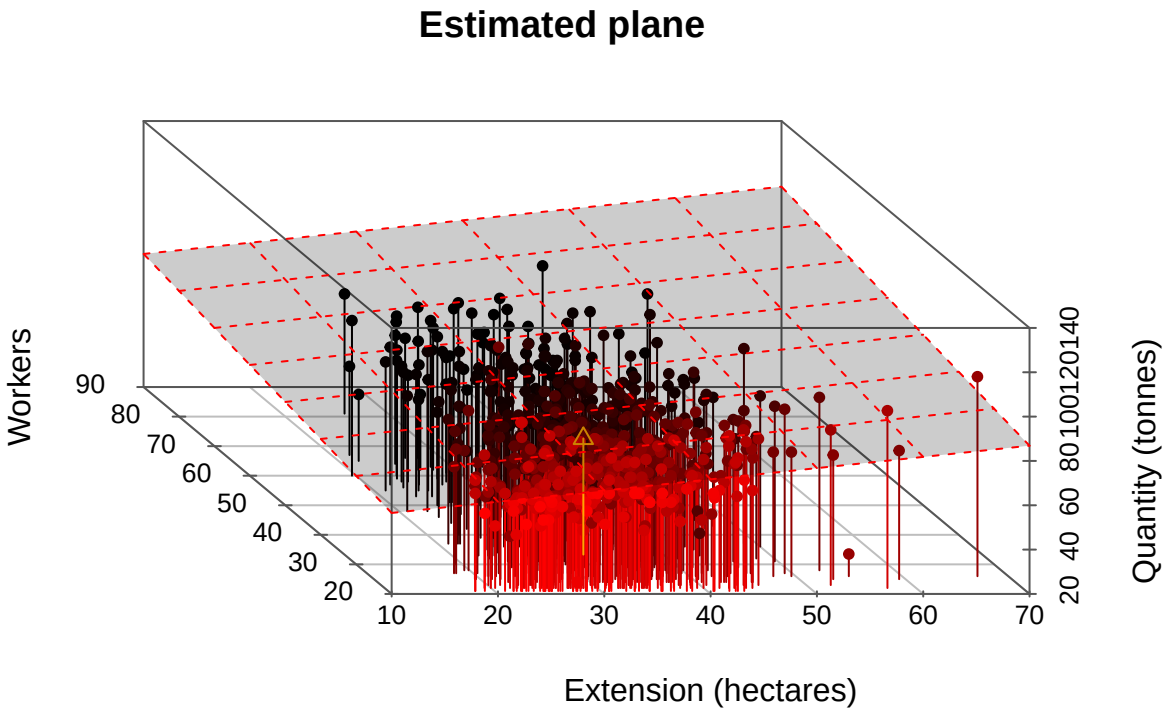
The residuals are generally **centered around zero**, which is consistent with the assumption of **mean-zero errors**. However, there is a **noticeable funnel shape**: for **lower values of `workers`**, the spread of residuals is wider and more variable, while for **higher values**, the residuals become more tightly clustered around zero.

This pattern indicates a **violation of the homoscedasticity assumption**, meaning the **variance of the residuals is not constant**. In particular, the model tends to make more **variable predictions when few workers are employed**, and more **consistent predictions as the workforce increases**.

In summary, the **linearity assumption seems plausible**, but the plot reveals a **potential issue with non-constant variance** (heteroscedasticity), which might affect the reliability of confidence intervals and hypothesis tests.

```
library(scatterplot3d)

s3d <- scatterplot3d(
  x = agriculture$extension,
  y = agriculture$workers,
  z = agriculture$quantity,
  type = "h",
  xlab = "Extension (hectares)",
  ylab = "Workers",
  zlab = "Quantity (tonnes)",
  highlight.3d = TRUE,
  col.axis = "gray34",
  col.grid = "gray",
  angle = 120,
```

```
  main = "Estimated plane",
  pch = 20
)

s3d$points3d(
  x = mean(agriculture$extension),
  y = mean(agriculture$workers),
  z = mean(agriculture$quantity),
  col = "orange",
  type = "h",
  pch = 2
)

s3d$plane3d(model15, col = "red", draw_lines = TRUE, draw_polygon = TRUE)
```



4.4 Open questions

- **Write the extended formula of the multiple linear regression model and comment on each component.** The general form of a multiple linear regression model is:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + \varepsilon_i$$

Where: - $\hat{Y}_i$ is the **response variable** for unit i. It is assumed to be continuous and normally distributed conditional on the explanatory variables. - X_{ij} are the **explanatory (independent) variables** or covariates

for unit i. - β_0 is the **intercept**, representing the expected value of Y when all covariates are zero. - β_j are the **regression coefficients**, each quantifying the **partial effect** of the corresponding variable X_j on Y, holding all other covariates constant. - ε_i is the **random error term**, assumed to be independent across observations, with:

$$\mathbb{E}[\varepsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2$$

The model assumes a **linear and additive** relationship between the covariates and the expected value of the response variable. Each coefficient β_j represents the marginal effect of variable X_j on the response, adjusted for the presence of other variables. The **error term** ε_i captures variability in Y_i that cannot be explained by the linear combination of covariates. The model is **parametric**, interpretable, and estimated typically using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals.

- **Write the multiple linear regression model in matrix notation and specify the dimension of each component.**

$$\hat{Y} = X\beta + \varepsilon$$

$\hat{Y}$ is the **response vector**,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

where n is the number of observations.

X is the **design matrix of covariates**,

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

The first column of ones accounts for the intercept.

β is the **vector of regression coefficients**,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^{(p+1) \times 1}$$

ε is the **vector of random errors**,

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

assumed to satisfy: $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \sigma^2 I_n$

This compact formulation is particularly useful for theoretical derivations (e.g., least squares solution: $\hat{\beta} = (X^T X)^{-1} X^T Y$) and computational efficiency.

- **Specify the assumptions required for the multiple linear regression model.** The **multiple linear regression model** relies on a set of core **assumptions** that ensure the validity of estimation, inference, and prediction. These assumptions apply to the error terms ε_i and to the structure of the data:

1. **Linearity**
   The relationship between the response variable \$Y \$ and the explanatory variables $X_1, X_2, \ldots, X_p$ is assumed to be **linear** in parameters:
   $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$
2. **Independence of Errors**
   The error terms ε_i are assumed to be **independent** across observations:
   $$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{for } i \neq j$$
3. **Homoscedasticity (Constant Variance)**
   The variance of the errors is **constant** for all observations:
   $$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{for all } i$$
4. **Normality of Errors (for inference)**
   The error terms ε_i are **normally distributed**:
   $$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
   This assumption is required to perform valid hypothesis testing and construct confidence intervals.
5. **No Perfect Multicollinearity**
   The explanatory variables must not be perfectly linearly related. The design matrix $\mathbf{X}$ must be of **full rank**, so that:
   $$\text{rank}(\mathbf{X}) = p + 1$$
   This ensures the uniqueness and existence of the least squares estimates.
6. **Exogeneity (Uncorrelated errors and covariates)**
   The errors ε_i are assumed to be **uncorrelated with the explanatory variables**:
   $$\mathbb{E}[\varepsilon_i | X_{i1}, X_{i2}, \ldots, X_{ip}] = 0$$

These assumptions are essential for ensuring the **unbiasedness**, **efficiency**, and **consistency** of the estimated regression coefficients, as well as the **validity of statistical inference**.

- **Describe the estimation method employed in the classical linear regression** In the classical linear regression model, the **estimation method** used is the **Ordinary Least Squares (OLS)**, which determines the values of the regression coefficients $\hat{\beta}$ by minimizing the sum of squared residuals, that is, the distance between observed and predicted values:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

In **matrix form**, the estimator is:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This method provides **unbiased** and **efficient** estimates under the classical assumptions, and the residual variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

which is used to compute standard errors and perform inference.

- **In which way can we decompose the variability of the response?** The **variability of the response variable** $ Y $ in a multiple linear regression model can be **decomposed** using the **total sum of squares** into two components:

$$\underbrace{SST}_{\text{Total Sum of Squares}} = \underbrace{SSR}_{\text{Regression (explained)}} + \underbrace{SSE}_{\text{Error (unexplained)}}$$

This decomposition reflects that:

- **SST** measures the total variation of Y around its mean:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- **SSR** is the variation **explained by the model**:

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- **SSE** is the **residual variation**:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

This leads to the **coefficient of determination** $R^2 = \frac{SSR}{SST}$, which quantifies the **proportion of variance explained** by the model.

- **Which is the least squares estimator of the intercept? How is it obtained?**
  The **least squares estimator of the intercept** $\hat{\beta}_0$ is obtained by minimizing the sum of squared residuals in the multiple linear regression model. It is given by the formula:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_p\bar{x}_p$$

This means that the intercept equals the mean of the response variable minus the weighted sum of the means of each covariate, where the weights are the corresponding estimated regression coefficients. It ensures that the **regression hyperplane passes through the centroid** of the data.

- **How do we interpret the values of the estimated regression coefficients for the covariates?**
  The **estimated regression coefficients** for the covariates in a multiple linear regression model represent the **partial effects** of each explanatory variable on the response variable. Specifically, each coefficient $\hat{\beta}_j$ (for $j = 1, \ldots, p$) indicates the **expected change in the response variable** Y for a **one-unit increase** in the covariate X_j, **holding all other covariates constant**. This interpretation is **conditional**, meaning the effect of a covariate is adjusted for the influence of the others included in the model.

- **What is the multiple R-squared and the adjusted R-squared? Why R-squared is adjusted?**
  The **multiple R-squared** measures the proportion of the variance in the response variable that is **explained by the covariates** in the regression model. It is defined as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where RSS is the residual sum of squares and TSS is the total sum of squares. However, **R-squared always increases** as more variables are added, even if they have little explanatory power.

The **adjusted R-squared** corrects for this by penalizing the inclusion of irrelevant predictors. It is defined as

$$\bar{R}^2 = 1 - \left(\frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} \right)$$

where n is the number of observations and p the number of predictors. It provides a more **reliable measure of model fit** when comparing models with different numbers of covariates, as it only increases if the new variable improves the model more than expected by chance.

- **What is the residual standard error, and how is it interpreted?**
  The **residual standard error (RSE)** is an estimate of the standard deviation of the error term ε in a linear regression model. It measures the average **distance between the observed values and the values predicted** by the model. Formally, it is computed as:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-p-1}}$$

where **RSS** is the residual sum of squares, n is the number of observations, and p is the number of predictors. A smaller RSE indicates that the model's predictions are, on average, closer to the actual data points. It is expressed in the same units as the response variable and helps assess the **overall accuracy of the model**.

- **How are defined the predicted (fitted values)? And the residuals?**
  In a multiple linear regression model, the **predicted values** (also called **fitted values**) represent the estimates of the response variable based on the regression equation. For each observation i, the predicted value is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

These values lie on the estimated regression hyperplane.

The **residuals** are the differences between the actual observed values y_i and the predicted values $\hat{y}_i$, and represent the **unexplained part of the variation** by the model:

$$e_i = y_i - \hat{y}_i$$

Residuals are essential for diagnosing the model's fit, checking assumptions like linearity and homoscedasticity, and identifying outliers or influential points.

- **Specify the properties of the residuals and explain how we expect the scatterplot of the residuals by unit number when the assumptions of the model are met.**
  The **residuals** $e_i = y_i - \widehat{y}_i$ in a multiple linear regression model are expected to satisfy several properties when the model assumptions are met:

1. **Zero mean**: The average of the residuals is zero, i.e., $\sum_{i=1}^{n} e_i = 0$.

2. **Independence**: Residuals are assumed to be independent of one another.

3. **Homoscedasticity**: The variance of the residuals is constant across all levels of the covariates (no funnel-shaped patterns).

4. **No autocorrelation**: There should be no systematic pattern or trend in residuals over time or ordered observations.

5. **No correlation with predictors**: Residuals should be uncorrelated with the covariates in the model.

6. **Normality (for inference)**: Residuals are assumed to be normally distributed if we are interested in constructing confidence intervals or performing hypothesis testing.

When these assumptions are met, the **scatterplot of residuals by unit number (observation index)** should appear as a random cloud of points centered around zero, without **any visible pattern, trend, or structure**. This randomness suggests that the model correctly captures the systematic variation in the response, and what remains is only unsystematic noise.

4.5 Multiple linear regression on body fat: univariate and multivariate analysis, correlation structure, multicollinearity, residuals, and inference on coefficients.

Data in `body.Rdata` are related to 248 men (see Johnson, R. (1996), Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4: 265-266)
Each man's percentage of body fat (*brozek*) was accurately estimated by an underwater weighing technique using Brozek's equation.

We like to explain body fat (brozek) according to several body measurements related to the following covariates:

- *age*: Age in years,
- *weight*: Weight (lbs),
- *height*: Height (in inches),
- *adipos*: Adipose index (kg/m^2),
- *neck*: Neck circumference (in cm),
- *abdom*: Abdomen circumference (cm) at the umbilicus and level with the iliac crest (in cm),
- *hip*: Hip circumference (in cm),
- *thigh*: Thigh circumference (in cm),
- *wrist*: Wrist circumference distal to the styloid processes (in cm).

```r
load("body.Rdata")
skim_without_charts(body)
```

4.5.0.1 16.1 Describe the observations first commenting on univariate descriptive measures and then using scatterplot matrices.

Table 13: Data summary

Name	body
Number of rows	248
Number of columns	10
Column type frequency:	
numeric	10
Group variables	None

Variable type: numeric

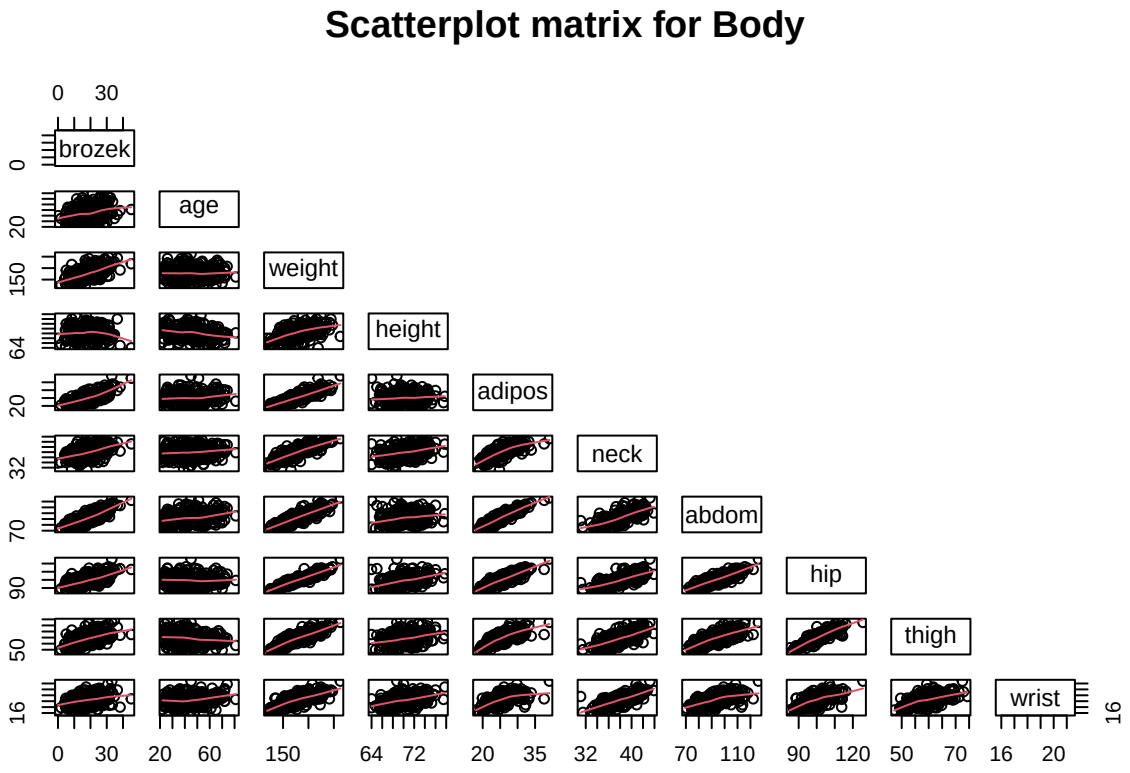
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
brozek	0	1	18.83	7.69	0.0	12.80	19.00	24.50	45.10
age	0	1	44.85	12.60	22.0	35.75	43.00	54.00	81.00
weight	0	1	178.11	27.13	118.5	158.19	176.12	196.81	262.75
height	0	1	70.30	2.61	64.0	68.25	70.00	72.25	77.75
adipos	0	1	25.33	3.34	18.1	23.10	24.95	27.30	39.10
neck	0	1	37.95	2.30	31.1	36.38	38.00	39.42	43.90
abdom	0	1	92.31	10.24	69.4	84.47	90.95	99.20	126.20
hip	0	1	99.66	6.47	85.0	95.47	99.30	103.28	125.60
thigh	0	1	59.27	4.92	47.2	56.00	59.00	62.30	74.40
wrist	0	1	18.22	0.92	15.8	17.60	18.30	18.80	21.40

The table summarizes key **univariate descriptive statistics** for 10 continuous variables related to the body measurements of 248 men.

- The **response variable**, `brozek` (body fat percentage), has a **mean of 18.83** with a **standard deviation (SD) of 7.69**, and ranges from **0 to 45.1**, suggesting moderate variability and the presence of individuals with very low and high fat percentages.
- **Age** varies widely (22 to 81 years), with a **mean of 44.85** and **SD = 12.6**, indicating a diverse adult population.
- **Weight** shows substantial variability (**SD = 27.13**, min 118.5 to max 262.75 lbs), while **height** is more tightly distributed (**mean = 70.3 inches**, **SD = 2.61**), likely reflecting the typical male population.
- **Adipos** (adiposity index) has a mean of **25.33**, which is within the overweight BMI range.
- Circumference-based measures (`neck`, `abdomen`, `hip`, `thigh`, `wrist`) show different levels of dispersion:
  - `abdomen` appears most variable (**mean = 92.31 cm**, **SD = 10.24 cm**), likely due to its strong link with central fat.
  - `wrist` is the least variable (**mean = 18.22 cm**, **SD = 0.92 cm**), which is expected as wrist size changes little across individuals.

Overall, the dataset is **complete (0 missing values)**, and the descriptive statistics suggest **sufficient variability** to allow effective modeling of body fat using the listed covariates.

```
plot(body,
     upper.panel = NULL,
     panel = panel.smooth,
     main = "Scatterplot matrix for Body")
```



The scatterplot matrix reveals **strong linear associations** among several variables, particularly between `brozek` (body fat) and other body measurements:

- **Brozek vs Abdomen** shows a **very strong positive linear relationship**, indicating abdomen circumference is a key predictor of body fat.
- `Adipos`, `hip`, and `weight` also exhibit **positive correlations** with `brozek`, though slightly weaker than `abdom`.
- Variables like `neck` and `thigh` show **moderate positive associations** with `brozek`.
- `Height` appears to have a **weaker or no clear linear trend** with `brozek`, possibly contributing little explanatory power.
- **Multicollinearity** is suggested by strong pairwise correlations among covariates themselves (e.g., `weight` with `hip`, `adipos`, `abdomen`), which may impact model estimation.

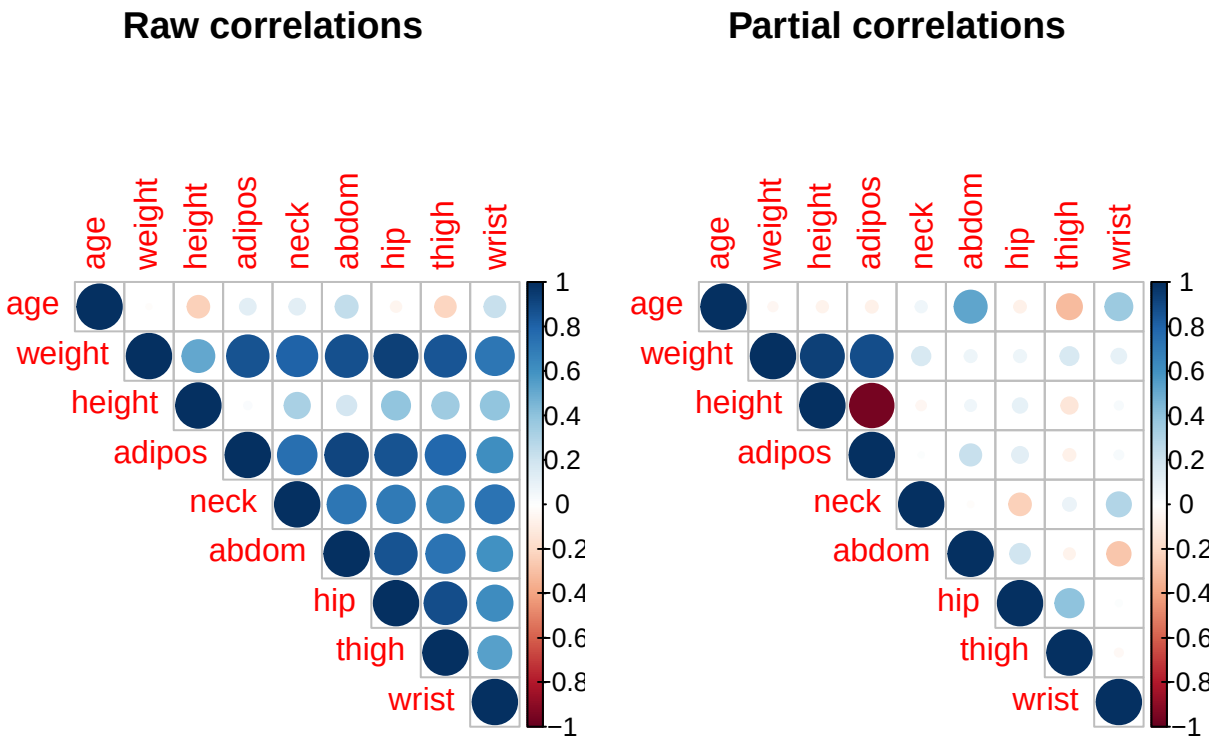
Overall, the scatterplot matrix supports the idea of fitting a **multiple linear regression** model, with a focus on `abdomen` and similar circumferences as the most promising predictors of body fat.

```
library(ggm)
library(corrplot)

covariates <- body[, c("age", "weight", "height", "adipos", "neck",
                       "abdom", "hip", "thigh", "wrist")]

S16 <- cov(covariates)

C1_16 <- parcor(S16)
round(C1_16, 3)
```

4.5.0.2 16.2 Depict the corrplot of raw and partial correlations

```
##           age weight height adipos   neck  abdom    hip  thigh  wrist
## age     1.000 -0.043 -0.065 -0.072  0.065  0.521 -0.071 -0.315  0.367
## weight -0.043  1.000  0.932  0.883  0.169  0.073  0.075  0.164  0.109
## height -0.065  0.932  1.000 -0.953 -0.047  0.061  0.105 -0.134  0.034
## adipos -0.072  0.883 -0.953  1.000  0.014  0.227  0.129 -0.072  0.040
## neck    0.065  0.169 -0.047  0.014  1.000 -0.017 -0.238  0.085  0.293
## abdom   0.521  0.073  0.061  0.227 -0.017  1.000  0.196 -0.061 -0.276
## hip    -0.071  0.075  0.105  0.129 -0.238  0.196  1.000  0.402  0.020
## thigh  -0.315  0.164 -0.134 -0.072  0.085 -0.061  0.402  1.000 -0.033
## wrist   0.367  0.109  0.034  0.040  0.293 -0.276  0.020 -0.033  1.000
```

```
C16 <- cor(covariates)

par(mfrow = c(1, 2))
corrplot(C16,
         type = "upper",
         title = "Raw correlations",
         mar = c(0, 0, 2, 0))
corrplot(C1_16,
         type = "upper",
         title = "Partial correlations",
         mar = c(0, 0, 2, 0))
```



The **Raw correlations** matrix (left) shows strong positive relationships among most body measurements, especially between variables like **weight**, **adipos**, **abdom**, **hip**, and **thigh**, which are all closely related to body size. These high correlations suggest **potential multicollinearity** if used together in a regression model.

In contrast, the **Partial correlations** matrix (right) reveals the *direct* associations between each pair of variables **after controlling for all the others**. Here, many of the previously strong correlations fade, indicating that much of the correlation observed in the raw matrix is mediated through other variables. For example, the partial correlation between **weight** and **height** remains strong, while **adipos** and **height** become negatively correlated, implying an adjusted inverse relationship.

This comparison highlights how **partial correlation isolates unique relationships**, helping identify which variables truly contribute independent information, which is especially useful for model building and interpretation.

```
model16 <- lm(brozek ~ age + weight + height + adipos + neck + abdom + hip + thigh + wrist, data = body
summary(model16)
```

4.5.0.3 16.3 Fit a multiple linear regression model to explain fat vs all the other variables. Comment on the summary of the residuals and on the estimated standard errors.

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + adipos + neck +
```

```
##     abdom + hip + thigh + wrist, data = body)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5420 -2.6465 -0.3223  2.9877  9.6408
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.12907   36.65618  -0.849  0.39661
## age           0.05273    0.02883   1.829  0.06860 .
## weight       -0.10537    0.10130  -1.040  0.29929
## height        0.25051    0.50616   0.495  0.62112
## adipos        0.63928    0.70057   0.913  0.36242
## neck         -0.28074    0.21104  -1.330  0.18470
## abdom         0.77469    0.08195   9.453  < 2e-16 ***
## hip          -0.18075    0.13311  -1.358  0.17580
## thigh         0.25233    0.12283   2.054  0.04105 *
## wrist        -1.38572    0.46968  -2.950  0.00349 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.938 on 238 degrees of freedom
## Multiple R-squared:  0.7472, Adjusted R-squared:  0.7376
## F-statistic: 78.16 on 9 and 238 DF,  p-value: < 2.2e-16
```

The residuals range from approximately -9.54 to +9.64, with a median close to zero (–0.32), and the interquartile range (IQR) is symmetric. This indicates that the model does not show strong skewness or bias in the errors. The **residual standard error is 3.938**, meaning that the typical deviation between the observed and fitted values is about 3.94 units of body fat percentage.

- **Standard errors** vary across predictors. A few of them (like `weight`, `height`, `adipos`) have relatively **large standard errors** compared to their coefficients, suggesting **imprecision** in their estimates — likely due to **multicollinearity**, which was also evident in the previous correlation analysis.
- **Significant predictors**:
  - `abdom` is **highly significant** ($p < 2 \times 10^{-16}$), suggesting it is the most powerful predictor of body fat percentage.
  - `wrist` and `thigh` are also **statistically significant** at conventional levels, with `wrist` showing a negative relationship.
  - `age` is marginally significant ($p = 0.068$), and all others (e.g., `weight`, `height`, `adipos`) are **not significant**, indicating they do not contribute meaningfully **once the others are accounted for**.

Overall, the **model fits well** (adjusted $R^2 = 0.7376$), explaining about **74%** of the variance in body fat, but **some predictors show collinearity and weak individual contributions**, which is reflected in their large standard errors and non-significant p-values.

4.5.0.4 16.4 Is the t-test of the regression coefficient related to weight based on the ratio between the estimated value and its estimated standard error? Provide a confidence interval for it doing yourself the calculation.

Yes, the **t-test for the regression coefficient of `weight`** is indeed based on the ratio between the **estimated coefficient** and its **standard error**. This test evaluates the null hypothesis $H_0 : \beta_{\text{weight}} = 0$ using the statistic:

$$t = \frac{\hat{\beta}_{\text{weight}}}{\text{SE}(\hat{\beta}_{\text{weight}})} = \frac{-0.10537}{0.10130} \approx -1.040$$

This matches the t-value shown in the regression output.

To **manually compute a 95% confidence interval** for the weight coefficient, we use:

$$\hat{\beta} \pm t_{n-k,0.975} \cdot \text{SE}(\hat{\beta})$$

Where: - $\hat{\beta}_{\text{weight}} = -0.10537$
- $\text{SE}(\hat{\beta}) = 0.10130$
- Degrees of freedom $df = 248 - 9 - 1 = 238$

```
-0.10537 + c(-1,1)*qt(0.975, 238)*0.10130
```

```
## [1] -0.30492913  0.09418913
```

$$[-0.305,\ 0.094]$$

Interpretation: Since the 95% confidence interval for the coefficient of `weight` includes zero, we do **not** reject the null hypothesis at the 5% level. This supports the regression output indicating that `weight` is not a statistically significant predictor of body fat when controlling for the other variables.

4.5.0.5 16.5 Discuss on the potential concerns (if any) of multicollinearity.

Yes, multicollinearity is a potential concern in this regression model. Multicollinearity occurs when two or more explanatory variables are highly linearly related, which can inflate the **standard errors** of the estimated coefficients, making it harder to detect statistically significant effects, even when they are meaningful in reality.

In this case, body measurements such as **weight, adipos, abdom, hip, thigh**, and **wrist** are all likely to be **strongly correlated**, as they describe different aspects of body size and shape. This is confirmed by the **high raw correlations** observed in the previous `corrplot`, and by the fact that several variables have **non-significant t-tests**, despite the overall model having a high R^2 **value (0.7472)** and a very significant **F-statistic**.

This pattern — good model fit but few individually significant predictors — is a classical signal of multicollinearity. As a result, the **estimated regression coefficients** become unstable and sensitive to small changes in the data, and their **interpretability decreases**.

To mitigate multicollinearity, one could: - Use **Principal Component Regression (PCR)** or **Ridge Regression**;
- **Remove redundant predictors** or combine them into indices;
- Evaluate **Variance Inflation Factors (VIFs)** to identify the most problematic variables.

```
data.frame(
  unit = c(23, 56, 103),
  observed = body$brozek[c(23, 56, 103)],
  fitted = fitted(model16)[c(23, 56, 103)],
  residual = residuals(model16)[c(23, 56, 103)]
)
```

4.5.0.6 16.6 Calculate and show the residuals for units 23, 56 and 103. The fitted values are similar to the observed values? Can you explain what happens if on average residuals are huge?

```
##     unit observed   fitted  residual
## 23    23     15.7 10.28998  5.410022
## 59    56     30.4 28.62489  1.775105
## 107  103     19.1 25.50873 -6.408728
```

- **Unit 56** has a small residual, indicating a good fit.
- **Units 23 and 103**, however, have **larger absolute residuals** (>5), suggesting that the model **underestimates** for unit 23 and **overestimates** for unit 103.

If residuals are **systematically large**, it may imply: - The model lacks important predictors,
- Functional form is incorrect (non-linearity, interaction effects),
- Or there are **outliers or influential points** driving poor predictions.

Large average residuals reduce **predictive power**, compromise **reliability**, and can **violate regression assumptions**, such as homoscedasticity and normality of errors.

4.6 Linear regression analysis on player data: model equation, coefficient interpretation, residual computation, hypothesis testing, and confidence intervals.

The intention is to explain playing minutes over the entire season of each professional basketball player as a function of certain explanatory variables. The estimated partial regression coefficients of the multiple linear regression model with some variables of interest are reported below.

$$\begin{aligned}
\beta_0 &= 358.848 \\
\beta_1 &= 0.6742 \\
\beta_2 &= 0.2855 \\
\beta_3 &= 303.81 \\
\beta_4 &= 504.95 \\
\beta_5 &= -3923.5 \\
\beta_6 &= 480.04 \\
\beta_7 &= 1350.3 \\
\beta_8 &= 891.67 \\
\beta_9 &= 722.95
\end{aligned}$$

4.6.0.1 17.1 Write the equation of the linear model.

$$\widehat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

4.6.0.2 17.2 Write the equation for the fitted values (y^i).

Substituting the estimated coefficients:

$$\widehat{y}_i = 358.848 + 0.6742 x_1 + 0.2855 x_2 + 303.81 x_3 + 504.95 x_4 - 3923.5 x_5 + 480.04 x_6 + 1350.3 x_7 + 891.67 x_8 + 722.95 x_9$$

4.6.0.3 17.3 Interpret the values of each estimated regression coefficient.

Each estimated regression coefficient $\hat{\beta}_j$ represents the **expected change in total playing minutes** for a one-unit increase in the corresponding variable x_j, **holding all other variables constant**. Here's the interpretation for each:

- $\hat{\beta}_1 = 0.6742$: A 1 percentage point increase in points scored is associated with an increase of about **0.67 minutes** of play.
- $\hat{\beta}_2 = 0.2855$: A 1 percentage point increase in free throw accuracy is associated with an increase of about **0.29 minutes**.
- $\hat{\beta}_3 = 303.81$: An increase of **1 rebound per minute** (a huge improvement) corresponds to an increase of **303.81 minutes** over the season.
- $\hat{\beta}_4 = 504.95$: Each additional **point per minute** is associated with an increase of about **505 minutes**.
- $\hat{\beta}_5 = -3923.5$: Each additional **foul per minute** leads to a **substantial decrease** (–3923.5 minutes), reflecting penalization for fouling.
- $\hat{\beta}_6 = 480.04$: Each additional **stolen ball per minute** increases playing time by **480 minutes**.
- $\hat{\beta}_7 = 1350.3$: Each **recovered ball per minute** is associated with an increase of **1350.3 minutes**, suggesting this is a highly valued skill.
- $\hat{\beta}_8 = 891.67$: Each **lost ball per minute** surprisingly shows a **positive effect**, which might suggest multicollinearity or that lost balls correlate with high usage players.
- $\hat{\beta}_9 = 722.95$: Each **assist per minute** adds about **723 minutes**, highlighting the value of playmaking.

Large coefficients on per-minute stats like rebounds, fouls, and assists suggest that small changes in these rates can imply significant differences in total minutes, because per-minute stats can vary widely across players.

4.6.0.3.1 17.4 The value of index R^2 is 0.5239. What does it mean?

An $ R^2 $ value of **0.5239** means that approximately **52.39% of the variability** in the total playing minutes of the basketball players is **explained by the regression model**, which includes variables such as scoring efficiency, rebounds, assists, and other in-game statistics.

In practical terms, this indicates a **moderate to strong fit**: the model captures a significant portion of the variation in playing time, but about **47.61% of the variation remains unexplained**, possibly due to factors not included in the model (e.g., coach strategy, injuries, player role, or other qualitative aspects).

```
358.848+(0.6742*0.5)+(0.2855*5)+(303.81*2)+(504.95*2)-(3923.5*0.5)+(480.04*0.2)+(1350.3*0.3)+(891.67*0.5
```

4.6.0.4 17.5 Knowing the following values of the covariates for player A calculate the fitted value, and compute the corresponding residual: $x_1 = 0.5$, $x_2 = 5$, $x_3 = 2$, $x_4 = 2$, $x_5 = 0.5$, $x_6 = 0.2$, $x_7 = 0.3$, $x_8 = 0.5$, $x_9 = 0.1$.

```
## [1] 1035.611
```

$$\hat{y}_A = 1035.611$$

To calculate the residual, we would need the observed value y_A. Since it's not provided, we cannot compute the residual directly. However, if the text provided the actual playing minutes for player A, we could compute:

$$z_A = y_A - \hat{y}_A$$

4.6.0.5 17.6 Disposing of the following estimated standard errors for each estimated coefficient: $SE(\hat{\beta}_0) = 44.695$, $SE(\hat{\beta}_1) = 0.0639$, $SE(\hat{\beta}_2) = 0.0388$, $SE(\hat{\beta}_3) = 77.73$, $SE(\hat{\beta}_4) = 43.26$, $SE(\hat{\beta}_5) = 120.6$, $SE(\hat{\beta}_6) = 224.9$, $SE(\hat{\beta}_7) = 212.3$, $SE(\hat{\beta}_8) = 180.87$, $SE(\hat{\beta}_9) = 110.98$. Calculate the value of the t test for testing the statistical hypothesis $H_0 : \beta_k = 0$ fixing a significance level $\alpha = 0.05$. Report and comment on the test result for each regression coefficient.

To test the null hypothesis $H_0 : \beta_k = 0$ for each coefficient, we compute the t-statistic using the formula:

$$t_k = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)}$$

```
beta17 <- c(358.848, 0.6742, 0.2855, 303.81, 504.95, -3923.5, 480.04, 1350.3, 891.67, 722.95)

se17 <- c(44.695, 0.0639, 0.0388, 77.73, 43.26, 120.6, 224.9, 212.3, 180.87, 110.98)

names(beta17) <- c("Intercept", "x1", "x2", "x3", "x4", "x5", "x6", "x7", "x8", "x9")

t_values17 <- beta17 / se17

round(t_values17, 2)
```

```
## Intercept        x1        x2        x3        x4        x5        x6        x7
##      8.03     10.55      7.36      3.91     11.67    -32.53      2.13      6.36
##        x8        x9
##      4.93      6.51
```

The critical value for a two-tailed t-test at significance level $ = 0.05 $ with a large sample (df $\approx \infty$) is approximately 1.96.

- All coefficients have $ |t_k| > 1.96 $, meaning they are **statistically significant at the 5% level**, except possibly $ _6 $, which is borderline with $ t 2.13 $.
- The strongest evidence comes from $ _5 , with a very large absolute $t - value$(-32.53 $), confirming a **strong negative effect** of fouls per minute on playing time.
- All other predictors (e.g., assists, rebounds, points per minute) also show significant positive associations with playing time.

These results suggest that most of the variables contribute **significantly** to the prediction of minutes played.

4.6.0.6 17.7 Assess the assumptions required to determine a confidence interval for each regression coefficient and considering a confidence level of 0.95. Report the lower and upper bounds of the confidence intervals for each regression coefficient.

To compute the **95% confidence intervals** for each regression coefficient, we need to:

1. Assume that the classical linear regression assumptions hold:
   - The model is **correctly specified** (linearity).
   - The error terms are **independent**, **normally distributed**, and have **constant variance**.
   - There is **no perfect multicollinearity** among covariates.
2. Use the formula for the confidence interval:

$$\hat{\beta}_k \pm t_{n-p,\,\alpha/2} \cdot SE(\hat{\beta}_k)$$

Where:

- $\hat{\beta}_k$ is the estimated coefficient,
- $SE(\hat{\beta}_k)$ is the standard error of the coefficient,
- $t_{n-p,\,\alpha/2}$ is the quantile from the Student's t-distribution with $n-p$ degrees of freedom,
- $n = 2679$, $p = 10 \rightarrow df = 2669$,
- For $\alpha = 0.05$, $t_{0.975,2669} \approx 1.960$.

```
t_crit17 <- qt(0.975, df = 2669)

ci_lower17 <- beta17 - t_crit17 * se17
ci_upper17 <- beta17 + t_crit17 * se17

conf_int17 <- data.frame(
  Coefficient = names(beta17),
  Estimate = round(beta17, 3),
  SE = round(se17, 3),
  CI_Lower = round(ci_lower17, 3),
  CI_Upper = round(ci_upper17, 3)
)

conf_int17
```

```
##           Coefficient  Estimate      SE  CI_Lower  CI_Upper
## Intercept   Intercept   358.848  44.695   271.208   446.488
## x1                 x1     0.674   0.064     0.549     0.799
## x2                 x2     0.286   0.039     0.209     0.362
## x3                 x3   303.810  77.730   151.393   456.227
## x4                 x4   504.950  43.260   420.123   589.777
## x5                 x5 -3923.500 120.600 -4159.979 -3687.021
## x6                 x6   480.040 224.900    39.044   921.036
## x7                 x7  1350.300 212.300   934.011  1766.589
## x8                 x8   891.670 180.870   537.010  1246.330
## x9                 x9   722.950 110.980   505.335   940.565
```

- All coefficients have confidence intervals **excluding 0**, which suggests they are **statistically significant** at the 5% level. Even the lowest bound of each interval is clearly above or below zero.
- The **intercept** lies between 271 and 446, meaning that even a hypothetical player with all zero stats would be expected to play a positive number of minutes, though this has limited interpretability in practice.
- The **most influential variables** in absolute terms are:
  - `x5` (fouls per minute), with a **strong negative** effect: each additional foul per minute reduces playing time drastically (CI: -4159.98 to -3687.02).
  - `x7` (recovered balls), `x8` (lost balls), and `x9` (assists), all with large **positive effects** on minutes played, indicating their importance in coach decisions.

Moreover: - The **narrow intervals** for variables like `x1` (points %) and `x2` (free throws %) reflect **precise estimates** due to likely lower standard errors. - The wider intervals for variables like `x6` (stolen balls) and `x7` indicate **more uncertainty**, possibly due to more variability or multicollinearity with other predictors.

In conclusion, all predictors appear to contribute significantly to explaining variation in playing time, and the model seems robust with reasonably tight confidence bounds.

4.7 Model diagnostics and selection: variance estimate, F-test, residual plots, multicollinearity via VIF, AIC-based selection, and inference on final model coefficients.

Consider the `ratings.Rdata` analysed in a previous exercise. Explain ratings of the firms according to profitability, capital structure and financial flexibility.

4.7.0.1 18.1 Report and comment on the estimate of the variance (σ^2).

We estimate the multiple linear regression model where the variable rating is explained with respect to profitability, capital structure, and financial flexibility.

```
load("ratings.Rdata")
```

We look at the estimate of the variance $\hat{\sigma}^2$ from the regression model fitted to explain **ratings** as a function of **profit**, **capital** and **flex**.

```
model18 <- lm(rating ~ ., data = ratings)

summary(model18)
```

```
##
## Call:
## lm(formula = rating ~ ., data = ratings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## profit        0.3277     4.4598   0.073  0.94198
## capital       3.9118     1.2484   3.133  0.00425 **
## f_flex       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

In this regression model, we are explaining the **firm rating** as a function of **profitability**, **capital structure**, and **financial flexibility**. The **residual standard error (RSE)** is extracted from the model summary and used to compute the **estimated residual variance**. As requested by the exercise, we focus on the estimated variance. The summary table obtained through the `summary()` function reports a value for the residual standard error equal to $\hat{S} = 10.13$ (Residual standard error). This represents the (unbiased) estimate for the standard deviation. Therefore, it is sufficient to square it to obtain the estimate for the variance:

$$\hat{\sigma}^2 = \mathrm{S}^2 = 10.13^2 = 102.63$$

The corresponding number of degrees of freedom is 26, computed as the sample size minus the number of regression coefficients (i.e. number of covariates p plus one for the intercept if included in the model).

Summarizing, we have assumed that the model errors are distributed as standard normal random variables, $\varepsilon_i \sim N(0, \sigma^2)$, or in a matrix notation $\sim N(0, \sigma^2 \cdot I)$, and the estimate of σ^2 is given by 102.62. This value also represents the estimated variance for the response variable Y according to the considered model.

The **estimated variance** $\hat{\sigma}^2 = 102.63$ reflects the **average squared deviation** between the observed firm ratings and the values predicted by the regression model. A value of around 100 indicates that, on average, the squared prediction error is high, and there is considerable variability not explained by the model.

4.7.0.2 18.2 Comment on the results of the observed value of F distribution, on the degrees of freedom and on the results of the test.

From the `summary()` function computed above we can also analyze and comment on the results of the F test having as null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ ((i.e. joint equality to zero of all regression coefficients except for the intercept). Note that the **null hypothesis** states that **none of the considered covariates helps to explain the response variable**. The alternative hypothesis states that at least one covariate helps to explain the response variable (but not necessarily all of them). The value of the F statistic is equal to 16.22, and the degrees of freedom are 3 for the explained component (p), and 26 for the residual (or unexplained) component ($n - p - 1$). The associated p-value is is extremely close to zero that leads us to reject the null hypothesis at each significance level:

$$F = 16.22 \quad \text{on} \quad (3, 26) \text{ degrees of freedom}, \quad \text{p-value} = 3.81 \times 10^{-6}$$

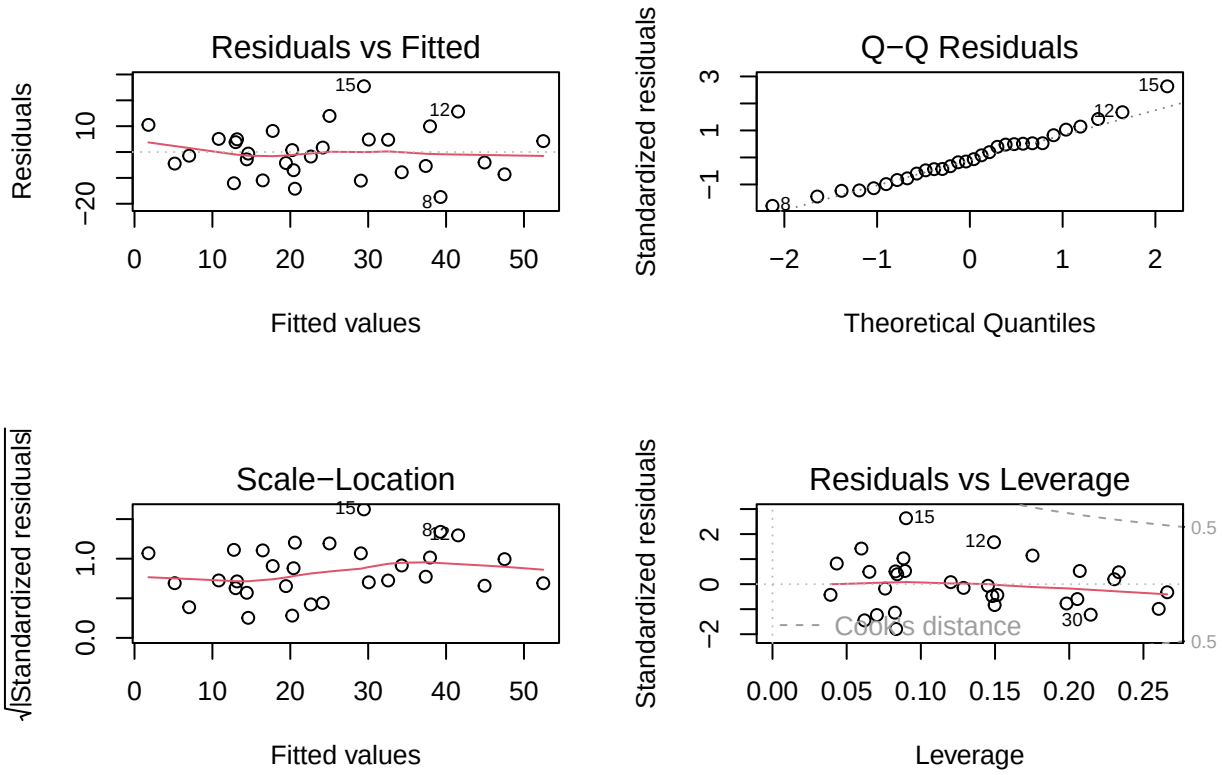
Given the **p-value « 0.05**, we **strongly reject the null hypothesis**, indicating that the model as a whole is **statistically significant**. That means at least one of the covariates significantly contributes to explaining the variability in ratings.

Talking about the degrees of freedom, we can say:

- **Numerator df = 3**: number of predictors.
- **Denominator df = 26**: total observations (30) minus the number of parameters (intercept + 3 predictors). The formula is $n - p - 1$ (here, $p = 3$ predictors), so $30 - 3 - 1 = 26$.

```
par(mfrow = c(2,2))
plot(model18)
```

4.7.0.3 18.3 Show all the diagnostic plots with the residuals. Comment on each of them.



1. **Residuals vs Fitted**

- **Purpose**: Detect non-linearity or non-constant variance.
- **Observation**: Residuals appear roughly randomly scattered around zero, but a **slight curvature** may suggest some **mild non-linearity**. However, no strong funnel shape is visible, indicating **acceptable homoscedasticity**.
- **Conclusion**: The linear model seems reasonably appropriate, though slight model misfit can't be ruled out.

2. **Normal Q-Q Plot**

- **Purpose**: Assess normality of residuals.
- **Observation**: Most points lie along the reference line, except **at the upper tail**, where a few points (e.g., unit 150) deviate.
- **Conclusion**: **Residuals are approximately normal**, with **minor deviations** that are not critical for inference in large samples, but should be noted.

3. **Scale-Location Plot**

- **Purpose**: Check for constant variance (homoscedasticity).
- **Observation**: The spread of residuals is relatively even, though **slight increases in spread** at higher fitted values are visible.
- **Conclusion**: **Homoscedasticity holds approximately**, though a transformation might slightly improve it.

4. **Residuals vs Leverage**

- **Purpose**: Identify influential observations.
- **Observation**: No points fall near or outside the Cook's distance lines, though unit **150 and 120 have higher leverage**.
- **Conclusion**: **No influential observations** are strongly distorting the model, though those with high leverage (like unit 150) may warrant individual review.

4.7.0.4 18.3-bis Report and comment on the values of the variance inflation factor of each covariate. The Variance Inflation Factor (VIF) allows us to measure the presence of (potentially excessive) collinearity among the explanatory variables included in a multiple linear regression model. Before proceeding with the actual calculation of the index, let us briefly recall that an indication of collinearity could already emerge from the analysis of raw and partial correlation matrices, when one of the coefficients between two different explanatory variables was particularly high (positive or negative). At this point, the VIF index allows us to determine and quantify in a much more precise way the actual presence of linearly correlated variables. To calculate the index, we use the `vif()` function, available in the `faraway` library.

```
require(faraway)
```

```
## Loading required package: faraway
```

```
##
## Attaching package: 'faraway'
```

```
## The following object is masked from 'package:GGally':
##
##     happy
```

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

```
vif(model18)
```

```
##   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 ($\text{VIF}_i = \frac{1}{1-R_i^2}$). 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.

4.7.0.5 18.4 Perform model selection using the Akaike information criterion. Comment explaining the procedure and the results of each selection step. Which is the variable providing more information?

We perform model selection using the AIC index:

$$AIC = -2 \cdot \ell(\theta) + 2 \cdot \#\text{par}$$

The `step()` function automatically calculates the index for a series of different models, comparing them, and then selects the best model, i.e. the one corresponding to the lowest AIC index. By default (`direction = both`), the `step()` function performs both forward selection and backward elimination. Starting with the provided model (in our case, the complete model), the function tries to add, one at a time, each of the not-included variables, and to eliminate, one at a time, each of the included variables. The (p) new models are sorted and the best one is selected as the new reference model. The whole procedure is then iterated until no further modification improves the AIC value.

```
step(model18)
```

```
## 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
```

The stepwise model selection using AIC starts with the complete model and the AIC index value is 142.64.

$$\texttt{rating} \sim \texttt{profit} + \texttt{capital} + \texttt{flex} \quad (\text{AIC} = 142.64)$$

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.

In the **second step**, the reduced model is:

$$\texttt{rating} \sim \texttt{capital} + \texttt{flex} \quad (\text{AIC} = 140.65)$$

Eliminating the other two variables does not result in any improvement. At this point, the new reference model includes only the `capital` and `flex` covariates. The procedure 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.

The final selected model is:

$$\texttt{rating} = \beta_0 + \beta_1 \cdot \texttt{capital} + \beta_2 \cdot \texttt{flex}$$

- The variable **`capital` provides the most information**, as seen by the largest increase in AIC when it's removed (**+9.09**). - **`flex`** is also informative (removal increases AIC by **+4.24**), but to a **lesser extent** than `capital`. - The variable **`profit` is redundant**, showing almost no reduction in residual sum of squares and no predictive contribution (only 0.55 variance explained).

This result aligns with both the **individual t-tests** in the model summary (where `profit` had $p = 0.94$) and the **AIC-based tradeoff** between model complexity and explanatory power.

In conclusion, capital structure and financial flexibility are key determinants of firm ratings in this dataset, while profitability does not add relevant information in this linear specification.

4.7.0.6 18.5 Report and interprete the estimated regression coefficients according to the model selected at the previous point. All the t-tests are statistically significant? Which are the confidence intervals for the estimated parameters?

We estimate the model selected at the previous point and comment on the partial regression coefficients.

```
model18_selected <- lm(rating ~ capital + f_flex, data = ratings)

summary(model18_selected)
```

```
##
## Call:
## lm(formula = rating ~ capital + f_flex, data = ratings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.343  -6.530  -1.164   4.844  25.618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.592      8.982  -3.072  0.00481 **
## capital        3.946      1.136   3.475  0.00174 **
## f_flex        19.887      7.959   2.499  0.01885 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.942 on 27 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6259
## F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

```
round(confint(model18_selected), 3)
```

```
##               2.5 % 97.5 %
## (Intercept) -46.021 -9.163
## capital       1.616  6.277
## f_flex        3.557 36.218
```

The fitted model

$$\widehat{\text{rating}} = -27.59 + 3.95 \cdot \text{capital} + 19.89 \cdot \text{f\_flex}$$

shows that both predictors have a statistically significant and positive effect on the firm rating. The coefficient for `capital` means that a one-unit increase in capital leads to an average increase of 3.95 in the rating. Similarly, the coefficient for `flex` indicates that, holding capital fixed, a one-unit increase in financial flexibility results in an average rating increase of 19.89. Both effects are statistically significant with p-values below 0.05 (0.00174 and 0.01885, respectively), meaning the null hypothesis that these coefficients are zero can be rejected at conventional significance levels. The intercept, while not directly interpretable due to the lack of practical meaning of a firm having zero capital and flexibility, serves to align the model and is also significant ($p = 0.00481$).

The 95% confidence intervals reinforce these findings: for `capital`, the interval is [1.62, 6.28] and for `flex`, it is [3.56, 36.22]. Since neither includes zero, we are reasonably confident that both predictors have a real, non-null effect on ratings. Capital appears more precisely estimated due to its narrower interval, while flex shows a larger, though more variable, impact. Overall, this model confirms that both capital structure and financial flexibility play an important role in determining firm ratings, and their effects are both statistically and economically significant.

4.8 Exploratory and regression analysis on savings data: correlations, model fitting, AIC-based selection, diagnostics, residuals, and coefficient interpretation.

Consider the data in the library `faraway` named `savings` referred to 50 different countries. These data are averages from 1960 to 1970 (to remove business cycles or other short-term fluctuations). Consider only the following:

- `sr`: aggregate personal saving divided by disposable income.
- `dpi`: per capita disposable income in U.S. dollars;
- `pop15`: is the percentage of population under 15 (pop15).

4.8.0.1 19.1 Explore the data using descriptive statistics. Comment on the results.

We preliminary load the data and select the required variables. The resulting dataframe has 3 variables and 50 observations. We inspect the data by computing and commenting on the main descriptive statistics through the `skimr` package.

```
require(faraway)

data(savings)
dat <- savings[, c(1, 2, 4)]

my_skim <- skim_with(base = sfl(),
                     numeric = sfl(hist = NULL, cv = EnvStats::cv))
my_skim(dat)
```

Table 15: Data summary

Name	dat
Number of rows	50
Number of columns	3

Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

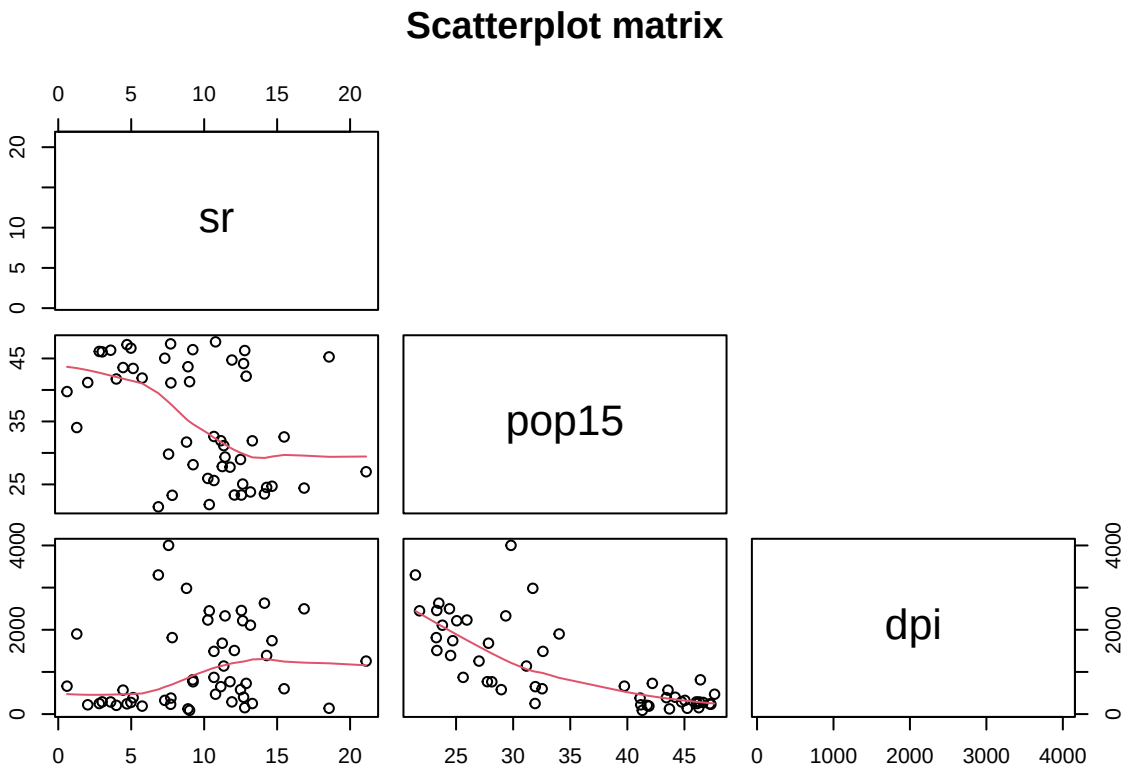
skim_variable	mean	sd	p0	p25	p50	p75	p100	cv
sr	9.67	4.48	0.60	6.97	10.51	12.62	21.10	0.46
pop15	35.09	9.15	21.44	26.22	32.58	44.06	47.64	0.26
dpi	1106.76	990.87	88.94	288.21	695.66	1795.62	4001.89	0.90

The saving rate, defined as saving divided by income is on average equal to 9.67: the minimum and maximum values are quite far from this central point, being equal to 0.6 and 21.1, respectively. Mean and median are quite similar, and 75% of the observation show a value for the saving rate that is lower than 12.6. Regarding the variability, the value of the standard deviation is quite small: the average variability around the mean is 4.5; the coefficient of variation CV $= \frac{\sigma}{\mu}$ is equal to 0.46, thus meaning that the variability of saving rate is around 46% of the mean. The percentage of population under 15 ranges from 21.4% to 47.6%; the corresponding variability seems to be quite low, as confirmed by the coefficient of variation equal to 0.26. The mean is equal to 35.1%, and the average variability around it is 9.2%. We can also notice that the distance between the median and the third quartile is quite large; on the contrary third quartiles and the maximum assume very close values: the countries having a percentage of population under 15 that is larger than the median tend to assume very high velues (i.e. there are very few countries with such percentage equal between 35% and 40%). Finally, considering the per capita income, we first notice that the range of variation is mach wider than those of the other two variables (it is worth considering dividing the variable by 1000 and having values expressed in thousands of dollars instead). The per capita income ranges from 89 to more than 4000. This reflects the different socio-economic conditions of the different considered countries. As typically happens, this variable is highly asymmetric (the skewness is equal to 0.9), with a heavy right tail: this behavior highlight the presence of a large number of countries with a low/very low per capita income, with few countries having very high incomes. Mean and median are consequently very different: the average per capita income is equal to 1106, but 50% of the considered countries have a value smaller than 696. Moreover, only 25% of countries have an per capita income hogher than 1796. The variability is quite high, as shown by the coefficient of variation (equal to 0.9); the average variability around the mean is equal to 991.

4.8.0.2 19.2 Show the scatter plot matrix with the panel smooth line and comment on each plot.

We inspect the association between each pair of variables through the scatter plot matrix.

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



- Considering the two scatter plots related to the `pop15` variable, 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 `sr` 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.

4.8.0.3 19.3 Report and comment on sample correlations and partial correlations.

Keeping in mind the results obtained at the previous point, we compute now the raw and partial correlation coefficients between the variables.

```
round(cor(dat), 3)
```

```
##           sr  pop15    dpi
## sr     1.000 -0.456  0.220
## pop15 -0.456  1.000 -0.756
## dpi    0.220 -0.756  1.000
```

```
round(parcor(cov(dat)), 3)
```

```
##           sr  pop15    dpi
## sr     1.000 -0.453 -0.213
## pop15 -0.453  1.000 -0.755
## dpi   -0.213 -0.755  1.000
```

- Considering the matrix of raw correlations, the highest coefficients is the one between the `pop15` and `dpi` variables, equal to -0.76. The same exact value is obtained for the partial correlations, highlighting that the interaction of the remaining variable (`sr`) has no effect in modifying the association between these two variables. It is very important to remark that, as shown by the corresponding scatter plot, this association is not linear, so the obtained value must be handled with care.
- A similar outcome and comment is valid for the pair of variables `sr` and `pop15` (even though with smaller values).
- Finally, regarding `sr` and `dpi`, we observe a significant modification between the raw and the partial correlation coefficient: while the former is positive (equal to 0.22, denoting a weak linear association), the latter becomes negative (equal to -0.21): the positive association between the two variables is due to the interaction of the percentage of population under 15; when this effect is ruled out, the two variables assume a negative association (an increase in the saving rate corresponds to an decrease in the per capita income and vice-versa).

4.8.0.4 19.4 Compare the observed values for Italy, Ireland, Japan and Switzerland. Comment. Which is the country showing the highest value of the population under 15 years? And the lowest?

We select the rows corresponding to the four countries we are interested in.

```
dat[c("Italy", "Ireland", "Japan", "Switzerland"), ]
```

```
##                sr pop15     dpi
## Italy       14.28 24.52 1390.00
## Ireland     11.34 31.16 1139.95
## Japan       21.10 27.01 1257.28
## Switzerland 14.13 23.49 2630.96
```

- We observe that Ireland is the country, among the four considered, with the highest percentage of population under 15, with more than 31%. Italy (24.5%) and Switzerland (23.5%), on the contrary, present the lowest values.
- We can also highlight that Switzerland is the Country with the highest per capita income, while Japan has the highest saving rate. For both these two variables Ireland presents the lowest values.

4.8.0.5 19.5 Consider the multiple regression model of savings as a function of other two covariates. Fit the model and calculate the variance inflation factor for each covariate.

We estimate the multiple linear regression model aimed at explaining the saving rate as a function of the remaining two variables. To check for potential collinearity between the two variables we compute the Variance Inflation Factor (VIF). Note that, since the scatter plot shows a non linear association between the two variables, we expect very small values of the VIF.

```
model19 <- lm(sr ~ pop15 + dpi, data = savings)

summary(model19)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + dpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1167 -2.6564 -0.0053  1.4831 10.9760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.7126386  4.1520364   5.470 1.69e-06 ***
## pop15       -0.3303269  0.0949204  -3.480  0.00109 **
## dpi         -0.0013107  0.0008767  -1.495  0.14159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.979 on 47 degrees of freedom
## Multiple R-squared:  0.2435, Adjusted R-squared:  0.2113
## F-statistic: 7.564 on 2 and 47 DF,  p-value: 0.001419
```

```
vif(model19)
```

```
##    pop15      dpi
## 2.335469 2.335469
```

The VIF is computed as the reciprocal of 1 minus the multiple coefficient of determination obtained by excluding the corresponding variable and allows us to measure the presence of (potentially excessive) collinearity among the explanatory variables.

$$\text{VIF}(\hat{\beta}_k) = \frac{1}{1 - R_k^2}$$

Here we obtain a value (the same for both variables) equal to 2.34. This small value (far from 10, which sometimes serves as a threshold) of VIF suggests the absence of multi-collinearity between the two variables. Another interpretation of this value is bound to the standard error of each regression coefficient β_k; in this case we can conclude that the estimated standard error of both $\hat{\beta}_1$ and $\hat{\beta}_2$ is about $\sqrt{2.34} = 1.53$ times greater than the one obtained in absence of collinearity. It is clear that there is no significant difference in our case.

4.8.0.6 19.6 Use the Akaike information Criterion to perform model selection and comment on the results at each step of the procedure.

We use the `step()` function to perform model selection on the basis of the Akaike Information Criterion (AIC). It considers an initial model (the complete model in our case) and proceeds iteratively adding or removing a single variable at a time with the aim of improving (i.e. reducing) the AIC value.

```
step(model19)
```

```
## Start:  AIC=141.01
## sr ~ pop15 + dpi
##
##         Df Sum of Sq    RSS    AIC
```

```
## <none>               744.12 141.01
## - dpi    1    35.387 779.51 141.33
## - pop15  1   191.741 935.87 150.47
```

```
##
## Call:
## lm(formula = sr ~ pop15 + dpi, data = savings)
##
## Coefficients:
## (Intercept)        pop15          dpi
##   22.712639    -0.330327    -0.001311
```

The output shows that the initial complete model, associated to an AIC value equal to 141.01, is the best one: we observe indeed that, by removing the `dpi` variable, the AIC index associated to the new model (slightly) increases to 141.33, while removing the `pop15` variable, the AIC index grows up to 150.47. Therefore, the complete model is the best considering AIC as a selection method. We also observe that the model removing `dpi` is only slightly inferior.

4.8.0.7 19.7 Comment on the residual standard error and on the R^2 and the adjusted R^2.

The multiple R-squared index (along with its adjusted value) measures the ability of the covariates included in the model to explain the total variability of the response variable. Therefore it is computed as the ratio between the variability explained by the model (hereafter denoted as SSE) and the total variability of the response variable (TSS).

```
summary(model19)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + dpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1167 -2.6564 -0.0053  1.4831 10.9760
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.7126386  4.1520364   5.470 1.69e-06 ***
## pop15       -0.3303269  0.0949204  -3.480  0.00109 **
## dpi         -0.0013107  0.0008767  -1.495  0.14159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.979 on 47 degrees of freedom
## Multiple R-squared:  0.2435, Adjusted R-squared:  0.2113
## F-statistic: 7.564 on 2 and 47 DF,  p-value: 0.001419
```

In the fitted model where the saving rate (`sr`) is explained by the percentage of population under 15 (`pop15`) and per capita disposable income (`dpi`), the **residual standard error (RSE)** is 3.979, indicating the average deviation of observed savings from the predicted values is about 4 units. Considering the observed range of the response variable (from 0.6 to 21.1), this value denotes a **moderate residual variability**, suggesting a non-negligible portion of the variance remains unexplained.

The **R-squared** is 0.2435, meaning that approximately 24% of the total variability in the saving rate across countries is explained by the two predictors. This implies that the linear relationship accounts for only a limited share of the observed differences in saving behavior. The **adjusted R-squared**, which penalizes for the number of predictors used, is slightly lower at 0.2113, reinforcing the idea that the model provides a modest fit.

From a statistical testing perspective, the **overall model is significant** (F-statistic = 7.564, p = 0.0014), so we reject the null hypothesis that both slope coefficients are jointly zero. However, looking at the individual contributions, only `pop15` is statistically significant at the 0.1% level ($p = 0.0011$), with a negative coefficient suggesting that higher proportions of young population are associated with lower savings. In contrast, `dpi` has a non-significant effect ($p = 0.14$), indicating no strong evidence of a linear relationship between disposable income and saving rate in this setting.

In conclusion, the model captures a meaningful association between demographics and savings, but leaves a large part of the variability unexplained, possibly due to omitted variables or non-linearities.

4.8.0.8 19.8 Write the equation of the estimated model

Let us denote by Y the response variable `sr`, and by X_1 and X_2 the explanatory variables `pop15` and `dpi`, respectively. The equation of the considered model is therefore:

$$\hat{y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon,$$

where ε denotes the error term. Considering the estimated coefficients, we obtain:

$$\hat{Y} = \mathbb{E}[Y] = 22.713 - 0.330 \cdot X_1 - 0.001 \cdot X_2.$$

This equation represents a plane in the tri-dimensional space.

4.8.0.9 19.9 Compute the fitted value for Italy and calculate its residual value.

The output of the model contains the fitted value and the residual value for each of the original sample units. We print both values along with the observed saving rate for Italy.

```
dat["Italy", 1]
```

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

```
round(model19$fitted["Italy"], 2)
```

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

```
round(model19$residuals["Italy"], 2)
```

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

- The estimated and observed values for the Italian saving rate are equal to $y = 14.28$ 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 computed for Italy is greater than 75% of the residuals of the other countries.

4.8.0.10 19.10 Report and comment on the estimated of ^2.

The estimated value of σ^2, in the following denoted as s^2, may be computed as the sum of the squared residuals divided by the number of degrees of freedom $(n - p)$.

```
n19 <- dim(dat)[1]
p19 <- dim(dat)[2]

s2_19 <- sum(model19$residuals^2)/(n19-p19); s2_19
```

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

```
s_19 <- sqrt(s2_19); s_19
```

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

The value of s, which is also reported in the summary of the model (see point 19.7), 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 s 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.

```
TSS19 <- (n19-1) * var(dat[, 1]); TSS19
```

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

```
SSR19 <- (n19-p19) * s2_19; SSR19
```

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

```
SSE19 <- TSS19 - SSR19; SSE19
```

```
## [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_19 <- SSE19/TSS19; R2_19
```

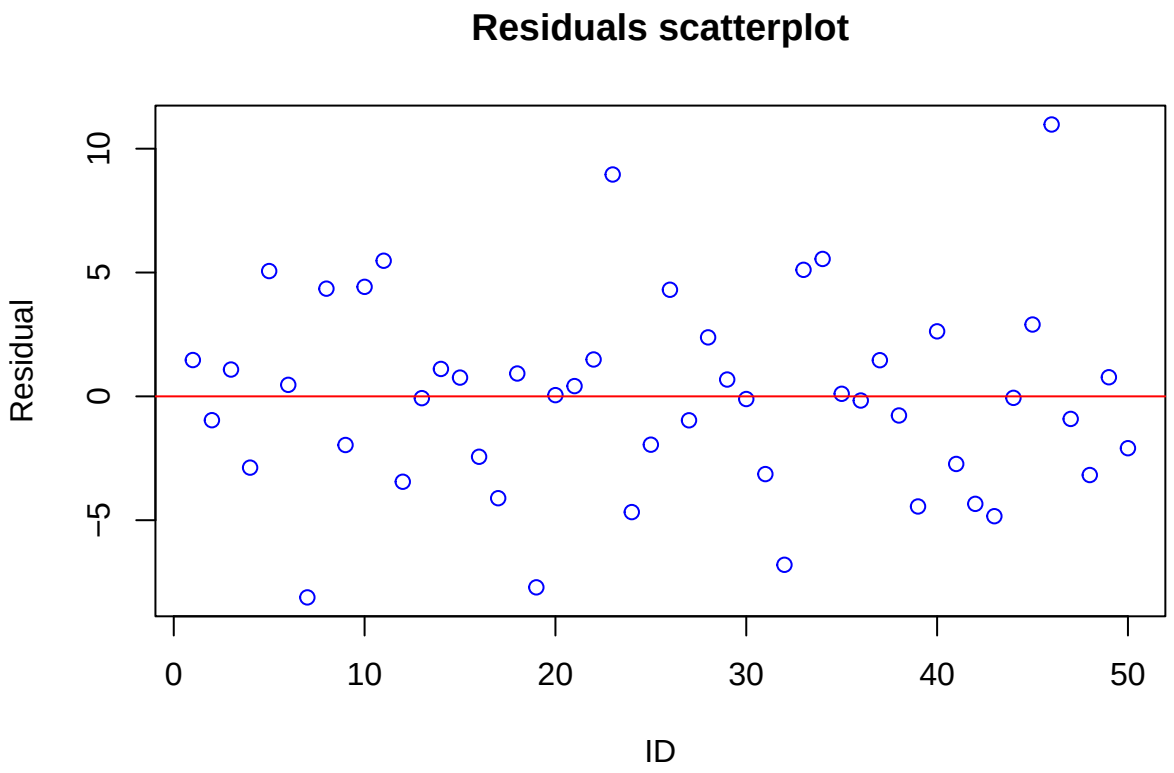
```
## [1] 0.2434909
```

4.8.0.11 19.11 Perform graphical inspection of the residuals and comment on each plot. Are there some violations of the assumptions made by the classical multiple linear regression model?

Before performing graphical analysis of the residual, we take a look at the descriptive statistics included in the summary of the model. The median is very close to 0, in accordance with the theoretical assumptions. The range of the distribution seems slightly larger than expected, going from a minimum of -8.12 to a maximum of 10.98.

1. The first plot displays on the cartesian plane the residuals automatically computed by the `lm()` function; we add the horizontal line at 0 to check if the residuals are approximately arranged in proximity and in a symmetric way with respect to this line.

```
plot(model19$residuals,
     main = "Residuals scatterplot",
     xlab = "ID", ylab = "Residual",
     col = "blue")
abline(h = 0, col = "red")
```



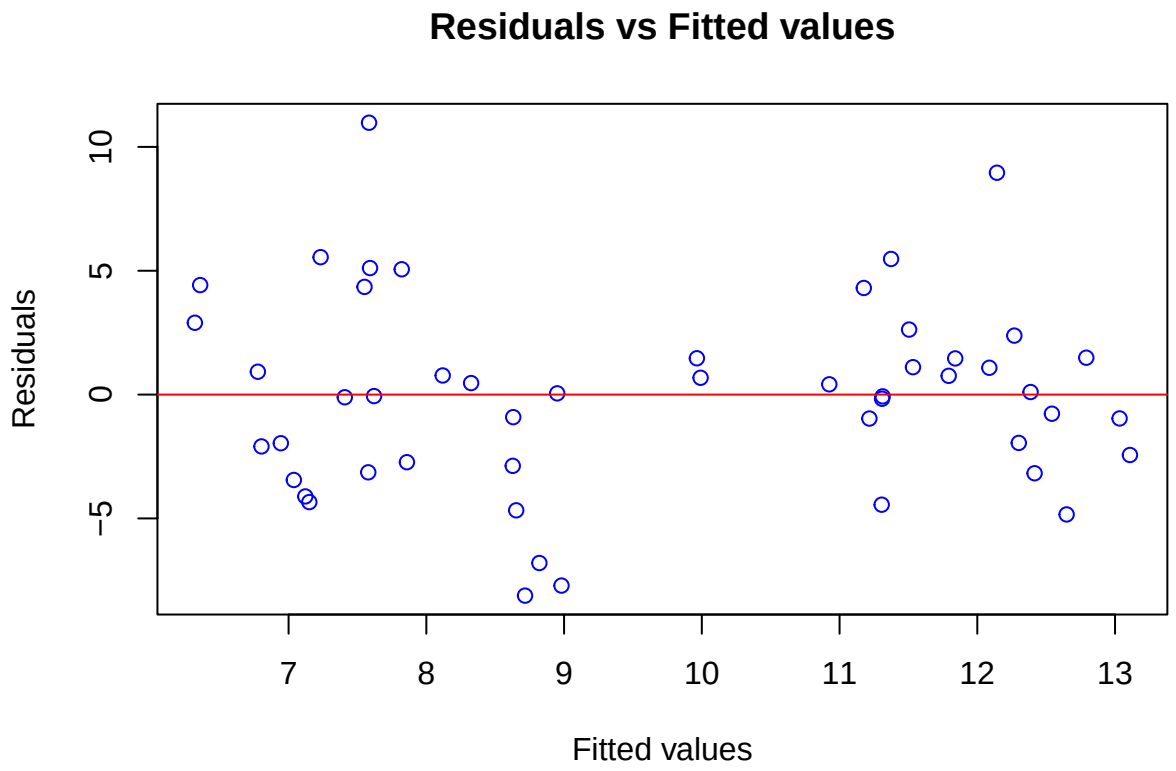
- Firstly, as already observed by the descriptive statistics returned by the `lm()` function, we notice that, although slightly broader than the expected, the range of variation of the residuals is quite limited, ranging approximately from a minimum of -10 to a maximum of 10. There are, therefore, no 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).

2. The second graph represents on the x-axis the values interpolated by the model, and on the y-axis, again, the residuals. Also in this case, we add the horizontal line at the value of 0.

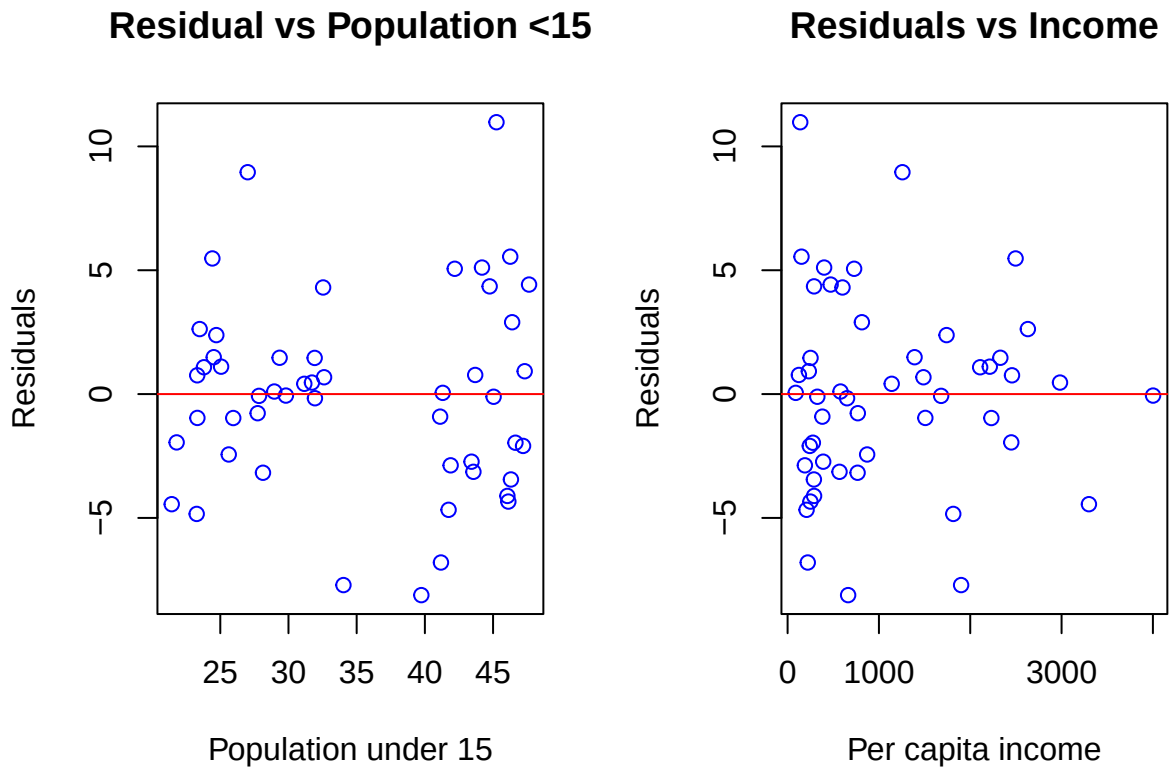
```
plot(model19$fitted,
     model19$residuals,
     main = "Residuals vs Fitted values",
     xlab = "Fitted values", ylab = "Residuals",
     col = "blue")
abline(h = 0, col = "red")
```



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.

3. The third pair of graphs display the explanatory variables on the x-axis and the residuals on the y-axis.

```r
par(mfrow = c(1,2))

plot(dat[, 2],
     model19$residuals,
     main = "Residual vs Population <15",
     xlab = "Population under 15", ylab = "Residuals",
     col = "blue")
abline(h = 0, col = "red")

plot(dat[, 3],
     model19$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.

4. The last pair of plots is used to assess normality of the residuals; therefore we represent the empirical distribution function (ECDF) and the QQ-plot, comparing them with the theoretical results of a standard Gaussian distribution. Note that we first have to standardize and/or studentize the residuals.

```
round(summary(model19$residuals), 3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.117  -2.656  -0.005   0.000   1.483  10.976
```

```
stand_res19 <- rstandard(model19)
round(summary(stand_res19), 3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.066  -0.685  -0.003   0.000   0.386   2.824
```

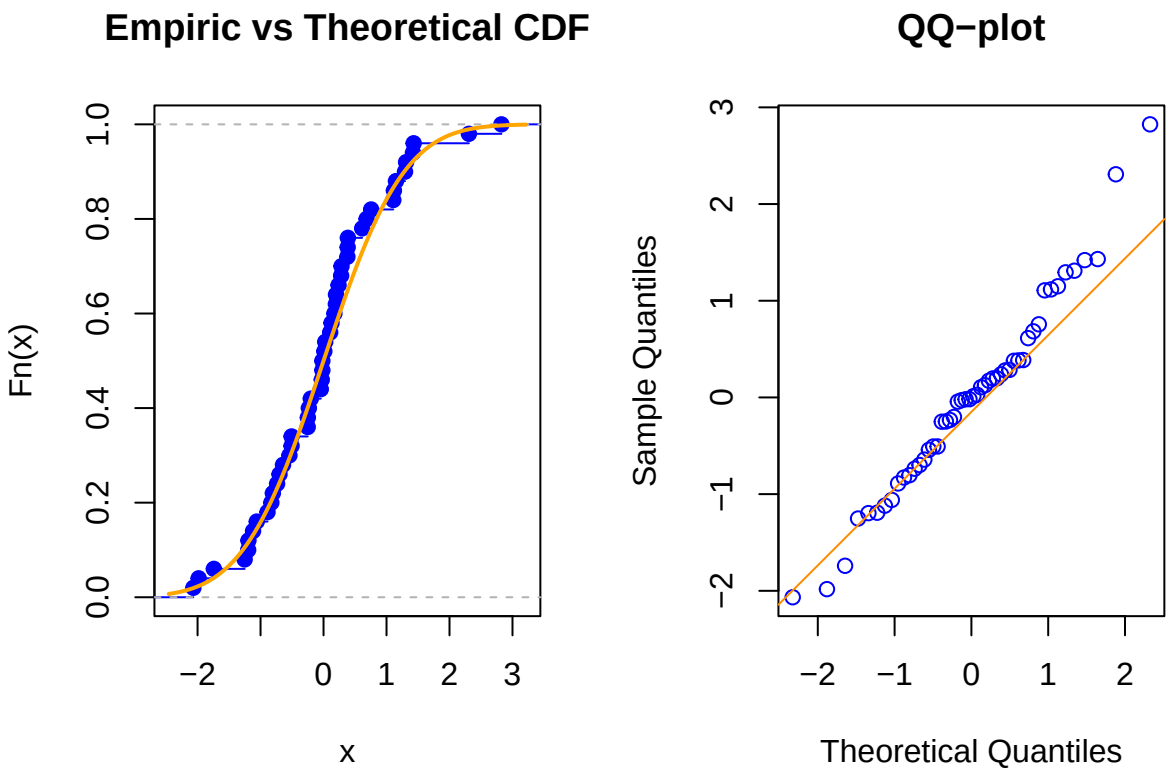
```
stud_res19 <- rstudent(model19)
round(summary(stud_res19), 3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.144  -0.681  -0.003   0.004   0.382   3.066
```

We notice that the values we obtain for standardized and studentized residuals are quite similar to each other; we only observe that, as expected, the distribution of studentized residuals has slightly heavier tails, since the student's t-distribution has heavier tails than a standard normal distribution. the original residuals on the contrary show a wider range of variation. This is a clear indication of the fact that the variance of the residual distribution is different from 1 (we still have to assess that the distribution is actually Gaussian).

We represent now the ECDF and QQ-plot of the (e.g.) standardized residuals in the same graphical window.

```
par(mfrow = c(1, 2))
plot(ecdf(stand_res19),
     main ='Empiric vs Theoretical CDF',
     col = "blue")
curve(pnorm(x),
      col = 'orange',
      lwd = 2,
      add = TRUE)
qqnorm(stand_res19,
       main = "QQ-plot",
       col = "blue")
qqline(stand_res19, 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.

4.8.0.12 19.12 Interpret each estimated regression coefficients and also describe the results to verify the null hypothesis that the coefficient is equal to zero.

Each coefficient reflects the marginal expected change in the saving rate (`sr`) for a one-unit increase in the respective covariate, holding the other variable constant.

- The **intercept** of 22.71 represents the predicted saving rate for a hypothetical country with both `pop15 = 0` and `dpi = 0`. While this situation is not realistic, the intercept serves as a baseline reference for the linear combination of the covariates.
- The **coefficient of `pop15`** is -0.33 and is **statistically significant** at the 1% level ($p = 0.00109$). This means that, holding disposable income constant, an increase of one percentage point in the population under 15 is associated with a decrease of 0.33 in the aggregate personal saving rate. The negative sign is economically plausible: countries with younger populations may have lower savings due to higher dependency ratios and consumption needs.

- The **coefficient of `dpi`** is -0.00131 and is **not statistically significant** ($p = 0.14$). This suggests that, after accounting for population structure, the effect of disposable income on savings is weak or possibly nonexistent in this linear formulation. Though the sign is negative, indicating a slight reduction in savings with higher income, we cannot reject the null hypothesis that $\beta_{\text{dpi}} = 0$ at conventional significance levels.

To formally verify the null hypotheses, we refer to the **t-tests** provided in the output:

- For `pop15`, the test statistic is $t = -3.48$ with $p = 0.0011$, leading us to **reject the null hypothesis** and conclude that `pop15` has a statistically significant effect on savings.
- For `dpi`, the test statistic is $t = -1.50$ with $p = 0.1416$, so we **fail to reject the null hypothesis**, and the evidence for an effect is not strong.

In summary, the model indicates that **demographic structure is a key determinant** of saving behavior across countries, while the role of income appears less clear in this specification.

4.9 Open questions

- **What particular points (leverage, anomalous and influential) can be determined by the graphical inspection of the residuals and by which graph?**

In multiple linear regression, the graphical inspection of residuals allows us to detect three types of potentially problematic observations:

1. **High-leverage points** These are observations that show extreme values in the explanatory variables. They are far from the center of the predictor space and can influence the position of the fitted regression line, even if their residuals are small. They are usually detected through the *Residuals vs Leverage* plot, where they appear far along the x-axis. Although not necessarily outliers in the response variable, they can exert substantial influence on the estimation process.
2. **Anomalous points (outliers)** These points exhibit large residuals, meaning that the actual response value differs substantially from the predicted value. They can be identified in the *Residuals vs Fitted* plot, where they lie far from the horizontal line at zero. Their presence may suggest problems in model specification, measurement errors, or simply variability not captured by the current set of predictors.
3. **Influential points** These are observations that both have high leverage and large residuals. Their presence can substantially alter the estimated regression coefficients. They are identified in the *Residuals vs Leverage* plot, particularly when they fall outside the Cook's distance contours. These points deserve special attention as they may distort inference if not properly considered.

The full diagnostic analysis should include the following standard plots:

- *Residuals vs Fitted*: to detect non-linearity and outliers.
- *Normal Q-Q*: to assess normality of residuals.
- *Scale-Location*: to evaluate the assumption of constant variance (homoscedasticity).
- *Residuals vs Leverage*: to detect influential and high-leverage points.

These visual tools are essential to assess whether the assumptions of the classical linear regression model are satisfied and whether the estimated model is robust to the presence of anomalous data.

- **Which are the measures we employ to detect these particular points numerically?**

In multiple linear regression, we use several numerical diagnostics to identify particular observations such as **high-leverage**, **outlying**, and **influential** points. The most common measures include:

- **Leverage (hat values)**: Computed from the diagonal elements of the hat matrix $H = X(X^TX)^{-1}X^T$, the leverage of observation i is denoted as h_i. It measures how far the values of the explanatory variables for observation i are from the mean of all explanatory variables. A common threshold is:

$$h_i > \frac{2(p+1)}{n}$$

  where p is the number of predictors and n is the sample size.

- **Standardized residuals**: These are residuals divided by their estimated standard deviation. Values exceeding ± 2 (or ± 3) suggest possible outliers.

- **Studentized (or externally studentized) residuals**: These are residuals adjusted for the effect of the observation on the model fit. They are more reliable for detecting outliers, especially when the model has high leverage points.

- **Cook's distance**: Measures the influence of an observation on all regression coefficients. It combines leverage and residual information. A typical rule of thumb flags a point as influential if:

$$D_i > \frac{4}{n}$$

- **DFBETAS**: Indicates the influence of observation i on the estimate of a specific regression coefficient. A value greater than $\frac{2}{\sqrt{n}}$ is often used as a threshold.

- **DFFITS**: Measures the influence of an observation on its own fitted value. Observations with:

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p+1}{n}}$$

  are considered influential.

These metrics are typically calculated using the `influence.measures()`, `hatvalues()`, `rstandard()`, `rstudent()`, and `cooks.distance()` functions in R.

- **Which are the properties of the maximum likelihood estimators?**

The **maximum likelihood estimators (MLEs)** have several desirable theoretical properties under standard regularity conditions. These properties are fundamental in statistical inference and ensure the reliability and efficiency of MLEs in large samples. Specifically:

- **Consistency**: The MLE converges in probability to the true value of the parameter as the sample size $n \to \infty$. That is, for an estimator $\hat{\theta}_{\text{MLE}}$,

$$\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0$$

  where θ_0 is the true parameter value.

- **Asymptotic Normality**: As the sample size increases, the distribution of the MLE approaches a normal distribution:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0))$$

  where $I(\theta_0)$ is the Fisher information.

- **Efficiency**: Among consistent estimators, the MLE achieves the **lowest possible asymptotic variance**, equal to the inverse of the Fisher information matrix. Therefore, the MLE is asymptotically efficient.
- **Invariance**: If $\widehat{\theta}_{\text{MLE}}$ is the MLE of θ, then for any function $g(\cdot)$, the MLE of $g(\theta)$ is $g(\widehat{\theta}_{\text{MLE}})$.
- **Sufficiency and Likelihood Principle**: MLEs are based on sufficient statistics when they exist, and their estimation depends only on the likelihood, not on any ancillary information.

These properties justify the widespread use of MLEs in both classical and modern statistical modeling.

- **Which is the distributional assumption for the error terms? And what does it implies for the conditional model of the response given the covariates?**

In the classical linear regression framework, the **error terms** ε_i are assumed to be **independent and identically distributed (i.i.d.)** with a **normal distribution of mean zero and constant variance**:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

This assumption implies that the response variable Y_i, **conditionally on the covariates** X_i, is also normally distributed:

$$Y_i \mid X_i \sim \mathcal{N}(X_i^\top \beta, \sigma^2)$$

In other words, the **conditional distribution of the response** is Gaussian with:

- **Mean**: $\mathbb{E}[Y_i \mid X_i] = X_i^\top \beta$
- **Variance**: $\text{Var}(Y_i \mid X_i) = \sigma^2$, which does **not depend on** X_i (i.e., **homoscedasticity**)

This implies that the regression model is fully characterized by a **linear conditional expectation** and a **constant conditional variance**, and that the randomness around the regression line is normally distributed. This assumption is crucial for the validity of **inference based on t and F tests**, as well as for constructing **confidence intervals and prediction intervals**.

- **What does it mean the assumption of homoschedastity? How the variance of the error terms is estimated?**

The assumption of **homoscedasticity** means that the **variance of the error terms is constant** across all observations, i.e.:

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{for all } i$$

This implies that the **spread of the residuals** does not depend on the value of the covariates: every observation has the same level of uncertainty (or noise) around the regression line.

The variance of the error terms, σ^2, is estimated using the **residual sum of squares (RSS)** divided by the degrees of freedom:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{RSS}{n-p-1}$$

where:

- y_i are the observed values,
- $\widehat{y}_i$ are the fitted values from the model,
- n is the number of observations,
- p is the number of covariates (excluding the intercept).

This estimator is **unbiased** under the assumption of homoscedasticity and is reported in the regression summary as the **residual variance** $\hat{\sigma}^2$, with its square root being the **residual standard error**.

- **Which is the distribution of the response variables when we assume a multivariate Gaussian distribution for the error terms?**

When we assume that the **error terms** follow a multivariate Gaussian distribution, i.e.,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

and the linear model is given by:

$$Y = X\beta + \varepsilon$$

then the **response variable** Y itself is also normally distributed:

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

This implies that the **conditional distribution** of the response vector Y, given the matrix of covariates X, is multivariate normal with:

- **mean** $\mathbb{E}(Y|X) = X\beta$
- **covariance matrix** $\text{Var}(Y|X) = \sigma^2 I_n$

In other words, the outcomes are jointly normally distributed around the linear predictor $X\beta$, with constant variance and no correlation across observations. This distributional assumption is crucial for deriving the **sampling distributions** of the estimators and for performing valid **inference** (e.g., hypothesis testing, confidence intervals).

- **Write the extended formula of the multiple linear regression model when there are five covariates and describe each component of the equation.**

The extended formula of a multiple linear regression model with five covariates is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i$$

where:

- Y_i is the **response variable** for the i-th observation.
- β_0 is the **intercept**, representing the expected value of Y when all covariates are 0.
- β_k for $k = 1, \ldots, 5$ are the **regression coefficients** for each covariate X_k; they measure the expected change in Y for a one-unit increase in X_k, holding all other variables constant.
- X_{ik} is the **value of the k-th covariate** for the i-th observation.
- ε_i is the **random error term**, assumed to follow a normal distribution with mean 0 and variance σ^2, accounting for unobserved variability.

This model linearly combines the contributions of each covariate to explain the variability in the outcome Y.

- **Write the multiple linear regression model in matrix notation and specify the dimension of each quantity.**

The multiple linear regression model in **matrix notation** is written as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where:

- $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is the **vector of response variables** for the n observations;
- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the **design matrix**, with the first column equal to 1 (for the intercept), and the remaining p columns containing the values of the covariates;
- $\beta \in \mathbb{R}^{(p+1)\times 1}$ is the **vector of regression coefficients**, including the intercept β_0;
- $\varepsilon \in \mathbb{R}^{n\times 1}$ is the **vector of error terms**, assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$.

Thus, the dimensions are:

- $\mathbf{Y}$: $n \times 1$
- $\mathbf{X}$: $n \times (p+1)$
- β: $(p+1) \times 1$
- ε: $n \times 1$
- **Specify the assumptions required for the multiple linear regression model.**

The multiple linear regression model relies on the following assumptions:

1. **Linearity**: The conditional expectation of the response variable is a linear combination of the covariates. That is, $\mathbb{E}(Y_i \mid \mathbf{X}_i) = \mathbf{X}_i^\top \beta$.
2. **Exogeneity**: The error terms have zero mean and are uncorrelated with the covariates: $\mathbb{E}(\varepsilon_i \mid \mathbf{X}_i) = 0$.
3. **Homoscedasticity**: The error terms have constant variance: $\mathrm{Var}(\varepsilon_i \mid \mathbf{X}_i) = \sigma^2$, for all i.
4. **Independence**: The observations are independent of each other.
5. **Normality** (for inference): The error terms ε_i are normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

These assumptions ensure that the least squares estimators are unbiased, efficient, and normally distributed in finite samples (under normality), and allow for valid hypothesis testing and confidence intervals.

- **Describe the estimation method employed in the classical linear regression.**

In classical linear regression, the **estimation method** used is **Ordinary Least Squares (OLS)**. This technique estimates the vector of regression coefficients β by minimizing the **residual sum of squares (RSS)**, which measures the total squared difference between observed and predicted values of the response variable:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2$$

In matrix form, this becomes:

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Here, $\mathbf{X}$ is the $n \times p$ design matrix (including a column of ones if an intercept is included), and $\mathbf{y}$ is the $n \times 1$ vector of responses.

Under the Gauss-Markov assumptions, this estimator is **BLUE** (Best Linear Unbiased Estimator), meaning it has the lowest variance among all linear and unbiased estimators.

- **In which way can we decompose the variability of the response?**

In multiple linear regression, the total variability of the response variable y can be decomposed as follows:

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{Total Sum of Squares (TSS)}} = \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}_{\text{Explained Sum of Squares (ESS)}} + \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\text{Residual Sum of Squares (RSS)}}$$

This decomposition expresses that the **total variability** in the observed data is partitioned into:

- **ESS**: the part explained by the regression model (variability of fitted values),
- **RSS**: the part left unexplained (variability of residuals).

This identity is at the core of the definition of the **coefficient of determination** R^2, which quantifies the proportion of variance in the response explained by the model:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- **Which is the least squares estimator of the intercept and how is it obtained?**

The least squares estimator of the intercept $\widehat{\beta}_0$ in a simple linear regression model is obtained by minimizing the sum of squared residuals between the observed values and the values predicted by the model. The estimator is:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the response and the covariate, respectively, and $\widehat{\beta}_1$ is the least squares estimator of the slope. This formula ensures that the regression line passes through the point $(\bar{x}, \bar{y})$, preserving the average structure of the data. In R, this estimate is returned by the `lm()` function, and can be accessed using the `coef()` function applied to the fitted model.

- **How do we interpret the values of the estimated regression coefficients for each covariate?**

The estimated regression coefficient for a covariate represents the **expected change in the response variable** associated with a **one-unit increase in that covariate**, **holding all other covariates constant**. Specifically, if $\hat{\beta}_j$ is the coefficient for covariate x_j, then:

$$\hat{\beta}_j = \frac{\partial \hat{y}}{\partial x_j}$$

This means that for every one-unit increase in x_j, the predicted response $\hat{y}$ is expected to increase (if $\hat{\beta}_j > 0$) or decrease (if $\hat{\beta}_j < 0$) by $\hat{\beta}_j$ units. In R, these coefficients can be obtained through the `summary()` function applied to the model object fitted with `lm()`, and their interpretation should always consider the context and the scale of the variables involved.

- **What is the multiple R-squared and the adjusted R-squared?**

The **multiple** R^2 is a measure of the proportion of variance in the response variable explained by the model. It is defined as:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where RSS is the residual sum of squares and TSS is the total sum of squares. The closer R^2 is to 1, the better the model explains the data.

However, R^2 **always increases** (or remains the same) when new covariates are added, even if they are not informative. To adjust for this, the **adjusted** R^2 penalizes model complexity and is computed as:

$$\text{Adjusted } R^2 = 1 - \left(\frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} \right)$$

where n is the number of observations and p is the number of predictors. This metric provides a more **reliable comparison between models** with different numbers of covariates. Both statistics are automatically reported by the `summary()` function in R for linear models.

- **Why R-squared is adjusted?**

The **adjusted** R^2 is used because the regular R^2 tends to **increase with each additional covariate**, regardless of whether that covariate is actually useful in explaining the variability of the response variable. This can lead to **overfitting**, where the model appears to fit better simply because it is more complex.

To correct for this, the adjusted R^2 introduces a **penalty for the number of predictors**:

$$\text{Adjusted } R^2 = 1 - \left(\frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} \right)$$

By accounting for the degrees of freedom, it **decreases** if a new variable does not improve the model enough, making it a **more accurate measure of model quality** when comparing models with different numbers of predictors.

- **What is the residual standard error, and how is it interpreted?**

The **residual standard error (RSE)** is an estimate of the **standard deviation of the error terms** in a linear regression model. It quantifies the typical distance between the observed values and the fitted values from the model. Formally, it is computed as:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-p-1}}$$

where:

- RSS is the residual sum of squares,
- n is the number of observations,
- p is the number of predictors (excluding the intercept).

The RSE tells us **how much, on average, the predicted values deviate from the actual values**. A smaller RSE indicates a better fit, but its interpretation also depends on the scale of the response variable. It is expressed in the same units as the response.

- **How are defined the predicted (fitted values)? And the residuals?**

The **fitted values** (also called predicted values) are the values of the response variable estimated by the regression model, obtained by plugging the observed covariates into the estimated regression equation. Formally, if $\hat{\beta}$ is the vector of estimated coefficients and $\mathbf{X}$ is the design matrix, the vector of fitted values $\hat{\mathbf{y}}$ is:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

The **residuals** are the differences between the observed values and the corresponding fitted values. For each observation i, the residual e_i is defined as:

$$e_i = y_i - \hat{y}_i$$

In R, the fitted values and residuals can be obtained using the functions `fitted()` and `residuals()` applied to a fitted model object.

- **Specify the properties of the residuals and explain how we expect the scatterplot of the residuals by unit number where the assumptions of the model are satisfied.**

When the assumptions of the classical linear regression model are satisfied, the **residuals** possess the following properties:

- They have **zero mean**, i.e., $\sum e_i = 0$, which implies that the residuals are centered around zero.
- They are **uncorrelated with the fitted values** and with each explanatory variable used in the model.
- They have **constant variance** (homoscedasticity): $\text{Var}(e_i) = \sigma^2 \ \forall i$, assuming that the error terms have equal variance.
- Under the assumption of normally distributed error terms, the residuals are approximately **normally distributed** as well.

If all assumptions hold, the **residual plot against the unit number** (i.e., the order of observations) should display a **random scatter around zero** without any evident structure, pattern, or trend. No systematic curvature, clustering, or heteroscedastic fan-shapes should be present. This randomness supports the idea that the model fits the data well and that the error terms behave as expected.

- **In which way we calculate confidence intervals for the regression coefficients?**

Confidence intervals for the regression coefficients in a multiple linear regression model are calculated using the standard formula based on the **t-distribution**, since the sample estimate of the standard error is used. Given a coefficient estimate $\hat{\beta}_k$, the confidence interval at level $1 - \alpha$ is:

$$\hat{\beta}_k \pm t_{n-p-1,\alpha/2} \cdot \text{SE}(\hat{\beta}_k)$$

where:

- $\hat{\beta}_k$ is the estimated coefficient,
- $\text{SE}(\hat{\beta}_k)$ is the standard error of the estimate,
- $t_{n-p-1,\alpha/2}$ is the quantile of the **Student's t-distribution** with $n - p - 1$ degrees of freedom (where n is the number of observations and p the number of predictors).

In R, the `confint()` function computes these intervals automatically for all model parameters, using the default confidence level (usually 95%) or a custom one if specified.

- **In a model where we try to explain diastolic blood pressure according to age of the patients and body mass index, and we estimate a confidence interval at 95% at a confidence level of 95% of (0.387; 0.634), how do we interpret these values?**

The confidence interval $(0.387; 0.634)$ refers to the estimated regression coefficient of one of the covariates in the model (for example, body mass index or age). This interval means that, with 95% confidence, the true value of the coefficient lies between 0.387 and 0.634.

If the coefficient corresponds, for example, to **body mass index (BMI)**, the interpretation would be:

> *For each one-unit increase in BMI, the diastolic blood pressure is expected to increase by between 0.387 and 0.634 mmHg, assuming age is held constant.*

Since the entire interval is **strictly positive**, we can also say that the effect of BMI on diastolic blood pressure is **statistically significant at the 5% level**.

- **Why we perform the F-test?**

We perform the **F-test** in multiple linear regression to evaluate whether the **model as a whole** provides a statistically significant explanation of the variability in the response variable.

Specifically, the F-test assesses the **null hypothesis**:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

against the alternative that **at least one** β_j is different from zero, where p is the number of covariates (excluding the intercept).

The test compares the **explained variance** (due to the model) to the **unexplained variance** (residuals). If the model explains a significant portion of the variability, the F-statistic will be large, and the **p-value** will be small (typically < 0.05), leading us to reject the null hypothesis and conclude that the model has **predictive power**.

- **How it is performed the F-test?**

The **F-test** in multiple linear regression is performed by comparing two sources of variability: the variability explained by the model and the residual (unexplained) variability. It uses the following formula:

$$F = \frac{\text{MS}_{\text{reg}}}{\text{MS}_{\text{res}}} = \frac{(\text{SSR}/p)}{(\text{SSE}/(n-p-1))}$$

where:

- SSR is the **sum of squares due to regression**,
- SSE is the **sum of squared errors (residuals)**,
- p is the number of covariates (excluding the intercept),
- n is the number of observations.

The numerator represents the **average explained variability**, and the denominator the **average unexplained variability**. The F-statistic follows an **F-distribution** with p and $n-p-1$ degrees of freedom under the null hypothesis.

In R, the F-test is automatically provided in the output of the `summary()` function on a linear model. The associated **p-value** indicates whether the full model significantly improves prediction compared to a null model (with only an intercept).

- **In a model where we try to explain diastolic blood pressure according to age of the patients and body mass index the result of the F-test is the following F-statistic 82.59 on 2 and 726 degrees of freedom p-value $<$ 0.00002. Which are the observed quantities and how we interpret these results?**

The output of the F-test in this case is:

$$\text{F-statistic} = 82.59 \quad \text{on} \quad 2 \text{ and } 726 \text{ degrees of freedom}, \quad \text{p-value} < 0.00002$$

This result tells us the following:

- The model includes **2 explanatory variables** (age and body mass index), hence the numerator degrees of freedom is 2.
- The residual degrees of freedom is **726**, corresponding to the total number of observations minus the number of estimated parameters (including the intercept), i.e. $n-p-1 = 729-2-1 = 726$.
- The **F-statistic of 82.59** quantifies how much better the full model (with age and BMI) explains the variability in diastolic blood pressure compared to a null model (with only an intercept).
- The associated **p-value $<$ 0.00002** is extremely small, meaning we reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ and conclude that at least one of the covariates significantly contributes to explaining the response variable.

In summary, both **age and BMI together provide a statistically significant improvement** in predicting diastolic blood pressure.

- **What is multicollinearity? Can you provide an example with an illustration?**

Multicollinearity is a situation in multiple linear regression when **two or more explanatory variables are highly linearly correlated**. This means that one predictor can be approximated by a linear combination of the others, which causes problems in estimating the unique contribution of each covariate.

When multicollinearity is present:

- The **standard errors of the estimated coefficients** become inflated.
- Coefficients may become **statistically insignificant**, even if they are theoretically relevant.
- The model becomes **unstable**, and small changes in the data can lead to large changes in coefficient estimates.
- The **interpretation of individual coefficients** becomes unreliable.

Example: Suppose we are modeling house prices using both the total square footage and the number of rooms.

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{TotalSquareFeet} + \beta_2 \cdot \text{NumberOfRooms} + \varepsilon$$

However, in most houses, a larger number of rooms naturally implies more square footage. Thus, these two variables are highly correlated. As a result, the model cannot easily determine how much of the effect on price is due to square footage versus number of rooms.

Graphical Illustration (describe in R or markdown):

You can visualize this in R using a **scatter plot matrix** (`pairs()` or `GGally::ggpairs()`), or compute the **correlation matrix** with `cor()` to observe near-perfect correlation (e.g., > 0.9) between predictors.

To **detect multicollinearity numerically**, we use the **Variance Inflation Factor (VIF)**. A VIF > 5 (or 10) typically indicates problematic multicollinearity.

- **Which measure is employed to assess the presence of excessive collinearity? And if it is too high which actions can be performed?**

The most commonly employed measure to assess **excessive collinearity** among predictors in a multiple linear regression model is the **Variance Inflation Factor (VIF)**. For a given covariate X_j, the VIF is defined as:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination obtained by regressing X_j on all the other predictors. A high VIF indicates that X_j is highly linearly correlated with the other predictors.

Interpretation of VIF values:

- VIF < 5: low to moderate collinearity, acceptable.
- VIF > 5 or especially > 10: severe multicollinearity that may distort regression estimates.

To compute VIFs in R, you can use the `vif()` function from the `car` package.

If VIF is too high, common remedies include:

- **Removing** one or more highly correlated covariates.
- **Combining** correlated variables into a composite index or principal component.
- **Centering** the variables (subtracting the mean) to reduce non-essential collinearity.
- Applying **regularization techniques** such as **ridge regression**, which penalize large coefficients and can handle multicollinearity better than ordinary least squares.
- **Variable selection is performed with information criteria. Which are the most popular and which is the principle behind their usage?**

The most popular information criteria used for variable selection in statistical modeling are the **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)**. Both criteria aim to balance **model fit** and **model complexity**, penalizing the addition of unnecessary predictors.

The **general principle** behind their usage is to avoid overfitting by selecting the model that best explains the data with the fewest parameters.

Formally, both are defined as:

- **AIC**:

$$\text{AIC} = -2 \cdot \log \mathcal{L} + 2p$$

- **BIC**:

$$\text{BIC} = -2 \cdot \log \mathcal{L} + p \cdot \log(n)$$

Where:

- $\log \mathcal{L}$ is the log-likelihood of the model,
- p is the number of estimated parameters (including the intercept),
- n is the number of observations.

AIC favors models with better predictive accuracy and is more permissive in including variables, while **BIC** imposes a stronger penalty for complexity and is more consistent in identifying the true model as the sample size grows.

In R, both criteria can be computed using the `AIC()` and `BIC()` functions. Automated model selection can be performed with `step()`, which uses AIC by default.

- **Automatic procedures are used jointly with the information criteria. Which are they and how they differ?**

The most commonly used automatic procedures for variable selection in regression models are **forward selection**, **backward elimination**, and **stepwise selection**. They differ in the direction in which covariates are added or removed and in how they explore the model space:

- **Forward selection** starts from the intercept-only model (null model) and adds variables one at a time. At each step, it selects the variable whose inclusion most improves the model according to a criterion like AIC or BIC. The process stops when no additional variable improves the model sufficiently.
- **Backward elimination** starts from the full model (including all candidate predictors) and removes one variable at a time. At each step, it eliminates the variable whose removal least worsens the model according to the chosen criterion. It stops when no further variable can be removed without significantly worsening the model.
- **Stepwise selection** combines both forward and backward strategies. Starting from an initial model, it alternates between adding and removing variables at each step based on whether they improve the information criterion. This allows for reevaluation of variables already included or excluded.

These procedures are implemented in R through the `step()` function. By default, `step()` performs **stepwise selection** using **AIC** but it can be customized to use **BIC** by setting the penalty parameter `k = log(n)` in the function call.

4.10 Out-of-sample prediction and uncertainty estimation: predicted rating score with 99% interval and bootstrap confidence intervals for regression coefficients.

4.10.0.1 20.1 Concerning data about ratings ('ratings.Rdata'). Considering the following values of profit (3.9), capital flexibility (6.02), and financial flexibility (1.43), calculate the predicted value of the score for a firm out of sample. Comment on the point estimate of the predicted score value for a new unit and on its prediction interval at a confidence level of 0.99.

The estimated model (computed in Exercise 13) is:

$$\hat{y}_{\text{rating}} = -28.88 + 0.33 \cdot \text{profit} + 3.91 \cdot \text{capital} + 19.67 \cdot \text{flex}$$

So, for the provided values of the covariates, the predicted score for a firm out of samples will be:

```
-28.88+(0.33*3.9)+(3.91*6.02)+(19.67*1.43)
```

```
## [1] 24.0733
```

So, the firm is expected to receive a rating about 24.07, given its financial structure. This value reflects the model's best linear prediction based on the covariates and indicates a moderate score, assuming the model holds for new firms.

Now we compute the prediction interval for a new observation using the `predict()` function. This means that, based on the regression model and the input covariates, the predicted rating for this new firm is approximately 24.08. However, this point estimate is only the center of a wider prediction interval, which quantifies the uncertainty in predicting a new, unobserved response.

```
new_obs20 <- data.frame(profit = 3.9, capital = 6.02, f_flex = 1.43)

predict(model13,
        newdata = new_obs20,
        interval = "prediction",
        level = 0.99)
```

```
##        fit       lwr      upr
## 1 24.07958 -10.72555 58.88471
```

The 99% prediction interval is:

$$[-10.73,\ 58.88]$$

This range is extremely wide, which suggests a high level of uncertainty in predicting a new observation. This is typical when:

- The residual variance $\hat{\sigma}^2$ is large,
- The new point lies in a region far from the center of the data (extrapolation),
- Or the model, while explaining a portion of the variability (e.g., with $R^2 \sim 0.65$), still leaves substantial unexplained variability.

With 99% confidence, the true rating of a new firm with those characteristics is expected to lie between -10.73 and 58.88. The large interval reflects the fact that we are predicting an individual (not mean) value, hence accounting for both model error and future variability. Despite the point estimate being moderate (24.08), extreme values are statistically plausible.

4.10.0.2 20.2 Considering these data provide a bootstrap confidence interval at 95% level of confidence. Plot also the bootstrap distribution of each regression coefficient. Compare the results with intervals provided by the 'confint' function. Comment on such a comparison.

In this exercise, we compute the 95% bootstrap confidence intervals for each regression coefficient and compare them to the classical intervals obtained with the `confint()` function. The bootstrap procedure is implemented using 1000 resamplings and the `boot()` function, applied to the linear model coefficients. The resulting bootstrap distributions are then plotted and visually assessed.

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, melanoma
```

```
boot_fun20 <- function(data, indices) {
  d <- data[indices, ]
  coef(lm(rating ~ profit + capital + f_flex, data = d))
}

set.seed(42)
boot_res20 <- boot(data = ratings, statistic = boot_fun20, R = 1000)

boot.ci(boot_res20, type = "perc", index = 1) # Intercept
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res20, type = "perc", index = 1)
##
## Intervals :
## Level     Percentile
## 95%   (-71.26,   6.46 )
## Calculations and Intervals on Original Scale
```

```
boot.ci(boot_res20, type = "perc", index = 2) # profit
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res20, type = "perc", index = 2)
##
## Intervals :
## Level     Percentile
## 95%   (-7.5820, 11.3606 )
## Calculations and Intervals on Original Scale
```

```r
boot.ci(boot_res20, type = "perc", index = 3) # capital
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res20, type = "perc", index = 3)
##
## Intervals :
## Level     Percentile
## 95%   ( 1.927,  5.534 )
## Calculations and Intervals on Original Scale
```

```r
boot.ci(boot_res20, type = "perc", index = 4) # flex
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res20, type = "perc", index = 4)
##
## Intervals :
## Level     Percentile
## 95%   ( 5.50, 34.28 )
## Calculations and Intervals on Original Scale
```

The bootstrap confidence intervals for the regression coefficients of the model predicting firm ratings from `profit`, `capital`, and `flex` have been computed using 1000 bootstrap replicates. The percentile method provides robust, non-parametric intervals that do not rely on normality assumptions.

For the **intercept**, the 95% bootstrap interval is very wide, ranging from -71.26 to 6.46. This suggests a high degree of uncertainty around the baseline level of the rating when all covariates are set to zero. The fact that the interval crosses zero reinforces the idea that the intercept is not of direct interpretative relevance and highly sensitive to the data range.

For the coefficient associated with `profit`, the interval is $[-7.58, 11.36]$. This confirms what classical inference also suggested: `profit` has a highly uncertain impact on the rating and is not statistically significant. The inclusion of zero in this interval means that we cannot reject the null hypothesis that the coefficient of `profit` is zero.

The coefficient for `capital` has a much tighter and entirely positive interval: $[1.93, 5.53]$. This supports the conclusion that capital structure has a statistically significant and positive effect on the firm's rating. It is a stable and robust predictor, as shown by both the bootstrap distribution and the classical t-test.

For `flex`, the interval is also wide but entirely positive: $[5.50, 34.28]$. This confirms the significance and positive association between financial flexibility and rating. Despite the larger uncertainty compared to `capital`, the direction and impact are consistent, and the interval's exclusion of zero validates its explanatory role in the model.

In conclusion, the bootstrap intervals are in line with those obtained through classical inference, but provide additional reassurance about the robustness of the findings. While `profit` remains non-significant, both `capital` and `flex` show reliable and meaningful contributions to the prediction of firm ratings.

We now compare these values with the classical t-based intervals obtained with:

```r
confint(model13)
```

```
##                  2.5 %    97.5 %
## (Intercept) -69.443503 11.689964
## profit       -8.839420  9.494902
## capital       1.345656  6.478026
## f_flex        1.933267 37.407820
```

The classical confidence intervals computed via `confint(model13)` provide the standard 95% t-based estimates under the assumption of approximately normal and homoscedastic residuals. Comparing these intervals to the bootstrap percentile intervals offers insight into the stability of the coefficient estimates and the validity of the model's assumptions.

For the **intercept**, the classical interval is $[-69.44, 11.69]$, which is very close to the bootstrap interval $[-71.26, 6.46]$. Both intervals are wide and include zero, indicating large uncertainty and a lack of significance in the intercept term. This result is expected, since the intercept depends on the reference level when all predictors are zero, which may not be meaningful or within the scope of the data.

The interval for `profit` from the classical method is $[-8.84, 9.49]$, and from the bootstrap method it is $[-7.58, 11.36]$. Both intervals are wide, centered around zero, and overlap almost entirely, confirming that the effect of `profit` is highly uncertain and not statistically significant. The model provides no evidence of a systematic relationship between profitability and rating.

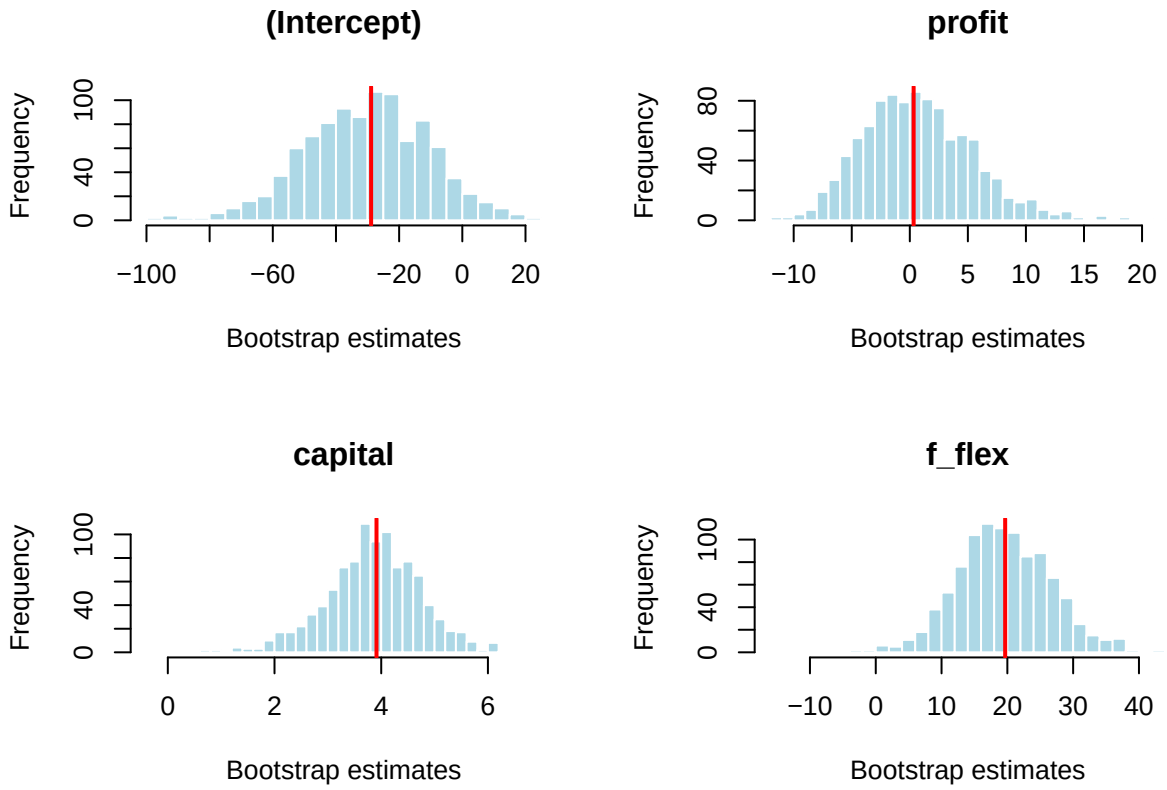
The interval for `capital` is $[1.35, 6.48]$ with the classical method, and $[1.93, 5.53]$ with the bootstrap method. Both intervals are entirely positive and relatively narrow, indicating a robust and statistically significant positive effect of capital structure on firm rating. The close alignment between methods confirms the stability and reliability of this predictor.

For `flex`, the classical interval is $[1.93, 37.41]$, and the bootstrap interval is $[5.50, 34.28]$. Again, both intervals exclude zero and are strongly positive, validating the importance of financial flexibility in explaining ratings. The slight difference in lower bounds is negligible in practical terms and falls within expected sampling variability.

In summary, the comparison confirms that **bootstrap and classical methods lead to consistent conclusions**. `capital` and `flex` are significant and positively associated with firm rating, while `profit` is not. Bootstrap results reinforce classical inference and provide additional robustness in cases where normality or sample size might be questioned.

The graphical inspection of the bootstrap distributions confirms these conclusions:

```r
par(mfrow = c(2, 2))
for (i in 1:4) {
  hist(boot_res20$t[, i],
       breaks = 30,
       col = "lightblue",
       border = "white",
       main = names(coef(model13))[i],
       xlab = "Bootstrap estimates")
  abline(v = coef(model13)[i], col = "red", lwd = 2)
}
```



The histograms show approximately symmetric and unimodal distributions for `capital` and `flex`, validating the assumptions underlying the standard inference. The distribution of `profit` is also centered around zero but more dispersed, reinforcing the finding of weak association with the response.

In conclusion, the bootstrap analysis confirms the reliability of the classical confidence intervals for well-behaved covariates, while offering a more robust and assumption-free check on parameters whose significance is borderline or questionable.

4.11 Prediction of average starting salary for new profile: point estimate and 90% prediction interval with interpretation.

```
newdata21 <- data.frame(exper = 10, educ = 8)

predict(model14,
        newdata = newdata21,
        interval = "prediction",
        level = 0.90)
```

4.11.0.1 21.1 With reference to the exercise where you fitted a multiple linear regression model to explain starting salary as a function of years of education and experience, suppose that you want to predict the average salary value for units out of the sample having 8 years of education and months of previous experience equal to 10. Provide a point estimate and a prediction interval with 90% confidence. Comment on the results.

```
##        fit     lwr      upr
## 1 4664.014 3570.85 5757.179
```

The point prediction for an individual with 8 years of education and 10 months of prior experience is approximately \$4664.01. This value represents the expected starting salary for a new unit with those characteristics, based on the multiple linear regression model estimated in Exercise 14.

The associated 90% prediction interval is $[\$3570.85, \$5757.18]$. This interval accounts not only for the uncertainty in estimating the mean response, but also for the inherent variability in individual outcomes around that mean. Its width reflects the residual variability of the model, which is captured by the residual standard error (636.2).

Given the relatively wide range, the result suggests that while we can reasonably predict the central tendency of the salary based on education and experience, individual-level variation remains substantial. Nonetheless, the lower bound is well above the intercept value, confirming the positive contribution of both covariates to salary prediction. The model thus provides a useful, though not highly precise, estimate for out-of-sample forecasting.

5 Week 6-7 - Interaction models and Model selection via AIC, Linear regression for prediction and inference, Binomial and logistic probability models, Odds, Odds ratios, and Case-control analysis, Logistic regression and ROC curve evaluation

5.1 Multiple regression on energy consumption: research questions, exploratory analysis, model selection via AIC, regression summary, and interpretation of parallel line model.

The data named `MASS::whiteside` (in the MASS library) concerns the consumption of natural gas (in feet) versus temperature in degrees Celsius, considering whether a house was renovated for energy efficiency (thermal insulation or not).

The measurement referred to the consumption of two consecutive winters and was taken before and after the energy efficiency intervention. The 56 weeks were divided into two groups of size 26 (before) and 30 (after).

```
library(MASS)
data("whiteside")
head(whiteside)
```

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

5.1.0.1 22.1 Specify the research questions which can be formulated considering the available data.

Given the structure of the `whiteside` dataset, which includes weekly gas consumption (`Gas`) and external temperature (`Temp`), along with an indicator variable (`Insul`) for whether the data were collected before or after thermal insulation was installed, the following research questions can be formulated:

- Is there a statistically significant relationship between **external temperature** and **gas consumption**? Specifically, does lower temperature correspond to higher gas usage, as expected?
- Does the **installation of insulation** significantly reduce gas consumption, after controlling for temperature?
- Can we detect a difference in the slope or intercept of the regression line (Gas vs Temp) between the *before* and *after* insulation periods, suggesting a structural change in consumption behavior due to the intervention?
- Is the effect of temperature on gas consumption different depending on the insulation status (i.e., is there an interaction between `Temp` and `Insul`)?

These questions aim to assess both the **effectiveness** of the insulation treatment and the **stability** of the relationship between consumption and temperature over time. The analysis will thus help evaluate whether thermal insulation reduces energy demand, especially in colder weeks.

5.1.0.2 22.2 Print levels and labels of the categorical covariate.

To inspect the **levels** and **labels** of the categorical covariate `Insul` in the `whiteside` dataset, we can use the following R command:

```r
levels(whiteside$Insul)
```

```
## [1] "Before" "After"
```

This tells us that the variable `Insul` is a **factor** with two levels:

- `"Before"`: referring to the period *before* the installation of thermal insulation,
- `"After"`: referring to the period *after* the insulation was installed.

If we want to confirm that `Insul` is indeed a factor and check its structure:

```r
str(whiteside$Insul)
```

```
##  Factor w/ 2 levels "Before","After": 1 1 1 1 1 1 1 1 1 1 ...
```

We can also tabulate the counts per group:

```r
table(whiteside$Insul)
```

```
##
## Before  After
##     26     30
```

This means:

- 26 observations in the `"Before"` group,
- 30 observations in the `"After"` group.

```r
skim_without_charts(whiteside)
```

5.1.0.3 22.3 Provide descriptive statistics and a figure to describe the sample data. Are they from an observational or a randomized experiment? Comment on each output.

Table 17: Data summary

Name	whiteside
Number of rows	56
Number of columns	3
Column type frequency:	
factor	1
numeric	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Insul	0	1	FALSE	2	Aft: 30, Bef: 26

Variable type: numeric

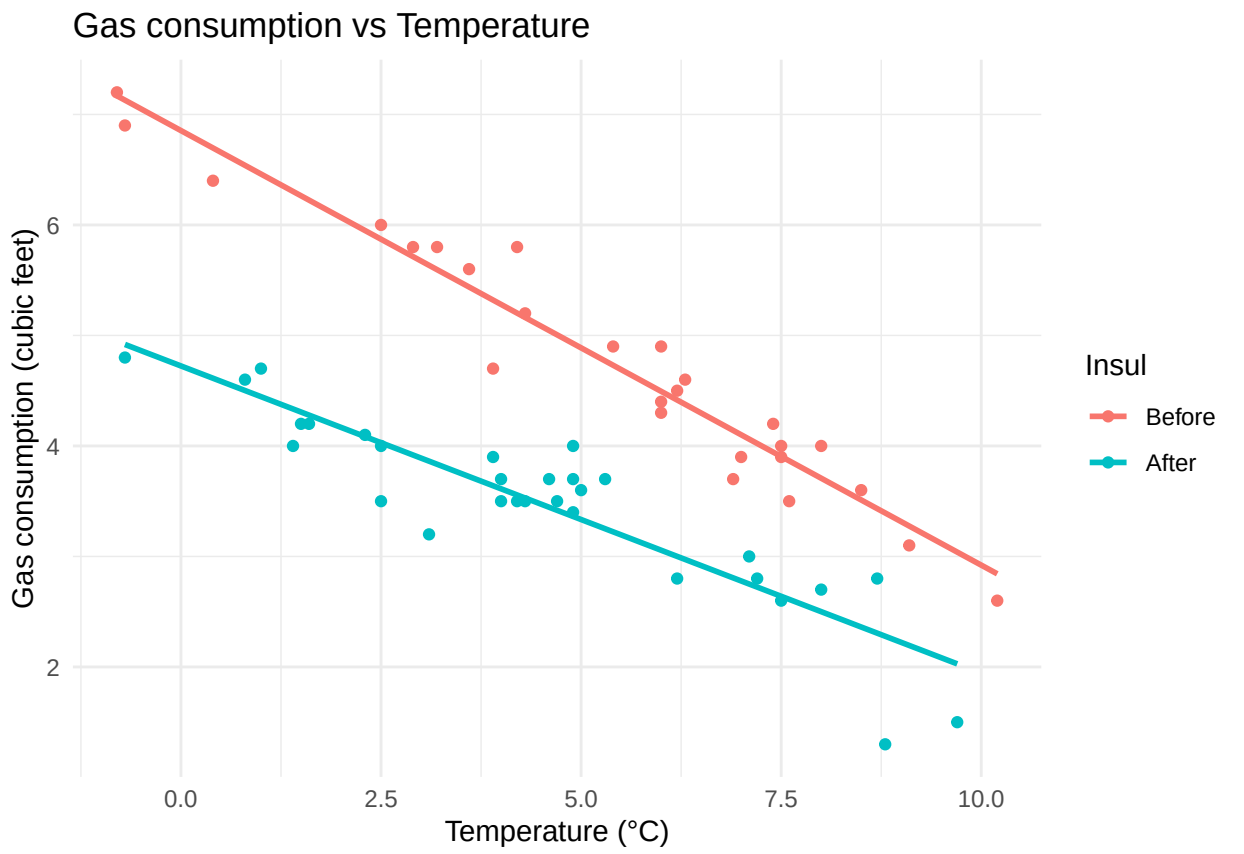
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Temp	0	1	4.88	2.75	-0.8	3.05	4.90	7.12	10.2
Gas	0	1	4.07	1.17	1.3	3.50	3.95	4.62	7.2

The descriptive statistics provide a clear summary of the two quantitative variables. The variable `Temp` ranges from -0.8 to 10.2 with a mean of 4.875 and a standard deviation of 2.75. This suggests moderate variability in temperature across the observed weeks, likely corresponding to winter temperatures. The variable `Gas` ranges from 1.3 to 7.2, with a mean of 4.07 and standard deviation of about 1.17. This implies relatively less dispersion in gas consumption than in temperature. Notably, the median gas consumption (3.95) is close to the lower quartile (3.50), indicating some skewness.

To better understand the relationship between the variables and the effect of insulation, it is informative to visualize the data by plotting gas consumption against temperature, separately for the "Before" and "After" insulation conditions:

```
ggplot(whiteside, aes(x = Temp, y = Gas, color = Insul)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Gas consumption vs Temperature",
       x = "Temperature (°C)",
       y = "Gas consumption (cubic feet)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This plot will reveal that gas consumption decreases with rising temperature, which is expected, but more importantly, we expect the slope to be flatter or the intercept lower in the "After" condition, signaling improved thermal insulation efficiency.

As for the experimental design, the data come from **an observational study with a clear intervention** (thermal insulation), rather than a randomized experiment. The intervention was applied to the same house, and data were collected before and after the intervention across two winters. This makes it a quasi-experiment or a natural experiment, where the treatment (insulation) is not randomly assigned, but its effect is studied using time as a surrogate for the intervention switch. This design allows us to infer causal effects under assumptions, although it lacks the robustness of full randomization.

```
model22 <- lm(Gas ~ Temp*Insul, data = whiteside)

summary(model22)
```

5.1.0.4 22.4 Estimate a multiple linear regression model to explain the consumption as a function of the other covariates. First, use the Akaike information criterion to select the model. Comment on the results provided by the step function.

```
##
## Call:
## lm(formula = Gas ~ Temp * Insul, data = whiteside)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
## Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
## InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
## Temp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

The estimated model includes both main effects and the interaction between temperature (`Temp`) and the insulation status (`Insul`). The results show that all included terms are highly significant. The intercept refers to the baseline case, which is the *"Before insulation"* condition. The coefficient of `Temp` (-0.393) indicates that for non-insulated houses, each additional degree in temperature is associated with a reduction in gas consumption of about 0.39 feet. The negative sign is consistent with expectations, as warmer weeks require less heating. The coefficient `InsulAfter` (-2.13) captures the effect of insulation on gas consumption at zero temperature: for insulated houses, gas consumption is already significantly lower even without considering temperature. The interaction term `Temp:InsulAfter` (+0.115) implies that the marginal effect of temperature on gas consumption is less negative after insulation, meaning that the rate at which gas use decreases with increasing temperature is lower for insulated homes.

The overall model fit is excellent: the R-squared is 0.9277, which indicates that the model explains more than 92% of the variability in gas consumption, and the residual standard error is much lower (0.323) compared to the simple model (0.86). This confirms that introducing the interaction between temperature and insulation significantly improves model performance. The `step()` function would confirm this choice as optimal in terms of the Akaike Information Criterion (AIC), since all included covariates reduce the residual deviance while keeping model complexity justified by their statistical relevance. Therefore, the selected model is statistically solid and interpretable, supporting both the inclusion of insulation status and its interaction with temperature to explain household gas consumption.

```
step(model22)
```

```
## Start:  AIC=-122.72
## Gas ~ Temp * Insul
##
##              Df Sum of Sq    RSS     AIC
## <none>                     5.4252 -122.72
## - Temp:Insul  1    1.3451 6.7704 -112.32

##
## Call:
## lm(formula = Gas ~ Temp * Insul, data = whiteside)
##
## Coefficients:
##     (Intercept)             Temp       InsulAfter  Temp:InsulAfter
##          6.8538          -0.3932          -2.1300           0.1153
```

The stepwise model selection based on the Akaike Information Criterion confirms that the optimal specification is the full interaction model `Gas ~ Temp * Insul`, which includes both the main effects of temperature

and insulation status, and their interaction. Starting from this full model (AIC = –122.72), the procedure evaluates a simplified version that excludes the interaction term. However, this exclusion results in a higher AIC (–112.32), indicating a worse trade-off between model complexity and fit. The improvement in model quality achieved by retaining the interaction term is thus substantial, confirming that the effect of temperature on gas consumption is not constant, but rather depends on whether insulation was installed. The selected model is statistically and substantively appropriate, as it captures how energy efficiency modifies the thermal sensitivity of gas usage.

5.1.0.5 22.5 Report and comment on all the results provided by the summary function on the model selected in the previous step.

The output of the `summary()` function provides several key elements for interpreting the fitted model. The model explains gas consumption (`Gas`) using **temperature (`Temp`)**, **insulation status (`Insul`)**, and their **interaction (`Temp:Insul`)**.

The intercept of the model is estimated at 6.85, which corresponds to the average gas consumption for an uninsulated house (Insul = "Before") when the temperature is 0°C. The coefficient for `Temp` is –0.393, indicating that for uninsulated houses, each additional degree Celsius leads to a reduction of approximately 0.393 cubic feet in gas consumption. This confirms the expected inverse relationship between outdoor temperature and energy usage.

The coefficient for `InsulAfter` is –2.13, meaning that, when temperature is fixed at 0°C, gas consumption in insulated houses is on average 2.13 units lower than in uninsulated ones. This quantifies the energy-saving effect of insulation independently of temperature.

The interaction term `Temp:InsulAfter` has a positive coefficient of 0.115, implying that the insulating intervention reduces the impact of temperature on gas consumption. In other words, insulated houses are less sensitive to cold temperatures: for every additional degree Celsius, the gas usage in insulated houses decreases by only*–0.393 + 0.115 = –0.278, compared to –0.393 for uninsulated houses.

All coefficients are highly significant (p-values < 0.001), meaning the effects are unlikely to be due to chance. The Residual Standard Error is 0.323, suggesting that the model fits the data very well. The **Multiple R-squared is 0.928**, indicating that approximately 93% of the variance in gas consumption is explained by the model. The **Adjusted R-squared is 0.924**, confirming the robustness of the fit even after accounting for the number of predictors. The **F-statistic** is 222.3 on 3 and 52 degrees of freedom, with a p-value $<$ 2.2e-16, which strongly supports the overall statistical significance of the model.

In conclusion, the model effectively captures both the direct effect of temperature and the modifying role of insulation, and it offers a highly accurate description of gas consumption behavior under varying thermal and structural conditions.

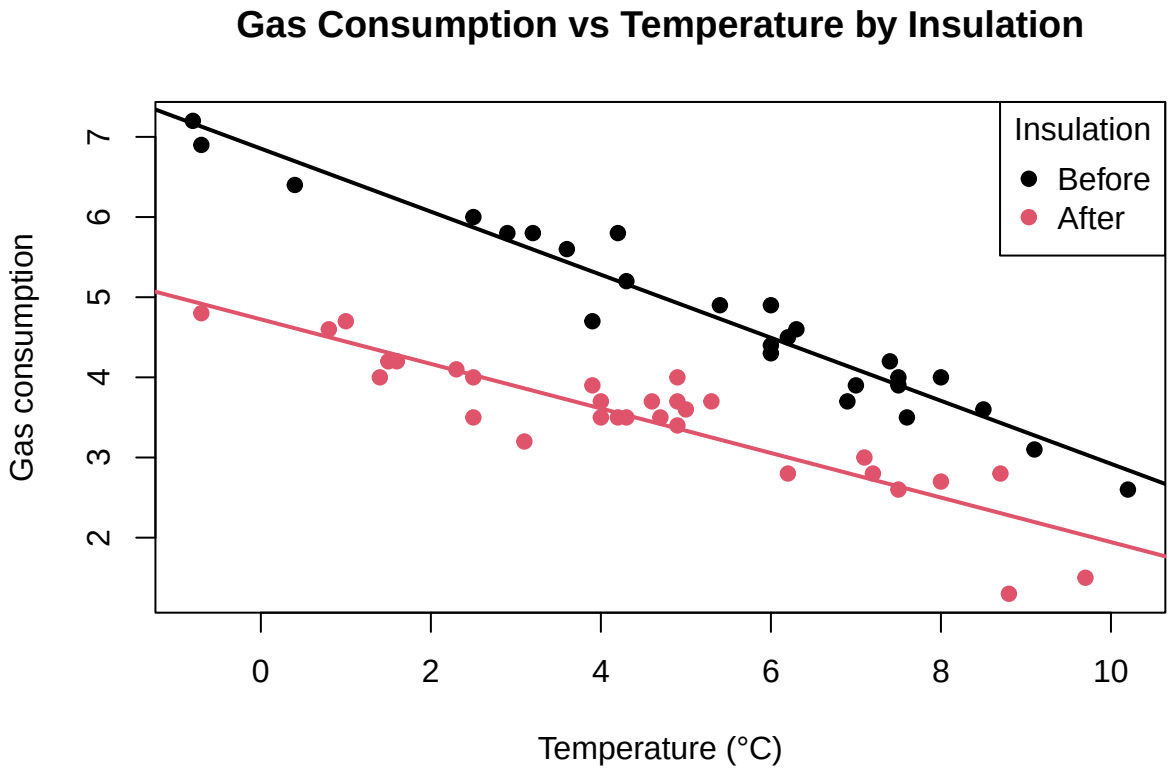
5.1.0.6 22.6 Add the estimated parallel lines to the data scatterplot and comment on the figure. Write the equations characterizing this model.

To visualize the fitted model `Gas ~ Temp * Insul`, we overlay the estimated regression lines on the scatterplot of gas consumption (`Gas`) vs temperature (`Temp`), separately for each insulation condition.

We start by plotting the data, colored by `Insul`, and then add the fitted lines based on the estimated coefficients.

```
plot(whiteside$Temp, whiteside$Gas, col = whiteside$Insul,
     pch = 19, xlab = "Temperature (°C)", ylab = "Gas consumption",
     main = "Gas Consumption vs Temperature by Insulation")
legend("topright", legend = levels(whiteside$Insul),
       col = 1:2, pch = 19, title = "Insulation")

# Coefficients from the model
coef <- coef(model22)
```

```r
# Equation for Insul = "Before"
abline(a = coef[1], b = coef[2], col = 1, lwd = 2)  # (Intercept), Temp

# Equation for Insul = "After"
abline(a = coef[1] + coef[3], b = coef[2] + coef[4], col = 2, lwd = 2)  # Adjusted Intercept and Temp
```



The scatterplot with the two fitted lines clearly illustrates how the insulation intervention modifies the relationship between temperature and gas usage. For the **"Before"** condition (uninsulated house), gas consumption is high at low temperatures and decreases sharply as temperature increases. For the **"After"** condition (insulated house), consumption is consistently lower across the range of temperatures and the slope is less steep, indicating better energy efficiency.

The estimated regression equations from the model are:

- **Before insulation**:

$$\hat{\text{Gas}}_{\text{Before}} = 6.85 - 0.393 \cdot \text{Temp}$$

- **After insulation**:

$$\hat{\text{Gas}}_{\text{After}} = (6.85 - 2.13) + (-0.393 + 0.115) \cdot \text{Temp} = 4.72 - 0.278 \cdot \text{Temp}$$

These equations confirm that at any given temperature, insulated homes use less gas, and their usage decreases more gradually with rising temperature, demonstrating the energy-saving effectiveness of the insulation intervention.

5.2 Regression analysis on diamond carats: model selection via AIC, coefficient interpretation, confidence and prediction intervals, and evaluation of residuals and collinearity.

Data in the file `diamonds.RData` (a subsample of the diamonds dataset in the ggplot2 package) are related to the following measurements:

- `carat`: purity of the diamond (in carats);
- `cut`: quality of the diamond's cut, measured through a categorical variable with categories ranging from 1 (lowest quality) to 5 (highest quality);
- `price`: price of the diamond, measured in 1000$;
- `length`: length of the diamond (in millimeters);
- `width`: width of the diamond (in millimeters);
- `depth`: The depth of the diamond (in millimeters).

Considering as response variable `carat`, do the points of the previous exercise.

- Calculate also a confidence interval for the estimated regression parameters at confidence level of 90%. Comment on the estimated intervals of price and depth.
- Calculate also a prediction interval at 95% confidence for a unit out of sample, assigning values of covariates of your choice. Comment on the reported values.

5.2.0.1 23.1 Data are related to some features and measurements taken on a group of 800 diamonds. A (posisble) research question concerns the estimation of the number of diamonds' carats relying on the available physical features, namely its dimensions (length, width and depth) and the quelity of its cut. Furthermore, it may also be interesting to access the association between the number of carats and the price.

5.2.0.2 23.2 The categorical variable is `cut`, measuring the quality of diamonds' cut. The levels of the variable may be obtained with the `level` function.

```
load("diamonds.Rdata")
levels(diamonds$cut)
```

```
## [1] "1" "2" "3" "4" "5"
```

The variable has 5 different levels, ranging from 1 (corresponding to the lowest quality) to 5 (associated with highest quality).

5.2.0.3 23.3 Data are from an observational study, in which each subject is observed and recorded without any kind of planned experiment. We use the `skimr` package to provide descriptive statistics of the data.

```
skim_without_charts(diamonds)
```

Table 20: Data summary

Name	diamonds
Number of rows	800
Number of columns	6

Column type frequency:	
factor	1
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cut	0	1	FALSE	5	5: 313, 4: 217, 3: 156, 2: 81

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
carat	0	1	0.84	0.48	0.23	0.42	0.72	1.06	2.74
price	0	1	4.18	4.02	0.35	1.10	2.72	5.71	18.78
length	0	1	5.83	1.11	3.94	4.83	5.77	6.57	8.87
width	0	1	5.83	1.10	3.96	4.85	5.78	6.58	8.90
depth	0	1	3.60	0.68	2.39	2.98	3.56	4.04	5.48

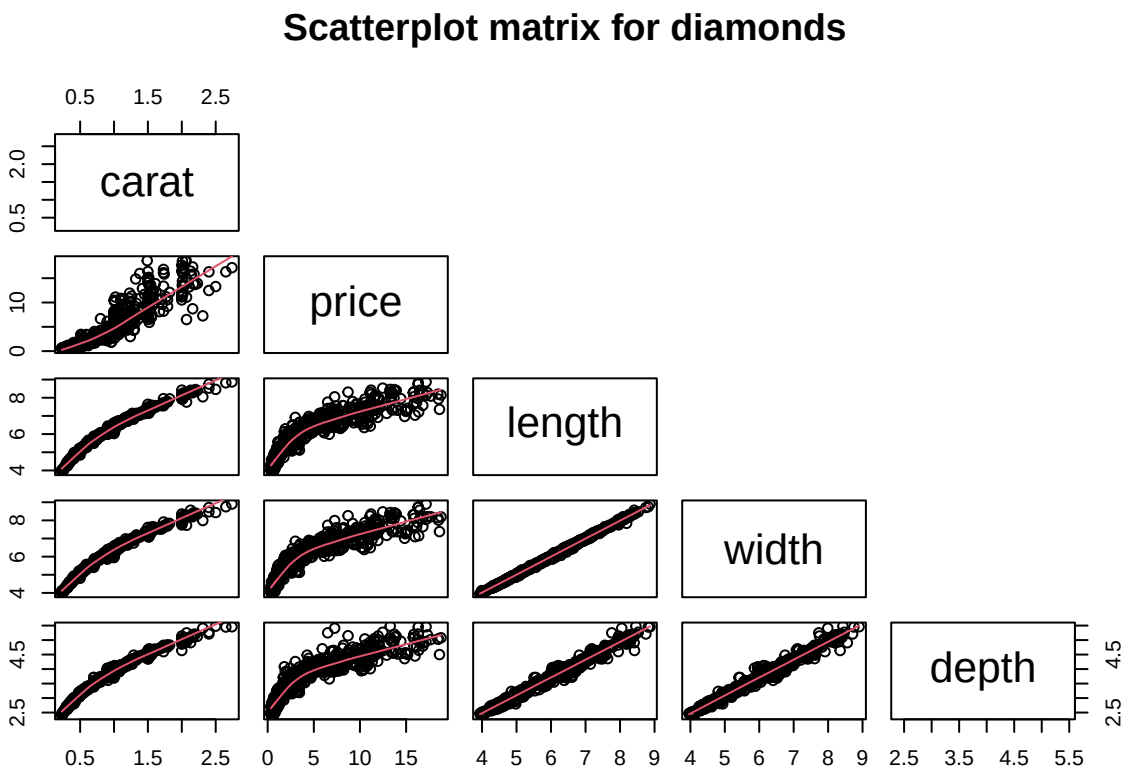
The carats range from 0.23 to 2.74, showing a very small range which corresponds to a quite low value for the standard deviation (0.48). On average, diamonds have 0.84 caarats. Half of them have less than 0.72 carats, and 75% of them is less than 1.06. We observe a very strong asymmetry of the distribution, with many diamonds having a low value for the carats, and (vice-versa) a small number of diamonds with a very high values for the carats. The price of diamonds is the variable with the highest variability. The average price is equal to 4180$, while the while the average variability around this value is approximately 4020$. This results in a coefficient of variation very close to 1: the variability of the price is about 100% of its mean. The most expensive diamonds reach the value of 18780$, while the cheapest ones only cost 350$. The variable is highly asymmetric (skewness equal to 1.49) with a very long right tail, representing diamonds with a very high price with respect to the rest of the distribution.

Concerning the dimensions of the diamonds, we observe that `length` and `width` take on very similar values, while `depth` is slightly inferior. On average, the considered diamonds are around 5.8 mm long; the median is very similar to this value, so that half of these diamonds are shorter than the average. Biggest diamonds almost reach 9 mm, while the smallest ones are slightly inferior to 4 mm. The variability is quite limited, as highlighted both by the range, and by the standard deviation: the average variability around the mean is equal to 1.1 mm. Computing the coefficient of variation we can emphasize that the variability is around 19% of the mean. The same comments hold also for the `width` variable. `depth` displays the smaller values among the three dimensions. None of the diamonds are more deep than 5.5 mm, and half of them do not exceed 3.5 mm. This is also the average value for a generic diamond. The distribution of each of these three variables (`length`, `width`, and `depth`) appears slightly asymmetric, presenting a heavy right tail, denoting the presence of a small number of diamods with particularly high dimensions.

Regarding the categorical variable `cut`, we observe that the majority of diamonds have the highest quality cut (313, corresponding to around 39%); on the contrary, only 33 diamonds (about 4%) have cuts of the lowest quality.

Finally we analyze the association between the continuous variables by displaying the scatterplot matrix.

```
pairs(diamonds[, -2],
      upper.panel = NULL,
      main = "Scatterplot matrix for diamonds",
      panel = panel.smooth)
```



We observe that all the variables are positively correlated. In particular, variables dealing with the dimensions of the diamonds (`length`, `width` and `depth`) appear to be strongly linearly associated. Further analyses are required to check the presence of collinearity among these three variables (variance inflation factor). The response variable `carat` is positively associated to all the explanatory variables, but the association seems not to be perfectly linear. In general, there is no presence of any points particularly distant from the rest of the distribution.

5.2.0.4 23.4 We use the `step()` function to perform model selection and identify a set of explanatory variables. We start with the complete model (retaining all the covariates) and using the option `direction = "both"` to perform both forward selection and backward elimination.

```
model23 <- lm(carat ~ ., data = diamonds)
step(model23, direction = "both")
```

```
## Start:  AIC=-3988.8
## carat ~ cut + price + length + width + depth
##
##          Df Sum of Sq    RSS     AIC
## - width   1   0.00363 5.3483 -3990.3
```

```
## <none>                5.3447 -3988.8
## - cut     4   0.09791 5.4426 -3982.3
## - length  1   0.06658 5.4113 -3980.9
## - depth   1   0.42124 5.7659 -3930.1
## - price   1   2.22049 7.5652 -3712.8
##
## Step:  AIC=-3990.26
## carat ~ cut + price + length + depth
##
##          Df Sum of Sq    RSS     AIC
## <none>                5.3483 -3990.3
## + width   1   0.00363 5.3447 -3988.8
## - cut     4   0.11920 5.4675 -3980.6
## - length  1   0.38083 5.7292 -3937.2
## - depth   1   0.41829 5.7666 -3932.0
## - price   1   2.26843 7.6168 -3709.4
```

```
##
## Call:
## lm(formula = carat ~ cut + price + length + depth, data = diamonds)
##
## Coefficients:
## (Intercept)         cut2         cut3         cut4         cut5        price
##    -1.14686     -0.03580     -0.04681     -0.03756     -0.05763      0.02966
##      length        depth
##     0.15915      0.27090
```

Starting from the complete model the value of the corresponding AIC index is equal to -3988.8. The function suggests then to exclude from the formulation of the model the `width` variable; this is associated to a very small amount of explained sum of squares (equal to 0.004) and its exclusion improves the AIC value up to -3990.3. This new model is finally selected according to the AIC index, since no other variable elimination (nor, obviously, the reintroduction of the `width` variable) helps to improve the value of the AIC index.

Remark: note that to check for the seemingly excessive collinearity between the two remaining explanatory variables `length` and `depth`, the model should also be evaluated from the VIF perspective.

5.2.0.5 23.5 We print the summary of the selected model.

```
model23_selected <- lm(data = diamonds[, -5], formula = carat ~ .)
summary(model23_selected)
```

```
##
## Call:
## lm(formula = carat ~ ., data = diamonds[, -5])
##
## 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 general, even though further (graphical) analyses are required, the theoretical assumptions seem to be fulfilled.

The estimated value for the intercept term is equal to $\hat{\beta}_0 = -1.15$: according to the model, the carats of a diamond with dimensions equal to 0 mm, with a price of \$0\$ and with cut of the worst quality (type 1) is predicted to have -1.15 carats (this interpretation has obviously no meaning in the practical context).

The corresponding standard error, equal to $\widehat{\text{SE}}_{\beta_0} = 0.036$, is used to assess the significance of the estimated parameter through the t-test. The test statistic is equal to

$$t = \frac{\hat{\beta}_0 - 0}{\widehat{\text{SE}}_{\beta_0}} = \frac{-1.15}{0.036} = -31.98,$$

and the p-value is extremely small ($\approx 10^{-16}$). We can reject the null hypothesis $H_0 : \beta_0 = 0$, concluding that the intercept term provides 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 $\beta_j = 0$ for $j = 5, 6, 7$. All three variables have a significant effect in the estimation of the diamonds' carats.

Regarding the interpretation of the corresponding (positive) estimated coefficients, we can, for example, state that the increase in the value of carats corresponding to a unitary increase in the price (i.e., +1000 units) is equal to 0.030 (if the other variables are kept fixed). Similarly, 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 taking 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 diamond with cut quality 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 t-tests.

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:

$$\frac{0.082}{0.84} \approx 0.098,$$

meaning that on average the percentage 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 around 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 coefficients as null hypothesis (the alternative hypothesis considers instead that at least one coefficient is statistically significant). The value of the F statistic is equal to 3785; the associated p-value is very close to 0 ($< 10^{-16}$), leading us to reject the null hypothesis at any conventional confidence level.

5.2.0.6 23.6 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 X_1, X_2 and X_3 the continuous explanatory variables `price`, `length` and `depth`, respectively, we obtain the following equations:

$\hat{Y} = -1.147 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3$, if cut quality is 1$\hat{Y} = -1.147 - 0.036 + 0.030 \cdot X_1 + 0.159 \cdot X_2 + 0.271 \cdot X_3$, if cut

Remark: geometrically, each equation represents a hyperplane of dimension 3 in the space $\mathbb{R}^4$ (that's why it is not possible to graphically represent them). All five hyperplanes are "parallel" to each other.

5.2.0.7 23.7 We use the `confint()` function to compute the confidence interval for each of the estimates regression coefficients β_j.

```
round(confint(model23, level = 0.9), 3)
```

```
##                5 %   95 %
## (Intercept) -1.204 -1.086
## cut2        -0.062 -0.002
## cut3        -0.072 -0.013
## cut4        -0.064 -0.008
## cut5        -0.082 -0.026
## price        0.027  0.033
## length       0.097  0.311
## width       -0.152  0.058
## depth        0.215  0.329
```

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.

5.2.0.8 23.8 To obtain a prediction for a new observation out of sample, we can simply use the `predict()` function, which requires the reference model and the values of the covariates for the new observation. These values should be specified in advance in the form of a new dataframe.

Hence, let us consider a new (out-of-sample) diamond with cut of type 4, price equal to 13000$, length and depth equal to 6.00 and 4.00mm, respectively.

```r
mean_width <- mean(diamonds$width, na.rm = TRUE)

new23 <- data.frame(cut = factor(4, levels = levels(diamonds$cut)),
                    price = 13.000,
                    length = 6.00,
                    depth = 4.00,
                    width = mean_width)

new23
```

```
##   cut price length depth   width
## 1   4    13      6     4 5.83105
```

```r
round(predict(model23, newdata = new23, interval = "prediction", level = 0.95), 3)
```

```
##     fit   lwr   upr
## 1 1.248 1.082 1.415
```

```r
round(predict(model23, newdata = new23, interval = "confidence", level = 0.95), 3)
```

```
##     fit   lwr  upr
## 1 1.248 1.207 1.29
```

- The point prediction is (obviously) the same with both specifications: for a new observation with specified covariate values, a value for the carats equal to 1.24 is predicted.
- Regarding the interval estimation, the prediction interval is quite narrow, with a lower bound of 1.082 and an upper bound of 1.415. 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 falls within the range of variation of the response variable.
- The confidence interval is even narrower (it does not take into account the variability of the new observation). In this case, the average carats for a diamond with the specified covariates falls between 1.207 and 1.29 with a 95% confidence level.

5.3 Basic probability calculation: estimating the chance of selecting a tolerable song from a playlist using simple relative frequency.

Your playlist of 200 songs has 5 which you cannot stand. What is the probability that when you hit shuffle a song you tolerate comes on?

The probability that a song I tolerate comes on is:

$$P(X = t) = \frac{\text{good songs}}{\text{bad songs}} = \frac{200 - 5}{200} = \frac{195}{200}$$

```r
195/200
```

```
## [1] 0.975
```

So, $P(X = t) = 0.975$.

5.4 Binomial probability calculation: chance of guessing exactly 2 correct answers out of 10 questions with 4 choices each.

While taking a multiple choice test, a student encountered 10 problems where she ended up completely guessing, randomly selecting one of one of the four options. What is the chance that she got exactly 2 of the 10 correct?

This is a binomial probability problem. We have:

- $n = 10$ questions
- $p = \frac{1}{4} = 0.25$ (probability of guessing the right answer between the 4 options)

We want to compute the probability of getting exactly $k = 2$ correct answers.

The binomial distribution formula is:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

So it is:

```
prob25 <- dbinom(2, size = 10, prob = 0.25)

round(prob25, 3)
```

```
## [1] 0.282
```

$P(X = 2) = 0.282$

5.5 Research design for chemical exposure and cancer: define binary response (cancer diagnosis) and list relevant covariates to control for confounding in analysis.

List possible covariates and response variable needed to respond to the following research question: is exposure to a particular chemical associate with a cancer diagnosis?

To address the research question *"Is exposure to a particular chemical associated with a cancer diagnosis?"*, the **response variable** and several **possible covariates** should be clearly identified to control for confounding effects and assess the strength of the association.

Response variable:

- `cancer_diagnosis`: a binary indicator (1 = diagnosed with cancer, 0 = not diagnosed) Alternatively, if more detail is available, this could be a categorical variable indicating type or stage of cancer.

Main explanatory variable (of interest):

- `chemical_exposure`: binary (exposed/not exposed), ordinal (low/medium/high), or continuous (measured concentration in blood, air, etc.)

Potential covariates (control variables):

- `age`: continuous or grouped (e.g., <40, 40–60, >60)

- `sex`: binary or categorical
- `smoking_status`: binary or categorical (never, former, current)
- `alcohol_consumption`: binary or continuous
- `occupational_exposure`: binary or categorical (e.g., industrial worker, office worker)
- `diet_quality`: ordinal or score-based
- `family_history_cancer`: binary (presence of family history)
- `socioeconomic_status`: ordinal or proxy (e.g., income, education level)
- `region_of_residence`: categorical (urban/rural or by geographic area)
- `duration_of_exposure`: continuous (e.g., years exposed)
- `body_mass_index (BMI)`: continuous or grouped
- `physical_activity`: ordinal (e.g., sedentary, moderate, active)

These covariates help reduce bias and allow for causal inference if the model assumptions hold or the study is well-designed (e.g., prospective cohort).

5.6 Contingency table analysis: compute odds and odds ratio to assess association between IVF use and congenital disabilities, with interpretation and visualization.

A study involved 9584 babies with congenital disabilities and 4792 babies without. Among the mothers of babies without congenital disabilities, 1.1% had used IVF (which is a technique in which sperm are injected directly into eggs) compared with 2.4% of mothers of babies with congenital disabilities. - Construct a contingency table and calculate the corresponding odds and odds ratios. - Comment on the reported and calculated results.

Construct the contingency table

```
disability_ivf <- matrix(c(230, 9354, 53, 4739),
                         nrow = 2,
                         byrow = TRUE)

colnames(disability_ivf) <- c("IVF", "No IVF")
rownames(disability_ivf) <- c("Disabilities", "No disabilities")

disability_ivf_table <- as.table(disability_ivf)
disability_ivf_table
```

```
##                  IVF No IVF
## Disabilities     230   9354
## No disabilities   53   4739
```

Calculate odds

- Odds of IVF use among babies with disabilities:

$$\text{Odds}_{\text{disabilities}} = \frac{230}{9354}$$

- Odds of IVF use among babies without disabilities:

$$\text{Odds}_{\text{no disabilities}} = \frac{53}{4739}$$

Compute the odds ratio (OR)

$$\text{OR} = \frac{\text{Odds}_{\text{disabilities}}}{\text{Odds}_{\text{no disabilities}}}$$

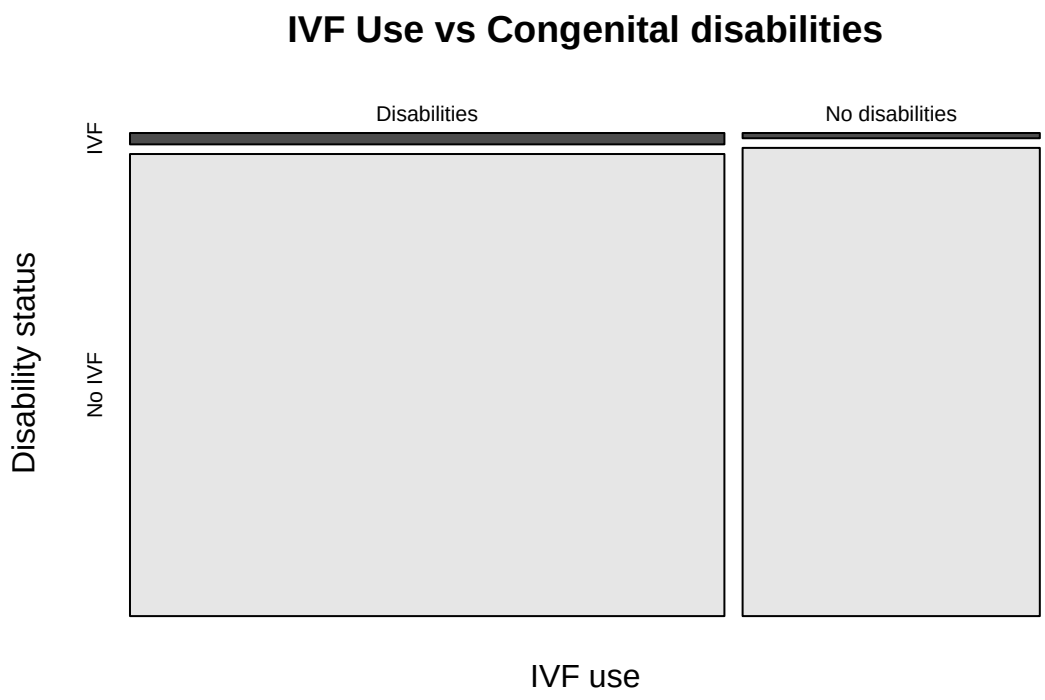
```
odds_disabilities <- 230 / 9354
odds_no_disabilities <- 53 / 4739

odds_ratio <- odds_disabilities / odds_no_disabilities
round(odds_ratio, 2)
```

```
## [1] 2.2
```

The odds ratio of OR $=\approx 2.2$ suggests that the odds of IVF use among mothers of babies with congenital disabilities are more than twice the odds among mothers of babies without. This highlights a potential association between IVF and congenital disabilities.

However, causality **cannot** be concluded from this alone. The observed association may be confounded by other factors (e.g., maternal age, genetics, IVF techniques, etc.).

```
mosaicplot(disability_ivf_table,
           main = "IVF Use vs Congenital disabilities",
           color = TRUE,
           xlab = "IVF use",
           ylab = "Disability status")
```



5.7 Logistic regression on lung cancer: binary and quantitative covariate analysis, model selection, coefficient interpretation, odds ratio confidence interval, prediction, classification performance, and ROC curve.

The data in `bird.Rdata` refer to a 1985 case-control study of patients. Each patient with cancer was matched with two control subjects (without cancer) by age and sex. The researchers wished to determine whether, after age, sex, socio-economic status, and smoking have been controlled for, the additional risk associated with bird keeping. The variables are the following:

- `female` (coded as 1 for female and 2 for male)
- `age` in years
- `highstatus` socioeconomic status (coded as 1 for high, coded as 0 for low)
- `yrsmoke` years of smoking prior to diagnosis
- `bird` indicator of birdkeeping (coded as 1 for yes and 0 for no) determined whether or not there were caged birds in the home for more than 6 consecutive months from 5 to 14 years before diagnosis or examination
- `cancer` indicator of lung cancer diagnosis (1 cancer, 0 no cancer)

We load the dataset and briefly inspect its first six rows.

```
load("bird.Rdata")
head(birds)
```

```
##   female age highstatus yrsmoke cigsday bird cancer
## 1      0  37          0      19      12    1      1
## 2      0  41          0      22      15    1      1
## 3      0  43          1      19      15    0      1
## 4      0  46          0      24      15    1      1
## 5      0  49          0      31      20    1      1
## 6      0  51          1      24      15    0      1
```

These rows concern 6 males, aged between 37 and 51 years old. All of them have been smokers for many years (at least 19) and currently smoke more than 10 cigarettes per day. All of them are suffering from lung cancer.

5.7.0.1 28.1 Provide descriptive statistics for binary indicators to check the association of each covariate with the response of interest (cancer). Comment on the results and try to provide possible reasons for absence or presence of association.

Starting with the binary explanatory variable `female`, we first compute the absolute frequencies table.

```
table(Female = birds$female, Cancer = birds$cancer)
```

```
##       Cancer
## Female  0  1
##      0 74 37
##      1 24 12
```

We observe the presence of 36 women and 111 men. Both for male and female observations, the number of people with cancer is exactly half of those without cancer.

We compute now the table of relative frequencies (obtained by dividing each element in the previous table by the total number of observations).

```
round(prop.table(table(Female = birds$female, Cancer = birds$cancer)), 4)
```

```
##       Cancer
## Female      0      1
##      0 0.5034 0.2517
##      1 0.1633 0.0816
```

The provided information is essentially the same as in the previous table. We notice that more than a half of the observations are men without cancer; women with cancer are only 8% of the total.

Finally we compute the frequencies relative to (conditionally) the row totals (this is obtained through the `margin = 1` argument).

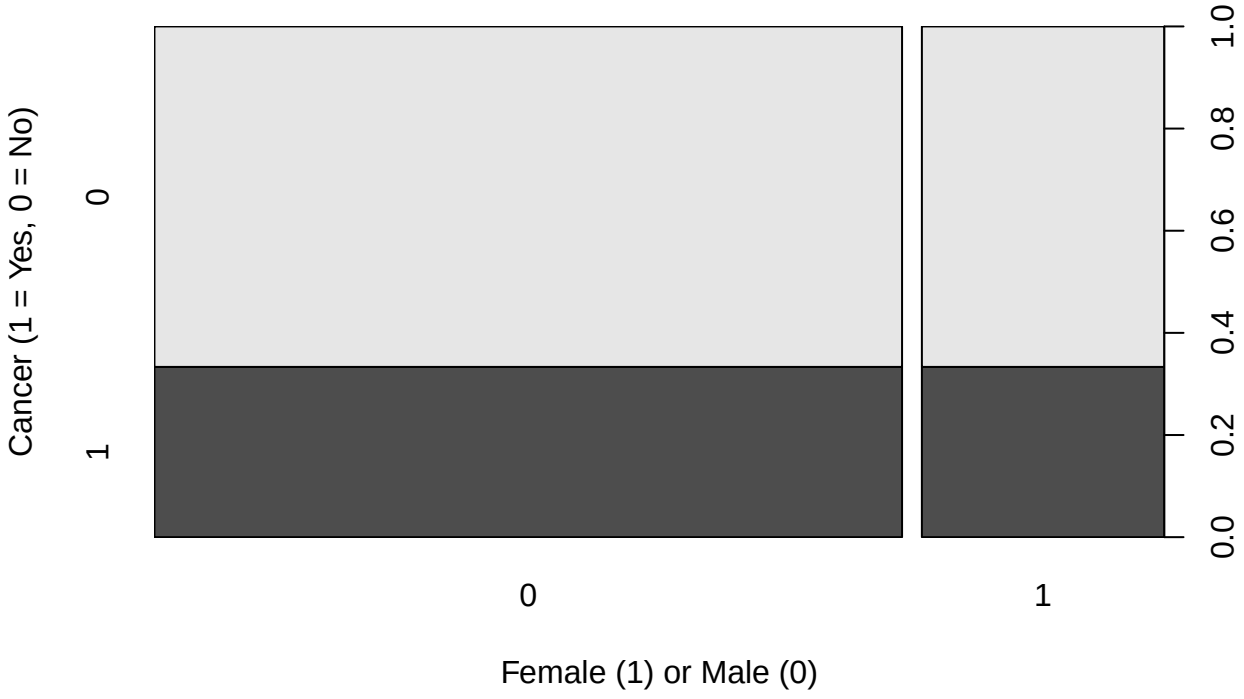
```
round(prop.table(table(Female = birds$female, Cancer = birds$cancer),
                 margin = 1), 4)
```

```
##       Cancer
## Female      0      1
##      0 0.6667 0.3333
##      1 0.6667 0.3333
```

This table shows the percentage of individuals with or without cancer among men (first row) and among women (second row). We note that, as previously observed, this percentage is exactly the same between the two groups of individuals, with approximately 33% of subjects having lung cancer and 67% not having lung cancer.

Note that the same results may also be obtained by plotting the frequencies of two variables.

```
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%.

Finally, we can calculate the odds (and odds ratio) to quantify the presence of cancer patients among women and men.

```
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.

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

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

```r
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, corresponding to around 23% of the total. Only 16 subjects (about 11% of the total) have cancer without having caged birds.

```r
round(prop.table(table(Birds = birds$bird, Cancer = birds$cancer), margin = 1), 3)
```

```
##      Cancer
## Birds     0     1
##     0 0.800 0.200
##     1 0.507 0.493
```

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.

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

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

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

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

```r
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.

Finally, we compute the same procedure for the `highstatus` variable.

```r
table(Status = birds$highstatus, Cancer = birds$cancer)
```

```
##       Cancer
## Status  0  1
##      0 65 37
##      1 33 12
```

```r
round(prop.table(table(Status = birds$highstatus, Cancer = birds$cancer)), 4)
```

```
##       Cancer
## Status      0      1
##      0 0.4422 0.2517
##      1 0.2245 0.0816
```

Only 12 individuals (corresponding to the 8% of the total) have cancer with an high socio-economic status. The number of individuals with cancer increase to 37 (25%) if the social status is low.

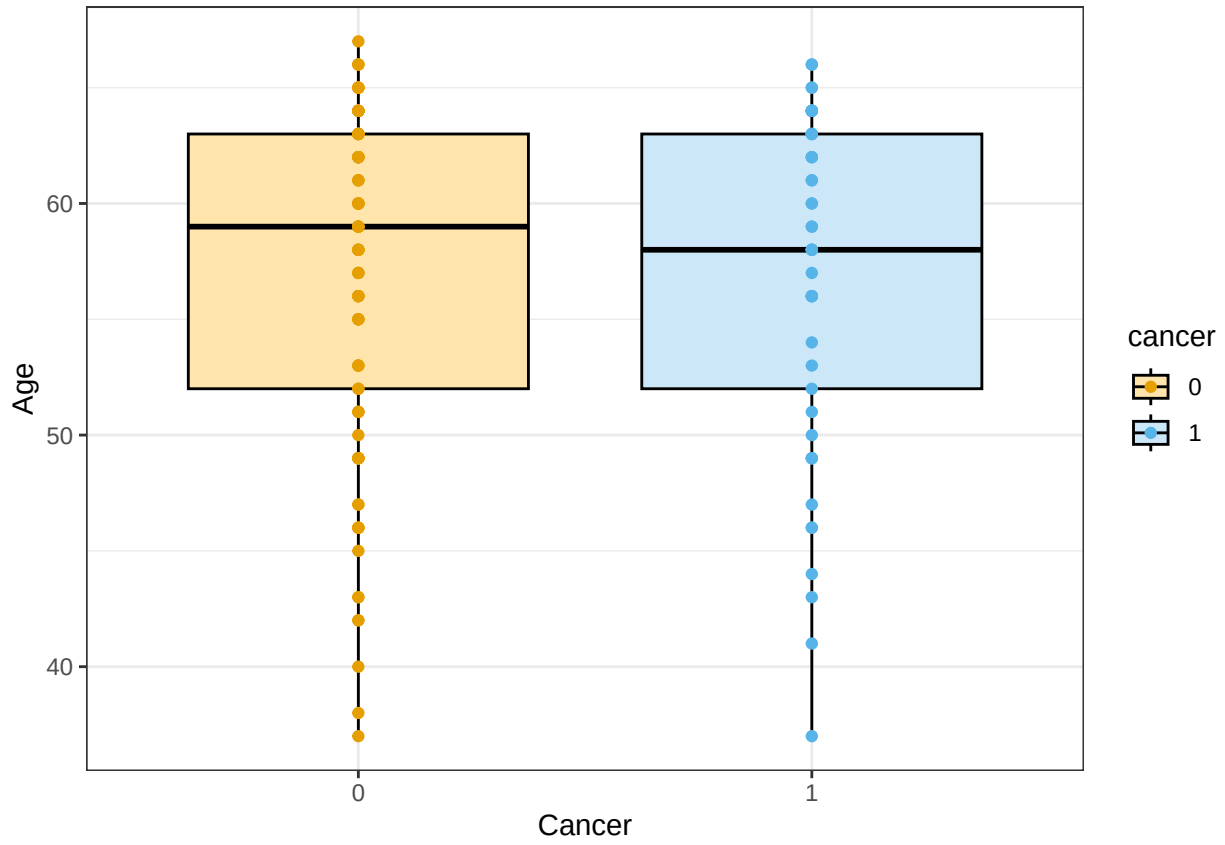
```r
round(prop.table(table(Status = birds$highstatus, Cancer = birds$cancer), margin = 1), 3)
```

```
##       Cancer
## Status     0     1
##      0 0.637 0.363
##      1 0.733 0.267
```

5.7.0.2 28.2 For quantitative variables produce a boxplot and summary statistics by cancer diagnosis.

Considering initially the continuous explanatory variable `age`, we represent a box plot for each level of the repsonse variable.

```r
require(ggplot2)
ggplot(birds, aes(x = cancer, y = age)) +
geom_boxplot(aes(fill = cancer), col = "black") +
geom_point(aes(col = cancer)) +
xlab("Cancer") + ylab("Age") +
scale_fill_manual(values=c("#ffe5ab", "#cce8f8")) +
scale_color_manual(values=c("#E69F00", "#56B4E9")) +
theme_bw()
```



The sample age distribution is almost identical among individuals with and without cancer, highlighting that there are not significant age differences between ill and healthy individuals. Only the median seems to be slightly lower for people having cancer; the difference is anyway negligible.

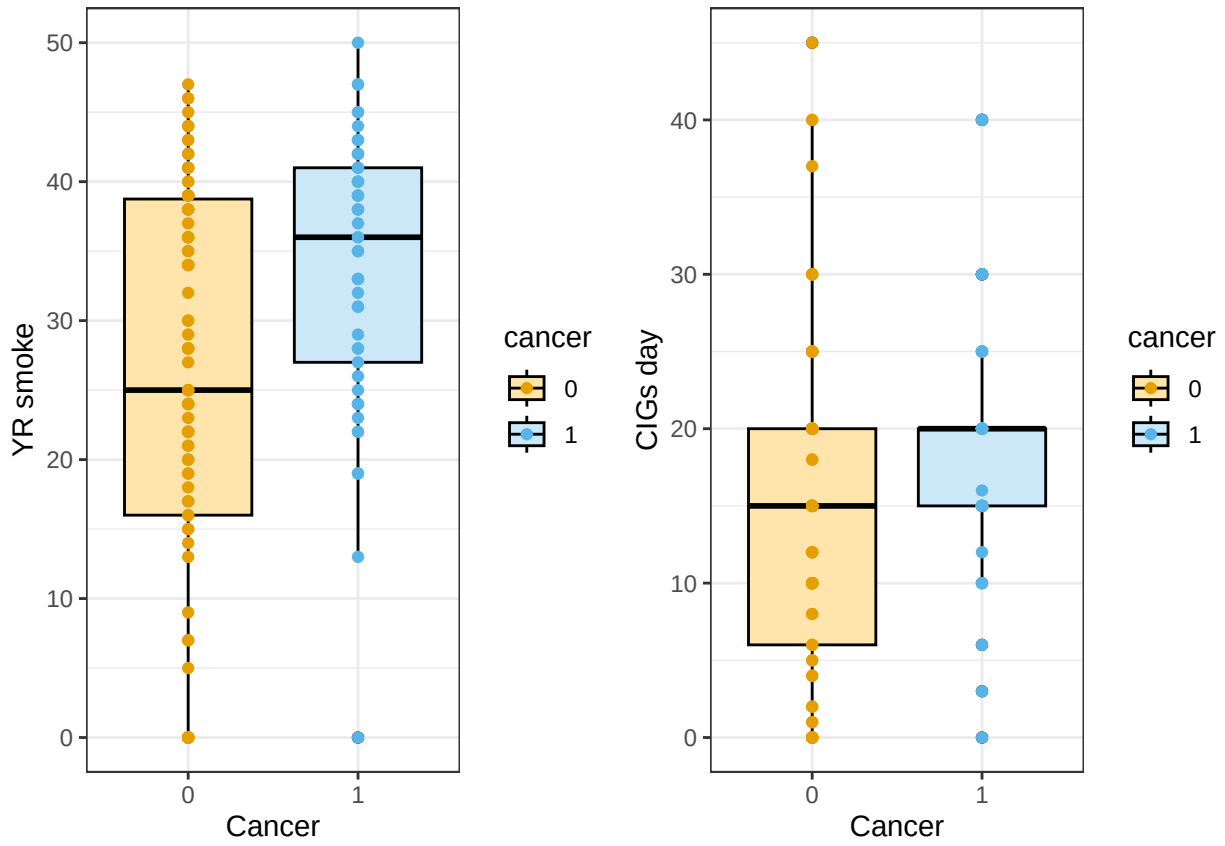
On the contrary, we expect the two explanatory variables related to smoke (`yrsmoke` and `cigsday`) to be associated with lung cancer.

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
p1 <- ggplot(birds, aes(x = cancer, y = yrsmoke)) +
geom_boxplot(aes(fill = cancer), col = "black") +
geom_point(aes(col = cancer)) +
xlab("Cancer") + ylab("YR smoke") +
scale_fill_manual(values=c("#ffe5ab", "#cce8f8")) +
scale_color_manual(values=c("#E69F00", "#56B4E9")) +
theme_bw()

p2 <- ggplot(birds, aes(x = cancer, y = cigsday)) +
geom_boxplot(aes(fill = cancer), col = "black") +
geom_point(aes(col = cancer)) +
xlab("Cancer") + ylab("CIGs day") +
scale_fill_manual(values=c("#ffe5ab", "#cce8f8")) +
scale_color_manual(values=c("#E69F00", "#56B4E9")) +
theme_bw()
```

```
grid.arrange(p1, p2, ncol = 2)
```



In the sample, the subjects having cancer are, on average, more likely to be long time smokers than individuals who do not have lung cancer. In particular, observing the median, we highlight that half of the subjects without cancer have smoken for less than (around) 25 years, while half of the subjects having cancer have smoked for a time less than 35 years. Also the other quartiles (included maximum and minimum) follow the same trend. Finally, we detect the presence of only one subject who has never smoked, but having lung cancer.

Considering the number of cigarettes smoked in a day, we notice the same difference between people having or not a lung cancer. For example we observe that 50% of subject having cancer smokes more than 20 cigarettes per day (median), while this percentage reduces to 25% for individuals not having cancer (third quartiles). Despite this general behavior, we highlight that the subject smoking the highest number of cigarettes (more than 40 per day) has not lung cancer.

This behavior is also confirmed by the descriptive statistics analyzed in the following.

```
by(birds[, c(2, 4, 5)], birds$cancer, summary)
```

```
## birds$cancer: 0
##       age         yrsmoke         cigsday
##  Min.   :37   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:52   1st Qu.:16.00   1st Qu.: 6.00
##  Median :59   Median :25.00   Median :15.00
##  Mean   :57   Mean   :24.96   Mean   :14.21
##  3rd Qu.:63   3rd Qu.:38.75   3rd Qu.:20.00
```

```
##  Max.   :67   Max.   :47.00   Max.   :45.00
## ------------------------------------------------------------
## birds$cancer: 1
##       age          yrsmoke         cigsday
##  Min.   :37.0   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:52.0   1st Qu.:27.00   1st Qu.:15.00
##  Median :58.0   Median :36.00   Median :20.00
##  Mean   :56.9   Mean   :33.63   Mean   :18.82
##  3rd Qu.:63.0   3rd Qu.:41.00   3rd Qu.:20.00
##  Max.   :66.0   Max.   :50.00   Max.   :40.00
```

Focusing on the comparisons of means, not displayed by the previous boxplots, we observe that age does not show significant differences: on average, cancer subjects are 56.9 years old, while those without cancer are 57 years old. On the contrary, variables related to smoking show significant differences. A cancer subject has spent on average 34 years smoking, about 9 more than an individual without cancer. Similarly, a cancer patient smokes on average 4.5 more cigarettes per day.

5.7.0.3 28.3 Fit a logistic regression model with cancer as response and perform variable selection. Comment on the results obtained according to the Akaike information criterion. State a conclusion in context.

We initially consider the complete model, in which all the explanatory variables are used to explain the presence or absence of lung cancer. We perform model selection using the `step()` function.

```
model28 <- glm(cancer ~ ., family = binomial(), data = birds) # glm -> generalized linear model, "." m

summary(model28)
```

```
##
## Call:
## glm(formula = cancer ~ ., family = binomial(), data = birds)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## female1      0.56127    0.53116   1.057 0.290653
## age         -0.03976    0.03548  -1.120 0.262503
## highstatus1  0.10545    0.46885   0.225 0.822050
## yrsmoke      0.07287    0.02649   2.751 0.005940 **
## cigsday      0.02602    0.02552   1.019 0.308055
## bird1        1.36259    0.41128   3.313 0.000923 ***
## ---
## 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: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```

```
step(model28)
```

```
## Start:  AIC=168.2
## cancer ~ female + age + highstatus + yrsmoke + cigsday + bird
##
##              Df Deviance    AIC
## - highstatus  1   154.25 166.25
## - cigsday     1   155.24 167.24
## - female      1   155.32 167.32
## - age         1   155.49 167.49
## <none>            154.20 168.20
## - yrsmoke     1   163.93 175.93
## - bird        1   165.87 177.87
##
## Step:  AIC=166.25
## cancer ~ female + age + yrsmoke + cigsday + bird
##
##           Df Deviance    AIC
## - female   1   155.32 165.32
## - cigsday  1   155.32 165.32
## - age      1   155.50 165.50
## <none>         154.25 166.25
## - yrsmoke  1   164.09 174.09
## - bird     1   165.90 175.90
##
## Step:  AIC=165.32
## cancer ~ age + yrsmoke + cigsday + bird
##
##           Df Deviance    AIC
## - cigsday  1   156.22 164.22
## - age      1   156.75 164.75
## <none>         155.32 165.32
## - yrsmoke  1   164.18 172.18
## - bird     1   168.35 176.35
##
## Step:  AIC=164.22
## cancer ~ age + yrsmoke + bird
##
##           Df Deviance    AIC
## - age      1   158.11 164.11
## <none>         156.22 164.22
## - bird     1   168.83 174.83
## - yrsmoke  1   172.53 178.53
##
## Step:  AIC=164.11
## cancer ~ yrsmoke + bird
##
##           Df Deviance    AIC
## <none>         158.11 164.11
## - yrsmoke  1   172.93 176.93
## - bird     1   173.17 177.17

##
```

```
## Call:  glm(formula = cancer ~ yrsmoke + bird, family = binomial(), data = birds)
##
## Coefficients:
## (Intercept)      yrsmoke        bird1
##    -3.18016      0.05825      1.47555
##
## Degrees of Freedom: 146 Total (i.e. Null);  144 Residual
## Null Deviance:       187.1
## Residual Deviance: 158.1     AIC: 164.1
```

The initial (complete) model is associated to a value of AIC index equal to 168.2. At the first step of the procedure, the function suggests that excluding each one of the variables `highstatus`, `cigsday`, `female`, and `age` ensures a model with a better fit and parsimony. It then selects a new model, which contains all explanatory variables apart from `highstatus`. The new value of the AIC index is equal to 166.25. In the same way, during the successive three steps of the procedure, the variables `cigsday`, `female`, and `age` are removed from the model. The final model keeps only two explanatory variables: `yrsmoke` and `bird`. The associated AIC index is equal to 164.1.

5.7.0.4 28.4 Carefully interpret each regression coefficient of the selected model at the previous step.

We print the summary of the selected model.

```
model28_selected <- glm(cancer ~ yrsmoke + bird, family = binomial(), data = birds)

summary(model28_selected)
```

```
##
## Call:
## glm(formula = cancer ~ yrsmoke + bird, family = binomial(), data = birds)
##
## 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
```

Before commenting on the estimated coefficient, we provide a brief explanation of the other information reported here.

- 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 $p_i = \mathbb{P}(Y_i = 1)$ is the estimated probability of having cancer. We have the following expression:

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

or, after applying the exponential function:

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

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, compared 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(model28_selected))
```

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

- The estimated parameter for the `yrsmoke` covariate is equal to $\hat{\beta}_1 = 00.058$, with $\exp(\hat{\beta}_1) = 1.060$. This value represents the (multiplicative) effect on the odds of having 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 multiplied by 1.060, thus slightly increasing.
- The estimated parameter for the `bird` variable is equal to $\hat{\beta}_2 = 1.476$, with $\exp(\hat{\beta}_2) = 4.373$. 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.

5.7.0.5 28.5 Provide a confidence interval for the odds ratio related to the estimated coefficient of bird. Comment on the result.

To compute the confidence interval of a certain estimated regression coefficient, we use the `confint()` function. The argument `parm = 3` allows us to select only the third estimated coefficient.

```
round(confint(model28_selected, parm = 3), 3)
```

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

```
##  2.5 % 97.5 %
##  0.718  2.277
```

We observe that the confidence interval does not include 0, highlighting that the corresponding variable is significant. We can also compute the exponential of the estimated confidence interval.

```
exp(confint(model28_selected, 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.

5.7.0.6 28.6 Can we conclude that birdkeeping causes increased odds of developing lung cancer?

In conclusion, birdkeeping seems to cause increased odds of developing lung-cancer (according to our model). Indeed, the odds of having lung-cancer is about 4.37 (OR) times higher for people having caged birds.

5.7.0.7 28.7

The estimated probabilities of having cancer ($\widehat{p}_i$) can be calculated by reversing the equation of the model. For the sample units, these probabilities are directly provided in the output of the `glm()` function.

```
prob28 <- model28_selected$fitted.values
round(head(prob28), 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 0.525. 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.

More generally, we can analyze the distribution of the estimated probabilities through the main descriptive statistics.

```
summary(prob28)
```

```
##    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.

```
birds[which.min(prob28), ]
```

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

```
birds[which.max(prob28), ]
```

```
##    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.

5.7.0.8 28.8 To provide the classification table we first binarize the predicted probabilities for the response variable. Choosing 0.5 as a cutoff, we set $\hat{y}_i = 0$ if $\hat{p}_i < 0.5$ and $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.5$.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dt28 <- data.frame(True = birds$cancer, Prob = model28_selected$fitted.values)
dt28 <- dt28 %>%
  mutate(Pred = ifelse(Prob < 0.5, 0, 1))

table(True = dt28$True, Pred = dt28$Pred)
```

```
##     Pred
## True  0  1
##    0 85 13
##    1 27 22
```

The classification performed according to the estimated logistic model correctly predicts 22 cases of presence of cancer and 85 cases of absence of cancer. Overall, it correctly predicts 107 cases over a total of 147. The correct classification rate is then equal to 0.73 (misclassification rate 0.27).

- **Sensitivity (true positive rate)**: among the 49 subjects having lung cancer, the estimated model correctly classifies only 22; sensitivity is computed as the ratio $\frac{22}{49} = 0.45$. It is a rather low value: less than half of the subject with "positive" category of the response variable is correctly classified.
- **Specificity (true negative rate)**: among the 98 subjects not having lung cancer, the estimated model correctly classifies 85; specificity is computed as the ratio $\frac{85}{98} = 0.87$. It is a quite high value: almost 90% of the subject with "negative" category of the response variable is correctly classified.

5.7.0.9 28.9 In the previous point, we have fixed to 0.5 the value of the cutoff to distinguish $\hat{y}_i = 1$ from $\hat{y}_i = 0$. In this point we use the Receiving Operating Characteristics (ROC) and the corresponding Area Under the Curve (AUC) to assess the classification performance of the estimated model without fixing a specific cutoff.

```
require(ROCit)
```

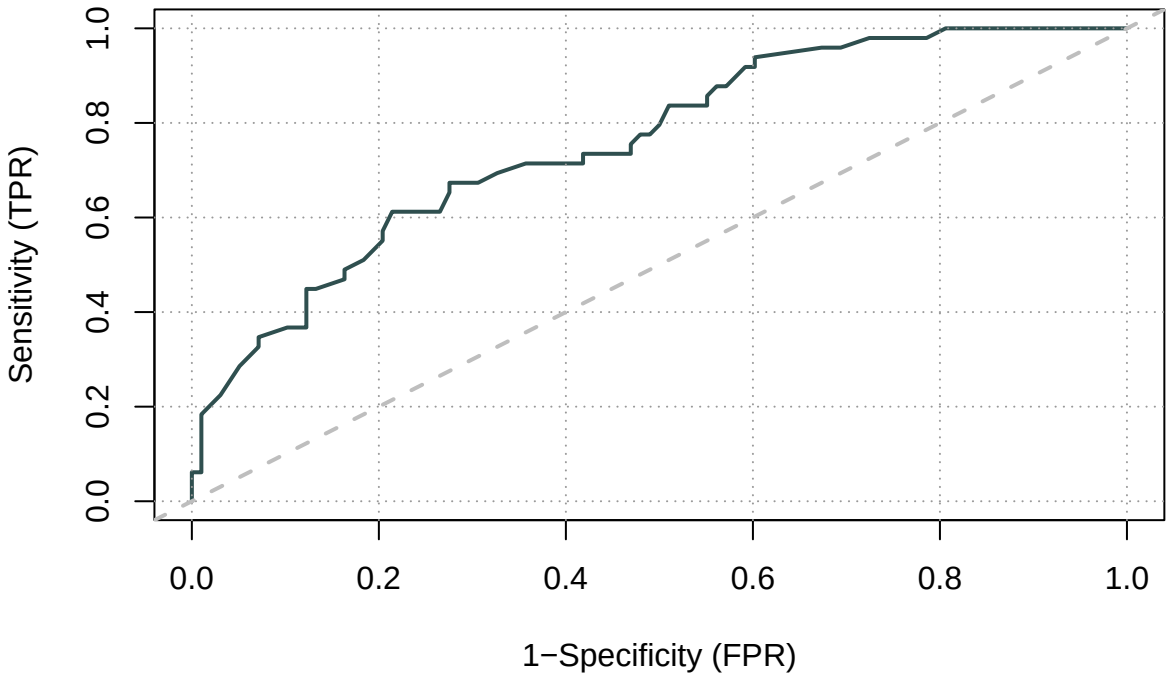
```
## Loading required package: ROCit
```

```
##
## Attaching package: 'ROCit'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

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

```
ROC.obj28 <- rocit(score = dt28$Prob, class = dt28$True)

plot(ROC.obj28, YIndex = FALSE, legend = FALSE)
```




```
ROC.obj28$AUC
```

```
## [1] 0.7582257
```

6 Week 8 - Gaussian mixture models for clustering (univariate and multivariate cases), Clustering with multivariate Gaussian mixtures under spherical covariance constraints, Gaussian mixture discriminant analysis for binary classification (EDDA framework), Model evaluation: train-test split, Cross-validation, ROC and AUC

6.1 Finite mixture modeling with Mclust: data exploration, model selection via BIC, parameter interpretation, classification, and visual assessment of clusters and density fit.

Consider the data in the file called `haemoglobin.Rdata`. These are the values observed on 70 individuals of a plasma component. Supposing potential groups of people with distinct values of such a component, we are interested in estimating these latent subpopulations.

6.1.0.1 30.1 Describe the data using descriptive statistics and graphics. Descriptive statistics are used to describe the data by analyzing its main characteristics.

```
load("haemoglobin.Rdata")
skim_without_charts(haemoglobin)
```

Table 23: Data summary

Name	haemoglobin
Number of rows	70
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

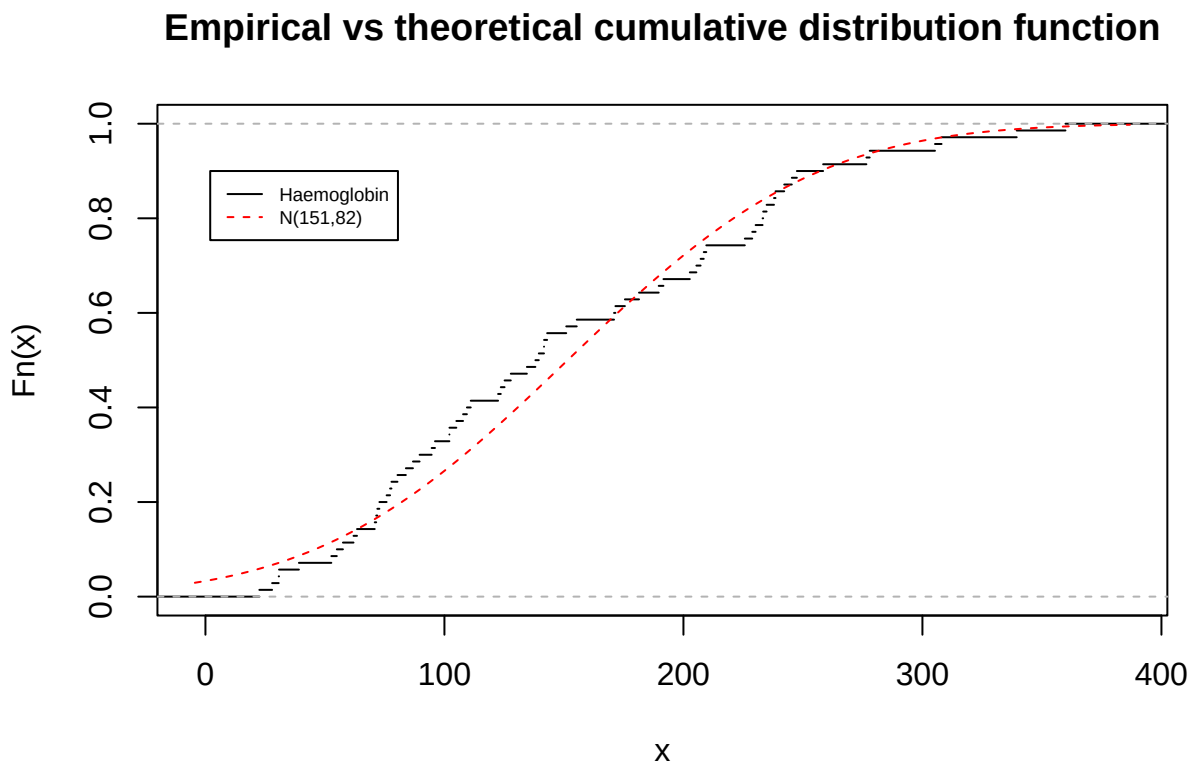
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
data	0	1	151.55	82.48	22.55	81.24	138.77	221.66	359.84

The dataset contains 70 rows (70 observations) and a single column, which records the observed haemoglobin. On average, the observations have a hemoglobin value of 152. The range of values is quite large, with an approximate minimum of 23 and a maximum of around 360. The variability is also quite high, as indicated by the value of the standard deviation: on average, the observations have a deviation from the mean of 82.5. Approximately calculating the coefficient of variation $\text{CV} = \frac{\sigma}{\mu}$, this translates to a variability of about 50% of the mean. The median is slightly different from the mean, indicating a possible asymmetry in the distribution.

We also provide a graphical representation of the data using the empirical distribution function. We compare this curve to the theoretical model of the normal distribution with mean and standard deviation equal to

those of the sample. It is clear that this will probably be an imprecise approximation, since the data come from a Gaussian mixture rather than a Gaussian variable.

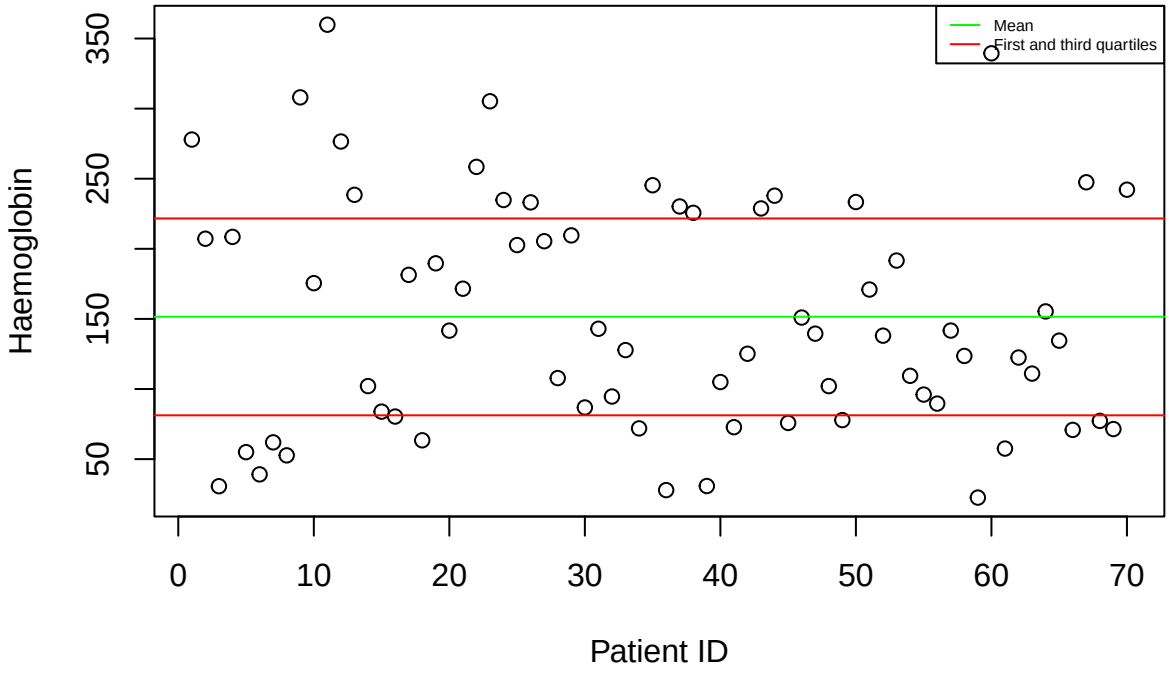
```
plot(ecdf(haemoglobin),
     do.points = FALSE,
     main = "Empirical vs theoretical cumulative distribution function",
     col = "black")

curve(pnorm(x, mean = mean(haemoglobin), sd = sd(haemoglobin)),
      add = TRUE,
      lty = 2,
      col = "red")

legend(2, 0.9,
       col = c("black", "red"),
       c("Haemoglobin", "N(151,82)"),
       lty = c(1, 2),
       cex = 0.6)
```



We observe that the two curves, empirical and theoretical, show some significant deviations; in particular, we note that the distribution of the observed data has a lighter left tail (fewer observations) than expected, while the right tail is consistent with the assumption of normality. Finally, we observe that the central region of the empirical distribution has more observations than expected (again, assuming normality).

Finally, we provide a graphical representation of the values on a Cartesian plane (scatter plot against the unit ID), comparing them with the mean and first and third quartiles of the distribution.

```r
plot(haemoglobin,
     xlab = "Patient ID",
     ylab = "Haemoglobin")

q <- quantile(haemoglobin, c(0.25,0.75))
m <- mean(haemoglobin)

abline(h = c(m, q[1], q[2]),
       col = c("green", "red", "red"))

legend("topright",
       c("Mean", "First and third quartiles"),
       col = c("green", "red"),
       lty = c(1, 1), cex = 0.5)
```



The points are randomly distributed on the plane (no particular systematic or structural patterns are evident). We note a slight asymmetry in their arrangement with respect to the mean value: in the lower part of the plot, there is a lower dispersion of points compared to the upper zone: the points are less distant from the third quartile line.

6.1.0.2 30.2 Estimate finite mixture model using 'mclust' package and performing model selection to choose the suitable number of clusters and the most parsimonious parameterization for the variance of each component through the Bayesian information criterion. The function `mclust::mclustBIC()` is used to estimate mixture models for an increasing number of components (from 1 to 9). By default, the function considers both the models that assume equal variance among the different

components (the more parsimonious model, specified with the letter "E") and the models with specific variance for each component (specified with the letter "V"). All estimated models are compared using the BIC criterion to select those that have a better fit.

```
require(mclust)
```

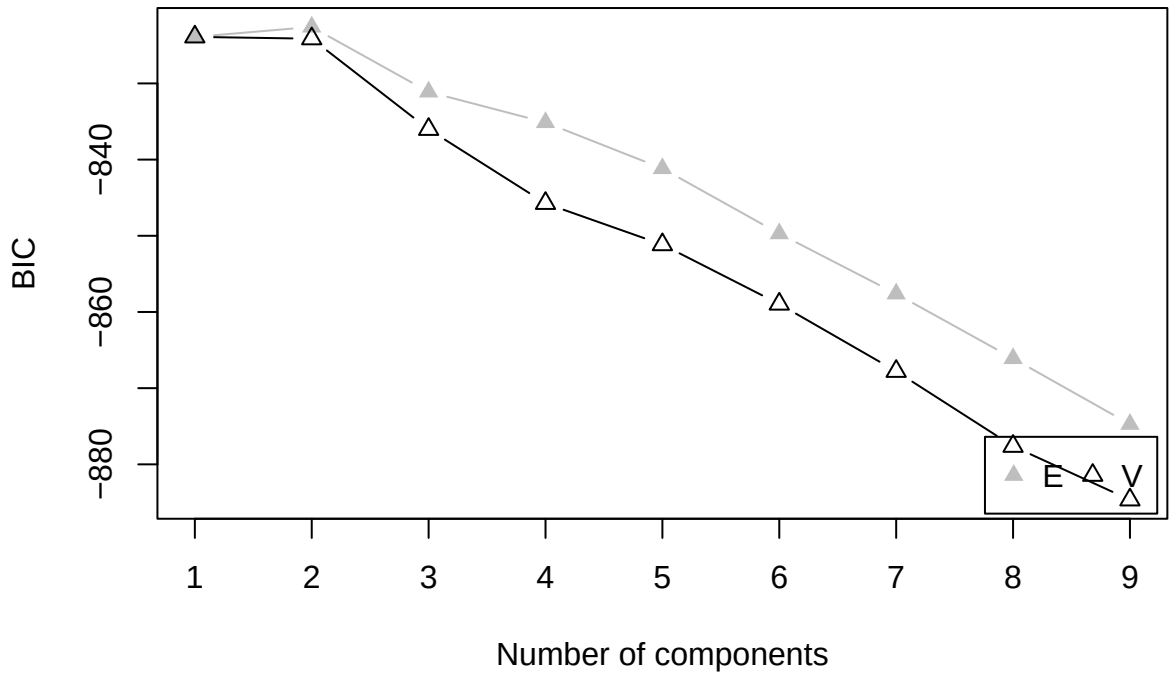
```
## Loading required package: mclust

## Package 'mclust' version 6.1.1
## Type 'citation("mclust")' for citing this R package in publications.

##
## Attaching package: 'mclust'

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

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

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

```
model30 <- mclustBIC(haemoglobin)

model30
```

```
## 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 correspond 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(model30)
```



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.

6.1.0.3 30.3 Comment on the estimated parameters referred to each subpopulation with reference to the data context.

We then estimate the selected model and comment on its parameters. The function used is `mclust::Mclust()`, which requires as input the data, the number of components, and the specification for the variance of the model.

```
model30_1 <- Mclust(haemoglobin,
                    G = 2,
                    modelNames = "E")

summary(model30_1)
```

```
## ----------------------------------------------------
## 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 output 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.

6.1.0.4 30.4 Comment on the estimated parameters of the mixture model with reference to the context of the data.

To obtain more information (on the parameters of the model), it is necessary to specify the option `parameters=TRUE`.

```
summary(model30_1, 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 mixing 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.

6.1.0.5 30.5 Provide a classification of each individual: how it is obtained? What we can learn from these results for the applicative problem at hand?

Classification of units is obtained through a **maximum a-posteriori procedure**: each subject is assigned to the cluster associated to the highest posterior probability. We print the results for the first six observations.

```
round(head(cbind(model30_1$z, Cluster = model30_1$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 observation 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.

We may also check for the cluster allocation with respect to the haemoglobin value.

```
round(head(cbind(Haemoglobin = haemoglobin, Cluster = model30_1$classification)), 4)
```

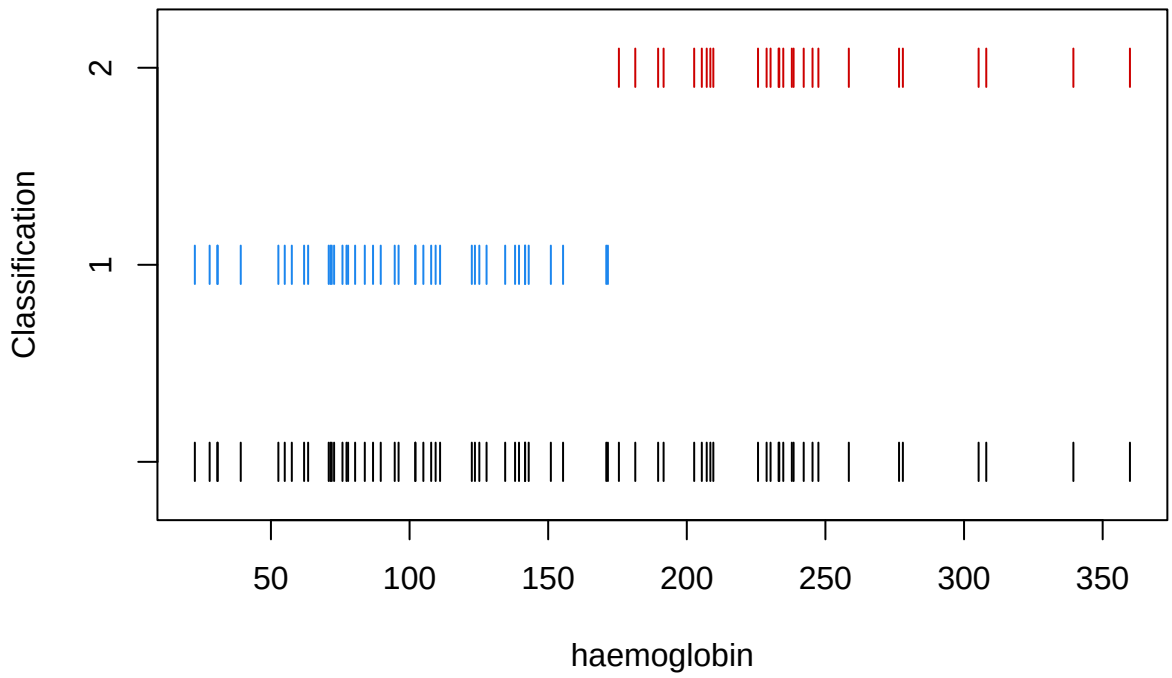
```
##      Haemoglobin Cluster
## [1,]    277.9355       2
## [2,]    207.1880       2
## [3,]     30.6607       1
## [4,]    208.4985       2
## [5,]     54.9764       1
## [6,]     39.1063       1
```

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

6.1.0.6 30.6 Draw the graph showing the clustering partition.

We plot the classification of units into the two groups by specifying the option `what = "classification"` in the plot function.

```
plot(model30_1, what = "classification")
```

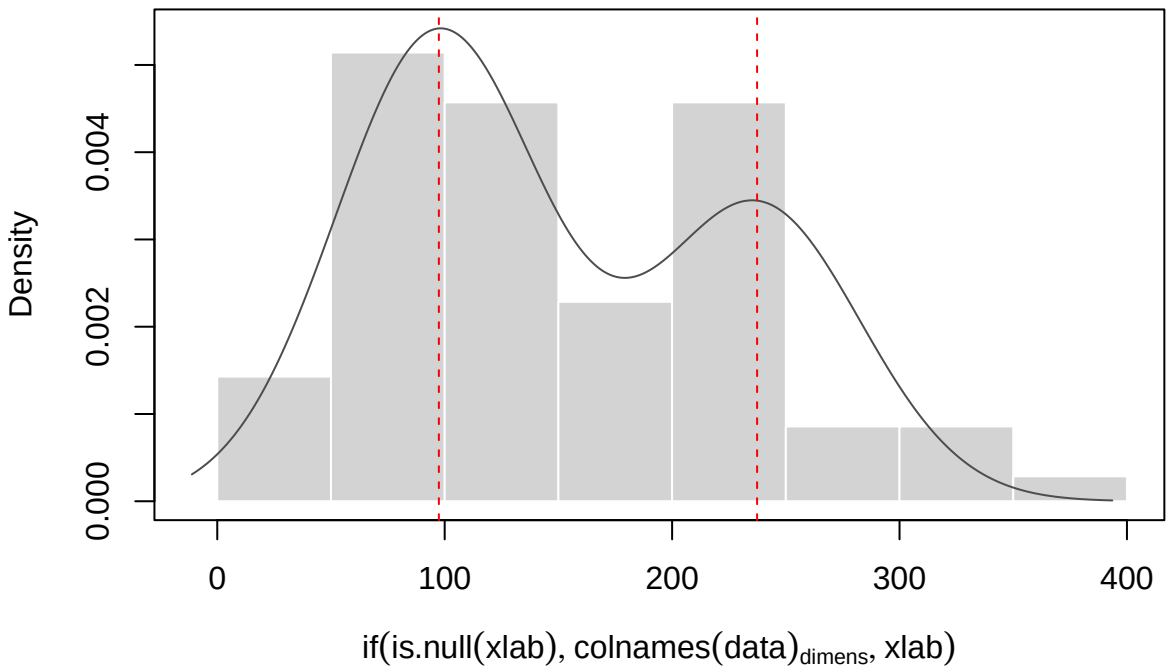


The plot represents the allocation class for each unit, based on the maximum a posteriori probability rule. We observe that the haemoglobin value that discriminates between belonging to one class or the other is around 170. The characterization of the first class as the one with lower haemoglobin values 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).

6.1.0.7 30.7 Plot the estimated mixture density along with the histogram of the observed data. Comment on the figure.

We also represent the plot that shows the mixture density. Again, it is sufficient to specify the `what = "density"` option to obtain the required plot.

```
plot(model30_1, what = 'density', data = haemoglobin,
     xlab = "Haemoglobin")
abline(v = model30_1$parameters$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).

6.2 Mixture modeling with 4 components: estimation with varying variance (model V), parameter interpretation, fish classification, density visualization, and out-of-sample prediction.

Fish length measurements (in hiches) of 256 snappers (described in Cassie, 1954) are available in the file `snapper.Rdata`. The heterogeneity present in the data may be due to the fact that fishes belong to different age groups but age is not reported in the data.

6.2.0.1 31.1 Describe the data using descriptive statistics and graphics.

We begin by exploring the dataset `snapper`, which contains the lengths (in inches) of 256 snappers. Since the dataset consists of a single quantitative variable, we rely on univariate descriptive statistics and exploratory plots to uncover potential patterns or heterogeneity—possibly linked to unobserved age groups.

```
load("snapper.Rdata")
skim_without_charts(snapper)
```

Table 25: Data summary

Name	snapper
Number of rows	256
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

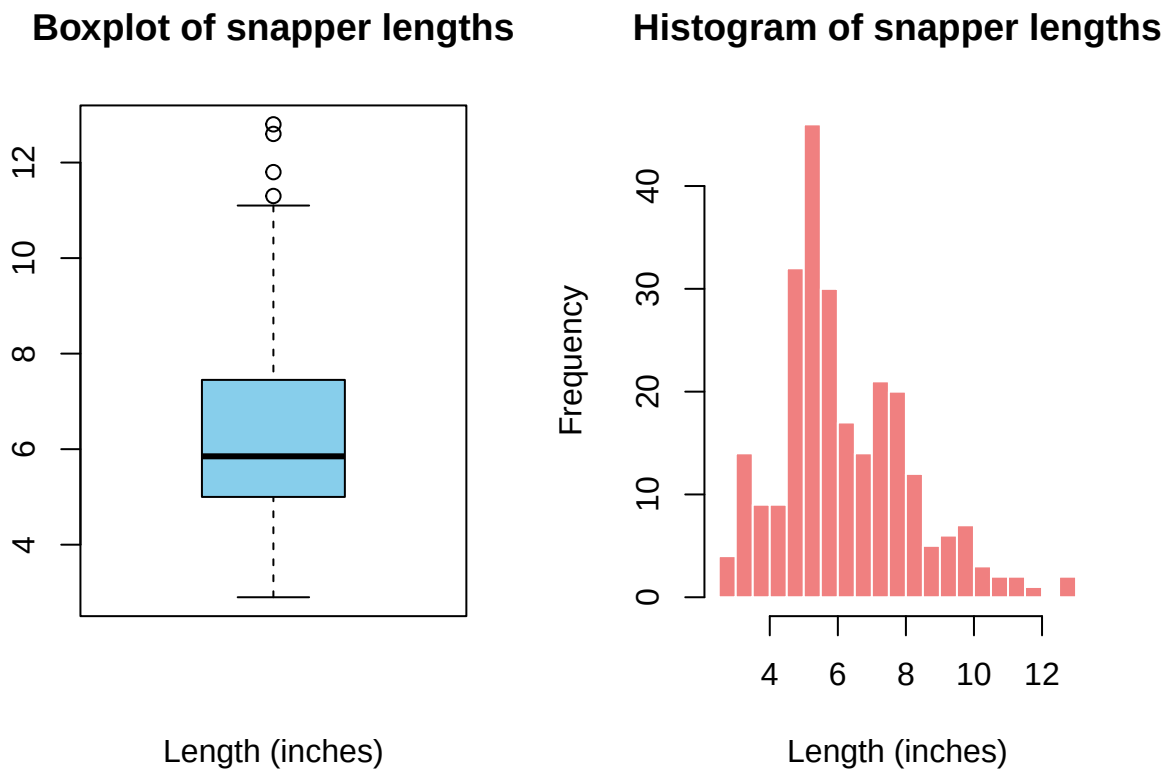
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
data	0	1	6.22	1.89	2.9	5	5.85	7.43	12.8

- The mean snapper length is approximately 6.22 inches, with a standard deviation of 1.89.
- The range spans from a minimum of 2.9 to a maximum of 12.8, indicating substantial variability across the sample.
- The interquartile range (from Q1 = 5 to Q3 = 7.425) shows that 50% of the fish fall within a 2.4-inch band, suggesting moderate concentration around the center.
- The distribution is positively skewed, as the maximum (12.8) is quite distant from the median (5.85) compared to the minimum.

To visualize the shape and spread of the distribution, we plot both a boxplot and a histogram.

```
par(mfrow = c(1,2))

boxplot(snapper,
        main = "Boxplot of snapper lengths",
        col = "skyblue",
        horizontal = FALSE,
        xlab = "Length (inches)")

hist(snapper,
     breaks = 30,
     col = "lightcoral",
     border = "white",
     main = "Histogram of snapper lengths",
     xlab = "Length (inches)",
     ylab = "Frequency")
```



The **boxplot** reveals a moderate level of variability in the data. The median length is around 5.85 inches, with an interquartile range (IQR) approximately between 5 and 7.4 inches. A few mild outliers are observed above 10 inches, suggesting that while most snappers are relatively small, a handful are substantially longer, likely corresponding to older or distinct subpopulations. The left whisker extends to around 2.9 inches, indicating the presence of a few particularly short individuals.

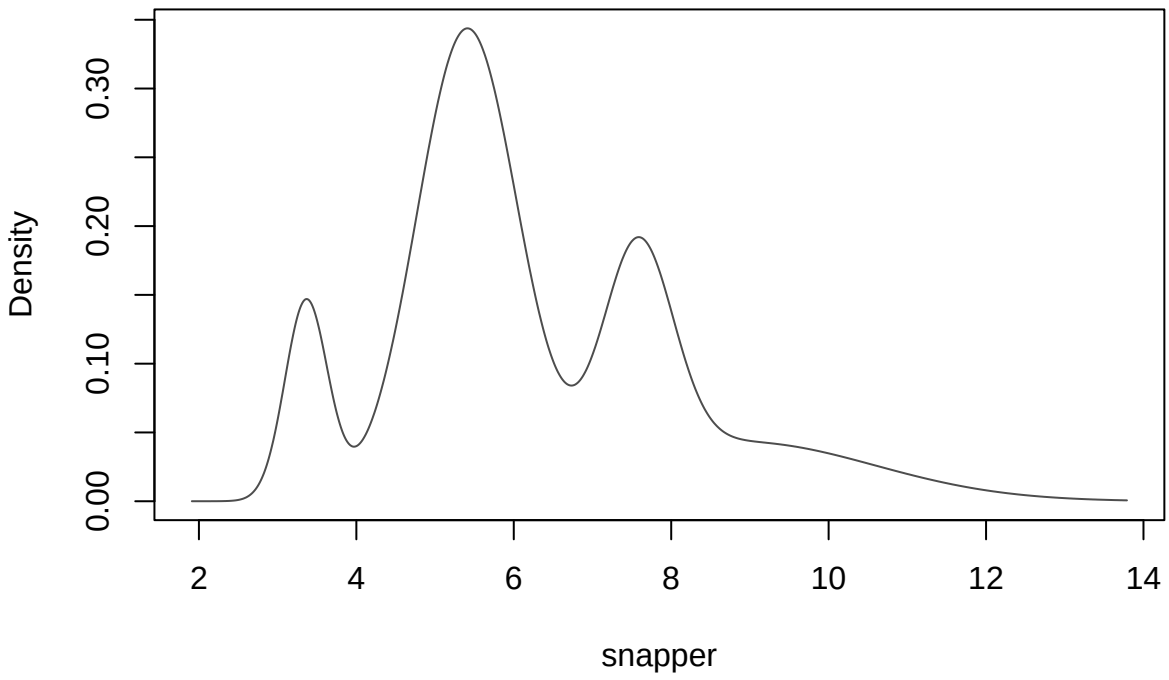
The **histogram** complements this with a more detailed view of the distribution shape. The histogram is right-skewed, with a clear concentration of fish lengths between 4 and 6 inches. This skewness may indicate the presence of latent subgroups, such as distinct age classes. The secondary mode in the 8–10 inch range supports this hypothesis, and motivates further investigation through clustering methods.

Overall, the data show heterogeneity, and the asymmetric shape and multimodality suggest the possible presence of multiple subpopulations not captured by a single normal distribution. This motivates the application of finite mixture models, which will be explored in the following steps.

6.2.0.2 31.2 Using the 'mclust' package estimate a mixture model with 4 components and varying variance among the components (model label V). Comment on the values provided by the 'summary' function.

We estimate a Gaussian finite mixture model on the snapper fish lengths using the `mclust` package. The model is constrained to use **4 components**, assuming each group has its own variance (model type `"V"`). The estimation is carried out with the `densityMclust()` function specifying both the number of components `G = 4` and the `modelNames = "V"` option.

```
model31 <- densityMclust(snapper, G = 4, modelNames = "V")
```




```
summary(model31, parameters = TRUE)
```

```
## ----------------------------------------------------
## Density estimation via Gaussian finite mixture modeling
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 4 components:
##
##  log-likelihood   n df       BIC       ICL
##       -489.2821 256 11 -1039.561 -1097.973
##
## Mixing probabilities:
##          1          2          3          4
## 0.09826312 0.54260677 0.17739121 0.18173891
##
## Means:
##        1        2        3        4
## 3.363060 5.404196 7.576137 8.902064
##
## Variances:
##          1          2          3          4
## 0.07338274 0.40836991 0.19732349 2.83984340
```

The Gaussian finite mixture model estimated via `densityMclust()` identifies **4 distinct subpopulations** within the snapper length data. The model used is a univariate mixture with **unequal variances** across the

components (`modelNames = "V"`). This allows each group to have its own level of spread, which is essential when modeling heterogeneous biological data such as fish lengths.

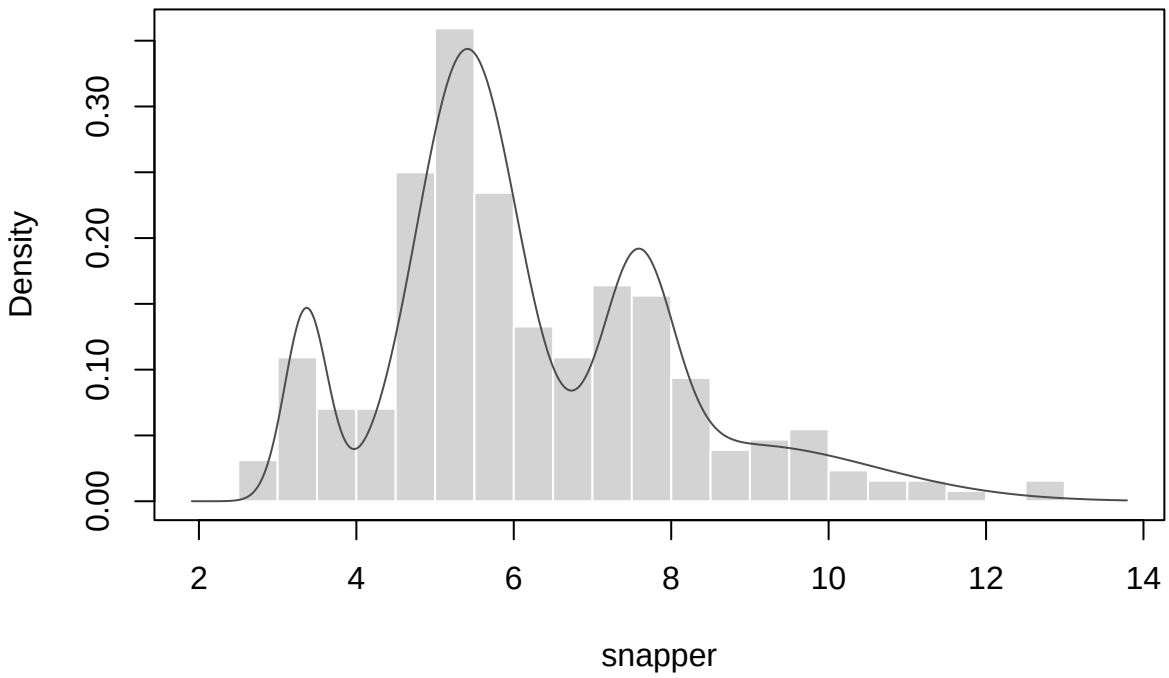
From the output:

- **Mixing probabilities** indicate the estimated proportion of the population in each group:
  - Cluster 1: ≈ 9.8
  - Cluster 2: $ 54.3% (largest group)$
  - Cluster 3: ≈ 17.7
  - Cluster 4: ≈ 18.2
- **Component means** suggest that fish lengths are centered around:
  - 3.36 inches (cluster 1)
  - 5.40 inches (cluster 2)
  - 7.58 inches (cluster 3)
  - 8.90 inches (cluster 4)
- **Variances** highlight heterogeneity among groups:
  - The fourth group (mean ≈ 8.9) has the **largest variance** (2.84), suggesting high variability in fish size within this cluster.
  - The first group is much more homogeneous, with a very low variance (0.07).

These findings are consistent with biological interpretations: the different Gaussian components likely reflect **age classes** or **growth stages** within the snapper population, although age was not explicitly recorded.

We can visualize the density and how well it fits the observed distribution.

```
plot(model31, what = "density", data = snapper, breaks = 20)
```



The **density plot** visually confirms this interpretation. The overall distribution is multimodal, with peaks aligning well with the estimated component means. This supports the idea that the total population is a mixture of distinct subgroups. The model offers a refined view of the distribution, decomposing it into biologically meaningful latent clusters.

This model provides evidence that the overall heterogeneity in snapper length can be decomposed into four distinct subpopulations, likely corresponding to **age classes** or **developmental stages**, as initially hypothesized.

6.2.0.3 31.3 Comment on the estimated parameters referred to each subpopulation with reference to the data context. The estimated parameters of the Gaussian finite mixture model fitted to the snapper dataset reveal meaningful substructure within the population, likely corresponding to distinct **age or growth classes**, despite age not being explicitly recorded.

The model identified **four subpopulations**, each described by its own mean and variance:

- **Component 1**:
  - **Mean**: 3.36 inches
  - **Variance**: 0.07
  - **Mixing proportion**: ≈ 9.8 This group contains the **smallest and most homogeneous fish**. The very low variance suggests they are of similar size and likely represent the **youngest age class** in the sample.
- **Component 2**:
  - **Mean**: 5.40 inches

  - **Variance**: 0.41
  - **Mixing proportion**: ≈ 54.3 This is the **largest group** in proportion. The moderate mean and variability suggest this is a **dominant age class**, likely representing juvenile to subadult snappers that are still in an active growth phase.

- **Component 3**:

  - **Mean**: 7.58 inches
  - **Variance**: 0.20
  - **Mixing proportion**: ≈ 17.7 This cluster likely corresponds to a more mature age group, with a relatively narrow spread in length, suggesting they are reaching growth stabilization.

- **Component 4**:

  - **Mean**: 8.90 inches
  - **Variance**: 2.84
  - **Mixing proportion**: ≈ 18.2 This final group represents the **largest and most variable fish**. The high variance may be due to inter-individual differences in growth among adults or the presence of outliers. Biologically, this could indicate older fish with more heterogeneous environmental or genetic backgrounds affecting their final size.

In summary, the model captures clear differentiation in the distribution of fish lengths, likely attributable to developmental stages. The unequal variances across groups justify the use of the `"V"` model specification, and the meaningful separation between means supports the assumption of latent heterogeneity in the sample.

6.2.0.4 31.4 Provide a classification of each fish. What we can learn from these results for the applicative problem at hand?

To classify each fish into one of the four latent subpopulations identified by the finite mixture model, we use the **posterior classification probabilities** obtained from the fitted `model31`. These posterior probabilities assign each fish to the component (cluster) for which it has the highest probability of belonging.

This classification step is performed by:

```
classification31 <- model31$classification

table(classification31)
```

```
## classification31
##   1   2   3   4
##  26 143  56  31
```

This gives the number of fish assigned to each of the four components. Each fish is assigned to the most likely subpopulation based on its length.

The classification highlights how the observed heterogeneity in fish lengths can be effectively explained by assuming that the population is composed of several distinct subgroups:

- These subgroups likely correspond to different age classes, despite age being unobserved.
- Classification enables indirect age estimation of each fish using only its length.
- This is particularly valuable in fisheries biology, where destructive age estimation (e.g., otolith sectioning) may be costly or infeasible.
- Moreover, it allows resource managers to track the composition of fish stocks, assess growth patterns, and inform sustainable harvest policies.

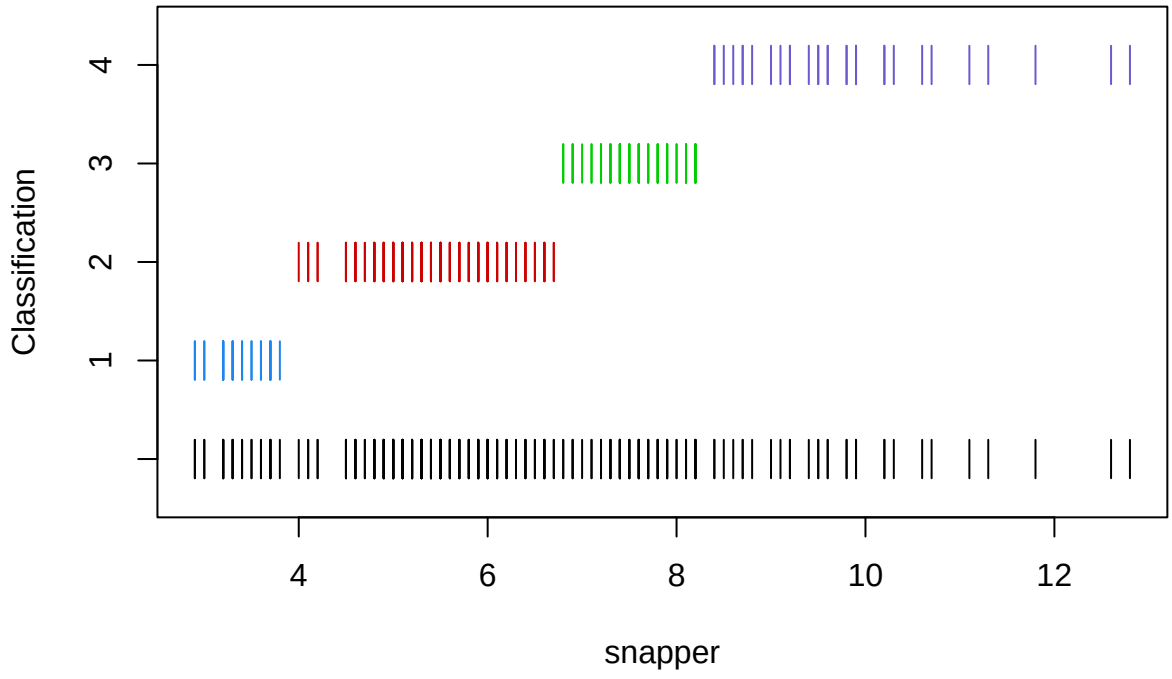
- The relative sizes of the subpopulations (as indicated by the number of fish assigned to each group) give an indication of recruitment and survival dynamics in the population.

This type of **unsupervised classification** is a powerful tool when **important latent traits (like age)** are unobserved but indirectly influence measurable quantities (like length). By fitting a **Gaussian mixture model**, we recover meaningful biological structure and provide **individual-level classifications** that can inform both ecological understanding and practical management.

6.2.0.5 30.8 Draw the graph showing the clustering partition and comment on the change value of each cluster.

To visualize the **clustering partition** and inspect how well the mixture model classifies the fish based on length, we draw the following **classification plot** using the `plot()` function from the `mclust` package:

```
model31_class <- Mclust(snapper, G = 4, modelNames = "V")
plot(model31_class, what = "classification")
```



The classification plot shows the assignment of each individual fish to one of the four Gaussian components estimated by the finite mixture model. On the x-axis we have the fish length (in inches), and on the y-axis the four identified latent classes.

Each tick mark represents a single fish, and its position on the x-axis reflects its length. The color grouping and horizontal position indicate the cluster to which the model has assigned the observation.

We observe that the model partitions the snapper population into four distinct subgroups along the continuous variable "length", each likely corresponding to a different age class or growth stage, which aligns well with the underlying biological hypothesis (i.e., the unobserved age heterogeneity of the fish population).

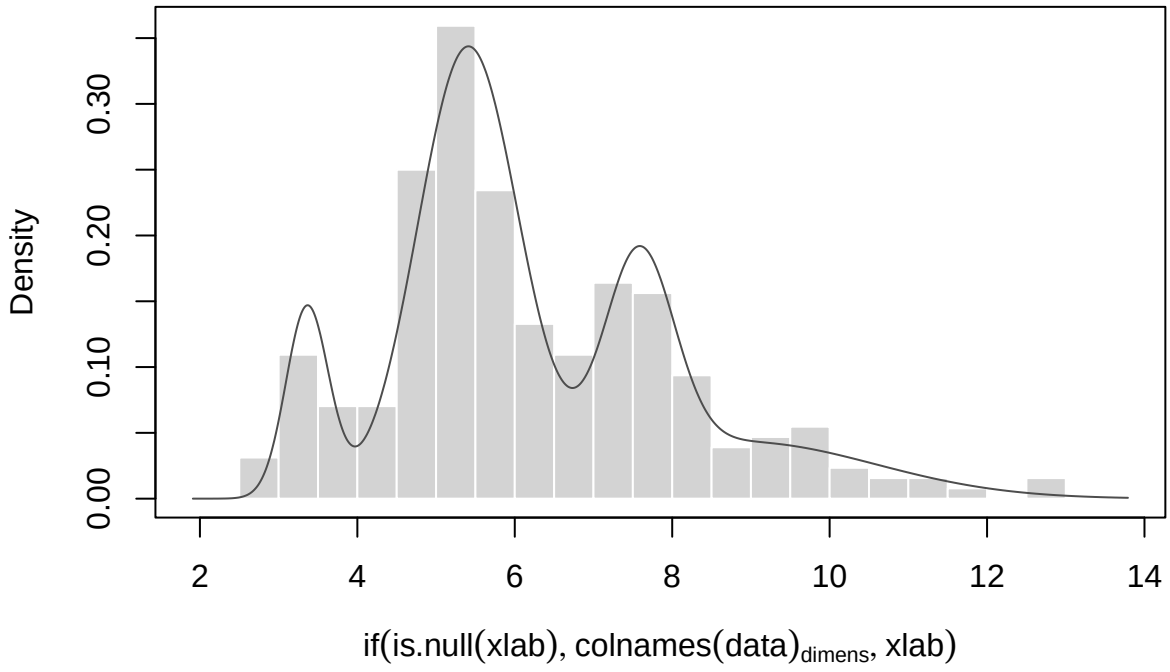
- **Cluster 1** (black): contains the majority of observations, ranging from roughly 3 to 8 inches, suggesting a dense core population of medium-size fish.
- **Cluster 2** (blue): corresponds to small fish (around 3.5–4.5 inches), likely the youngest age group.
- **Cluster 3** (green): covers fish in the intermediate range (7–9 inches), probably an older juvenile or sub-adult group.
- **Cluster 4** (purple): contains the largest fish (>9 inches), representing the oldest or fully grown group, with greater within-group variability as also indicated by the density tail in the histogram.

This classification confirms the multimodal shape observed in the earlier histogram and supports the hypothesis of unobserved heterogeneity due to age differences. The model provides not only a partition of the population but also an interpretable framework to guide further ecological or biological analysis.

6.2.0.6 31.5 Plot the estimated mixture density along with the histogram of the observed data. Comment on the figure.

To address this point, we visually compare the empirical distribution of the snapper lengths with the estimated mixture density.

```
plot(model31_class, what = "density", data = snapper, breaks = 20)
```



This command overlays the estimated **mixture density** on top of the histogram of observed data. The histogram shows the frequency distribution of fish lengths, while the superimposed density curve reflects the **model-based smooth estimate** obtained by the Gaussian finite mixture model.

The estimated density curve follows the shape of the histogram quite closely, successfully capturing the **asymmetry and multimodality** in the data. In particular:

- The **first peak** (around 3–4 inches) corresponds to the smallest fish, likely representing the youngest age group.

- A **second, higher peak** (around 5.5 inches) captures the bulk of the fish population.
- **Minor peaks** around 7.5 and 9 inches account for larger and potentially older individuals.

The density estimate smoothly connects these peaks and provides a **flexible, probabilistic description** of the data distribution that accounts for underlying heterogeneity. This supports the biological hypothesis that the observed lengths are a **mixture of subpopulations**, potentially corresponding to unmeasured age classes.

Thus, the mixture model is appropriate both for estimating the distribution and for clustering fish by size, offering insights that go beyond simple summary statistics.

6.2.0.7 31.6 Predict the classification for new fish (out of sample) having length equal to 7. Comment on the results.

To classify a new fish with length equal to 7 inches using the estimated Gaussian finite mixture model `model31_class`, we apply the `predict()` function as follows:

```
predict(model31_class, newdata = 7)
```

```
## $classification
## [1] 3
##
## $z
##                  1         2         3         4
## [1,] 9.824306e-40 0.1408037 0.6454258 0.2137705
```

The classification output for the new fish with length 7 inches confirms that it is assigned to **cluster 3**, with the following estimated posterior probabilities:

- Cluster 1: ≈ 0
- Cluster 2: 14.1%
- **Cluster 3: 64.5%** (*highest*)
- Cluster 4: 21.4%

This means the fish is most likely to belong to **component 3**, but with some uncertainty: there's a moderate chance (≈ 21) that it might belong to cluster 4 and some (14%) to cluster 2. The model shows good discriminatory ability, but this borderline case illustrates how classification uncertainty increases near overlapping regions of the components.

From a biological perspective, this length value lies **between the centers of two clusters**, likely representing overlapping growth stages. Thus, the probabilistic output allows us to **quantify uncertainty** rather than forcing a hard, binary decision—this is one of the key strengths of mixture modeling for classification.

6.3 Finite mixture clustering with spherical models (EII, VII): model selection, parameter estimation, classification plots, posterior probabilities, density visualization, and bootstrap confidence intervals.

`tyr.Rdata` contains values related to the measurements of patients affected by thyroid disease recorded about two medical records of total serum thyroxine (totser) and basal hormone (basal).

6.3.0.1 32.1. Describe the sample data.

The dataset collects data on the level of serum thyroxine and basal hormone in 122 patients affected by thyroid disease. We use the `skimr` package to describe the data and analyze its main descriptive statistics.

```
load("tyr.Rdata")
skim_without_charts(tyr)
```

Table 27: Data summary

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

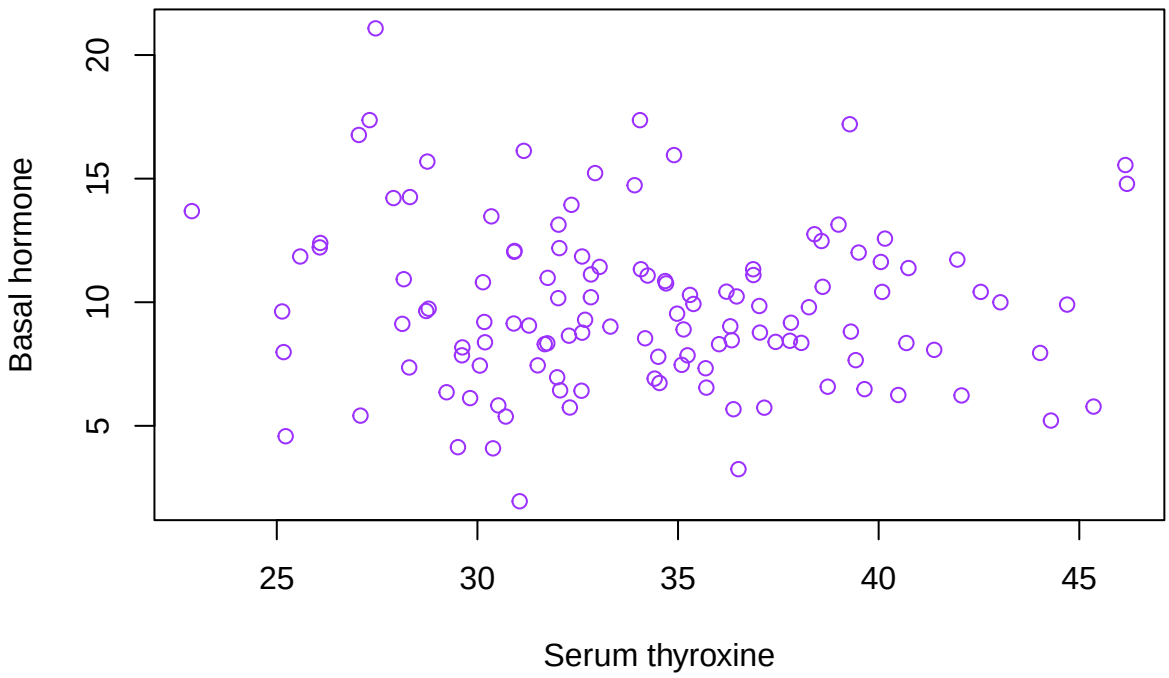
Variable type: numeric

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.

We also depict the scatter plot in order to show associations between the two variables in the sample data.

```
plot(tyr$totser, tyr$basal,
     xlab = "Serum thyroxine",
     ylab = "Basal hormone",
     col = "purple1")
```



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.

6.3.0.2 32.2 Use a finite mixture model of Gaussian distribution to find clusters in the reference population. Perform model selection to choose the best model among those with the following spherical structures "EII" and "VII". Report the results and comment especially on the selected covariance structure. The choice of the number of components and model specification is performed using the `mclustBIC()` function. In the multivariate case, among the possible model specifications, we focus on the assumption of spherical components (assuming that there is no correlation between variables). The encoding for this situation is `EII` (in the case where the variability is the same for all components) or `VII` (in the case where each component has a specific variability).

```
sel32 <- mclustBIC(tyr, modelNames = c("VII", "EII"))

sel32
```

```
## 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.

6.3.0.3 32.3 Fit the mixture model selected at the previous step and comment on the estimates according to the applied context.

We now estimate the model selected in the previous step.

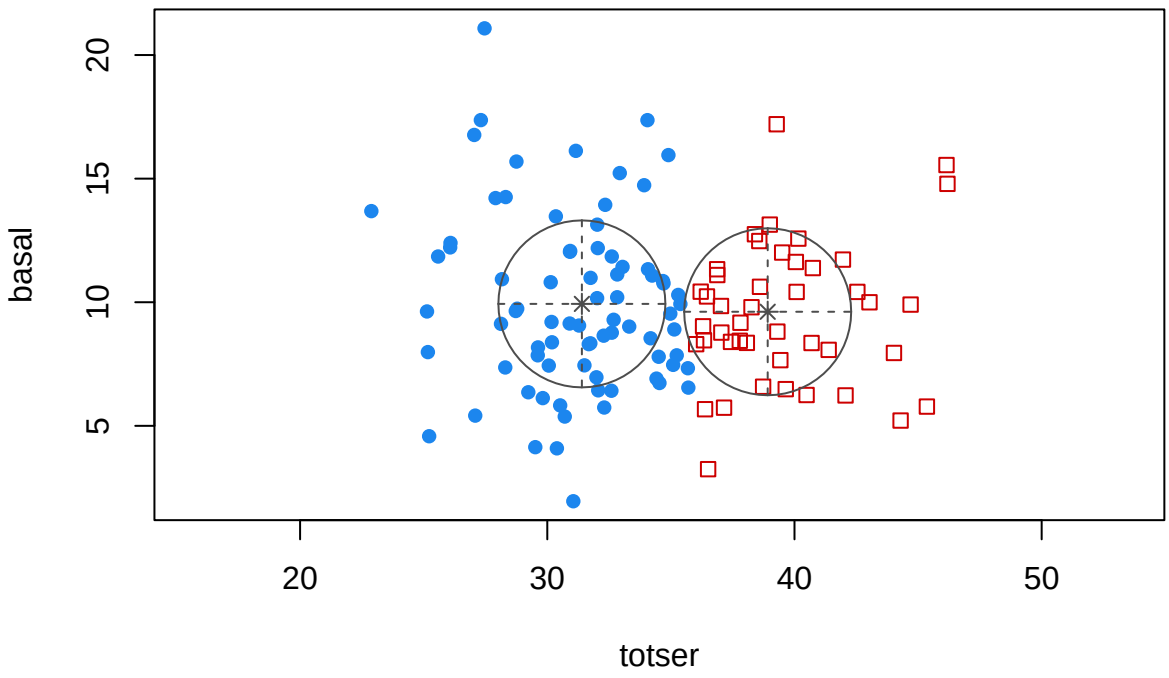
```
model32 <- Mclust(tyr, G=2, modelNames = "EII")
summary(model32, parameters = TRUE)
```

```
## ----------------------------------------------------
## 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.

6.3.0.4 32.4 Depict the graph illustrating the classification. Comment. The plot of the classification of observations is obtained by specifying the `what = "classification"` option in the plot command. Note that by using the `asp = 1` option, it is possible to fix the same scale for both the x-axis and the y-axis; this option is relevant in this case because the model being considered assumes the same variability for both components in both groups.

```
plot(model32, "classification", asp = 1)
```



- 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 located in a central position with respect to the two clusters. This behavior highlights a high uncertainty in the classification.

6.3.0.5 32.5 Show the estimated posterior probabilities for units with id 23, 45, 33, 19. Comment.

We print the estimated posterior probabilities for units with id 23, 45, 33, 19.

```
id32 <- c(23, 45, 33, 19)
prob32 <- model32$z[id32, ]
rownames(prob32) <- id32
prob32
```

```
##         [,1]        [,2]
## 23 0.0428784 0.957121599
```

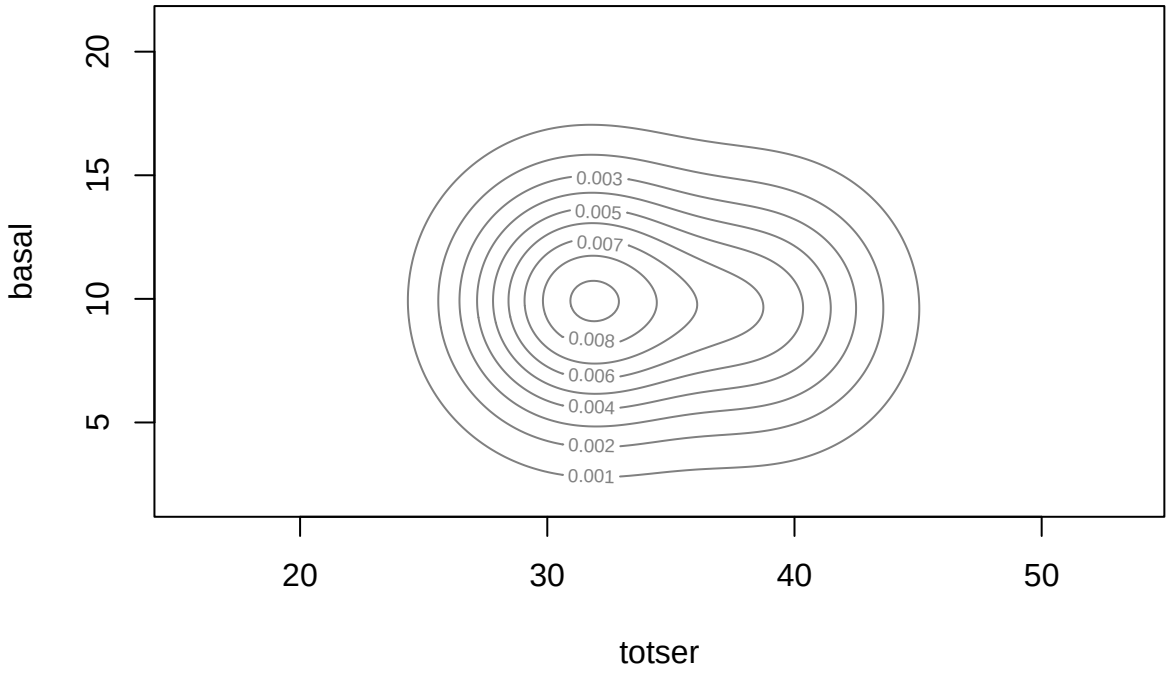
```
## 45 0.9223871 0.077612865
## 33 0.9941897 0.005810301
## 19 0.3266274 0.673372631
```

Units 33 and 45 share a similar behavior: both have a very high estimated probability to be allocated in the first group (equal to 0.99 and 0.92, respectively). On the contrary, unit with id 23 is classified in group 2 with probability equal to 0.95. Finally, the last unit (id 19) shows a more uncertain behavior: the probability to be allocated in the first and second groups is equal to 0.33 and 0.67, respectively.

6.3.0.6 32.6 Depict the graph illustrating the estimated density and comment on it. Show also the image plot and the perspective plot, as well as the highest density regions plot. Comment.

Using the `what = "density"` option, it is possible to obtain the plot of the density of the mixture.

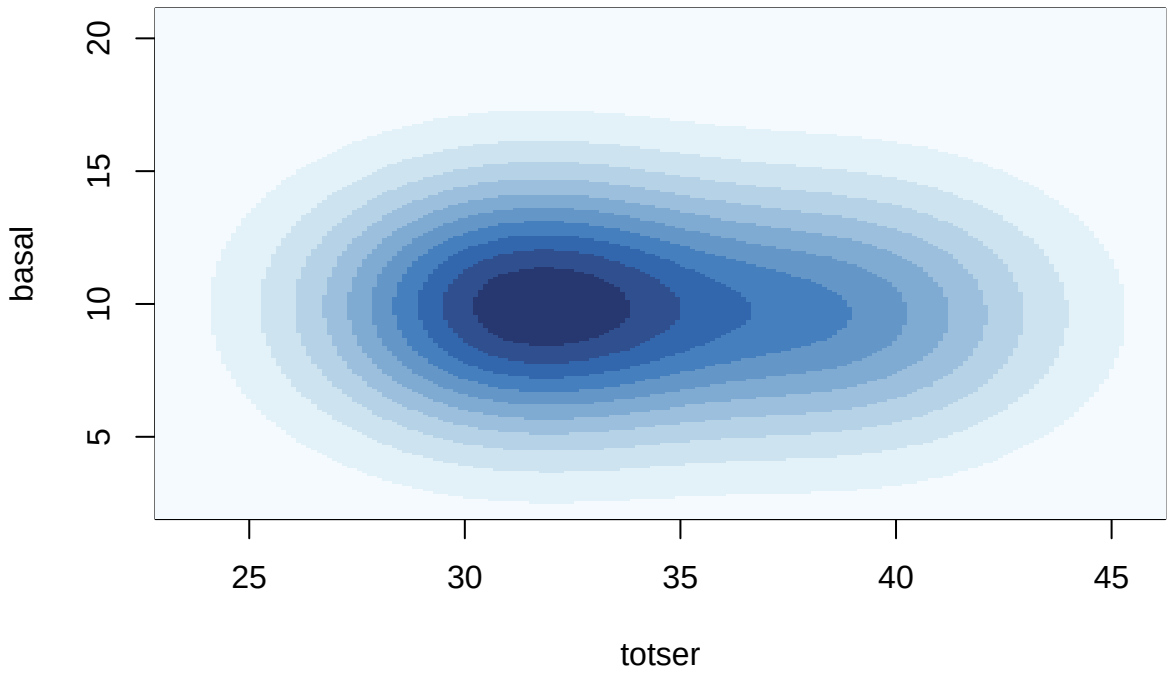
```
plot(model32, "density", asp = 1)
```



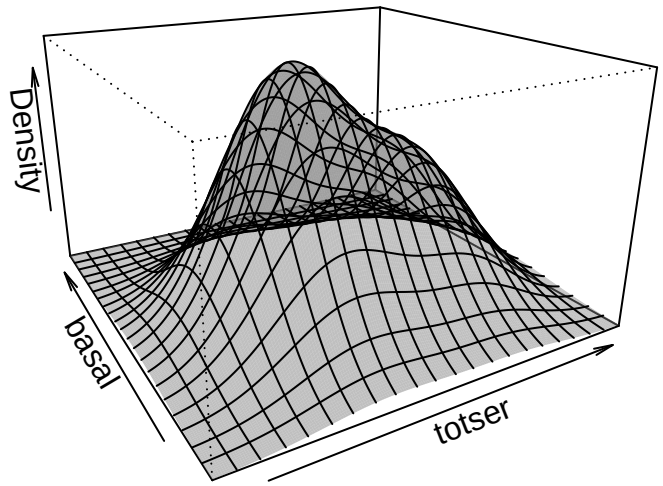
The shape of the density is represented by means of contour lines (projections of the intersections of the 3D surface with planes parallel to the xy-plane and at different heights). It can be observed that the figure is slightly elongated in the horizontal direction, with a more pronounced extension 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.

The following three plots (image plot, perspective plot, and density regions plot, respectively) provide the same information.

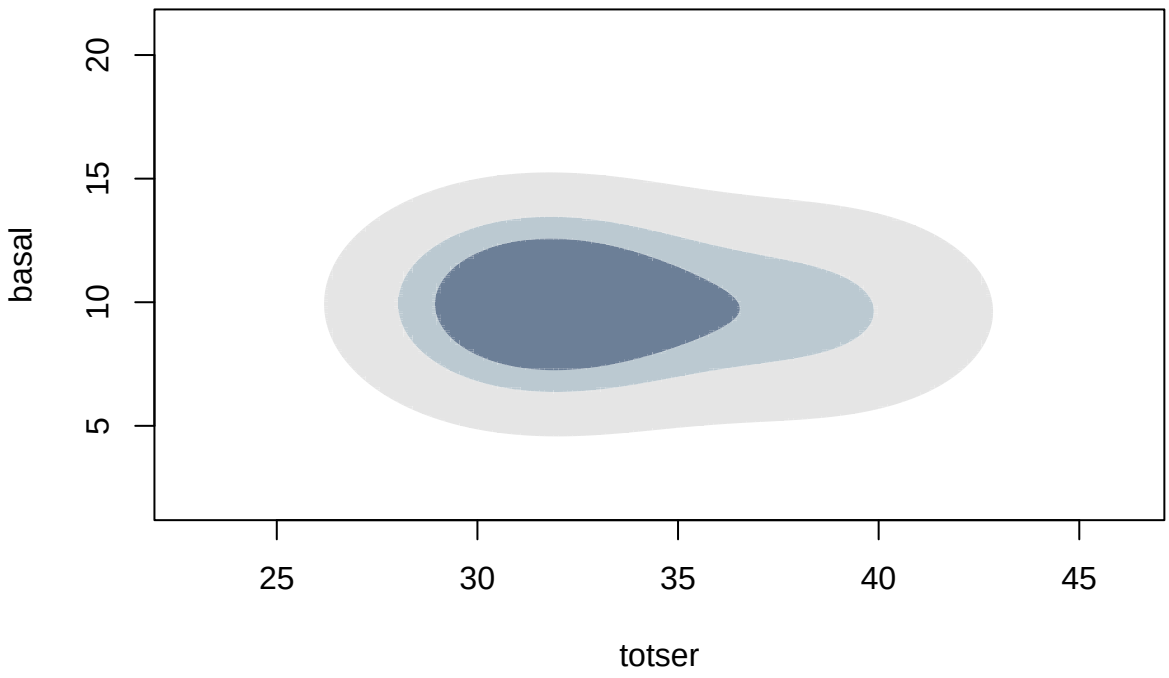
```
plot(model32, "density", type = "image")
```




```
plot(model32, "density", type = "persp")
```




```
plot(model32, "density", type = "hdr")
```



6.3.0.7 32.7 Calculate the bootstrap standard errors and report and comment the estimated bootstrap confidence intervals obtained with the percentile method for the means of each cluster. Comment on the results reporting also the plot of the bootstrap distribution. To obtain bootstrap confidence intervals for the model parameters, the `mclust::MclustBootstrap` function is used. The only argument required is the estimated model; then, the summary function requires the option `what = "ci"` to obtain the confidence intervals (there might also be a similar option `what = "se"` to obtain the standard errors).

```
set.seed(17)
mc_boot32 <- MclustBootstrap(model32)
summary(mc_boot32, 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%  0.4286354 0.1940786
## 97.5% 0.8059214 0.5713646
```

```
##
## Means:
## [,,1]
##         totser     basal
## 2.5%  29.89607  9.108385
## 97.5% 32.68035 11.023581
## [,,2]
##         totser     basal
## 2.5%  37.07816  8.759168
## 97.5% 41.36121 10.721163
##
## Variances:
## [,,1]
##          totser     basal
## 2.5%   8.783639  8.783639
## 97.5% 13.600364 13.600364
## [,,2]
##          totser     basal
## 2.5%   8.783639  8.783639
## 97.5% 13.600364 13.600364
```

- We observe that the confidence interval (at 95%) for the weight of the first component is quite wide, with 0.43 and 0.81 as lower and upper bounds, respectively. It is worth nothing that this interval also includes values below 0.5, which would result in a mixture where the heaviest component would no longer be the first but the second. Comments regarding the interval for the weight of the second component are symmetric.
- As for the means, the confidence intervals are generally quite narrow; for instance, the interval for the mean of level of serum thyroxine among subjects in the first group ranges from 29.9 to 32.7: in other words, in 95% of 999 bootstrap samples, the calculated value for this parameter falls within this interval.

6.4 Classification with finite mixture models: data exploration, train-test split, discriminant analysis with Gaussian mixtures, performance metrics (accuracy, sensitivity, specificity), cross-validation, and ROC analysis.

Data `dtbreast1.Rdata` report data related to measurements on breast mass: radius mean, symmetry mean and texture mean and a diagnosis of malignant (M) and benign (B) cancer.

```
load("dtbreast1.Rdata")
```

6.4.0.1 33.1 Describe the sample data evaluating the observed association of each measurement with the diagnosis.

We begin the analysis by describing the distribution of the three quantitative features collected in the dataset: `Radius_mean`, `Symmetry_mean`, and `Texture_mean`. These variables refer respectively to the average radius of the cell nuclei, the mean asymmetry in the tumor shape, and the average variation in the texture of the mass—features that are often relevant in distinguishing between benign and malignant breast masses.

```
skim_without_charts(dtbreast1)
```

Table 29: Data summary

Name	dtbreast1
Number of rows	569
Number of columns	4
Column type frequency:	
factor	1
numeric	3
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
wdbc.Diagnosis	0	1	FALSE	2	B: 357, M: 212

Variable type: numeric

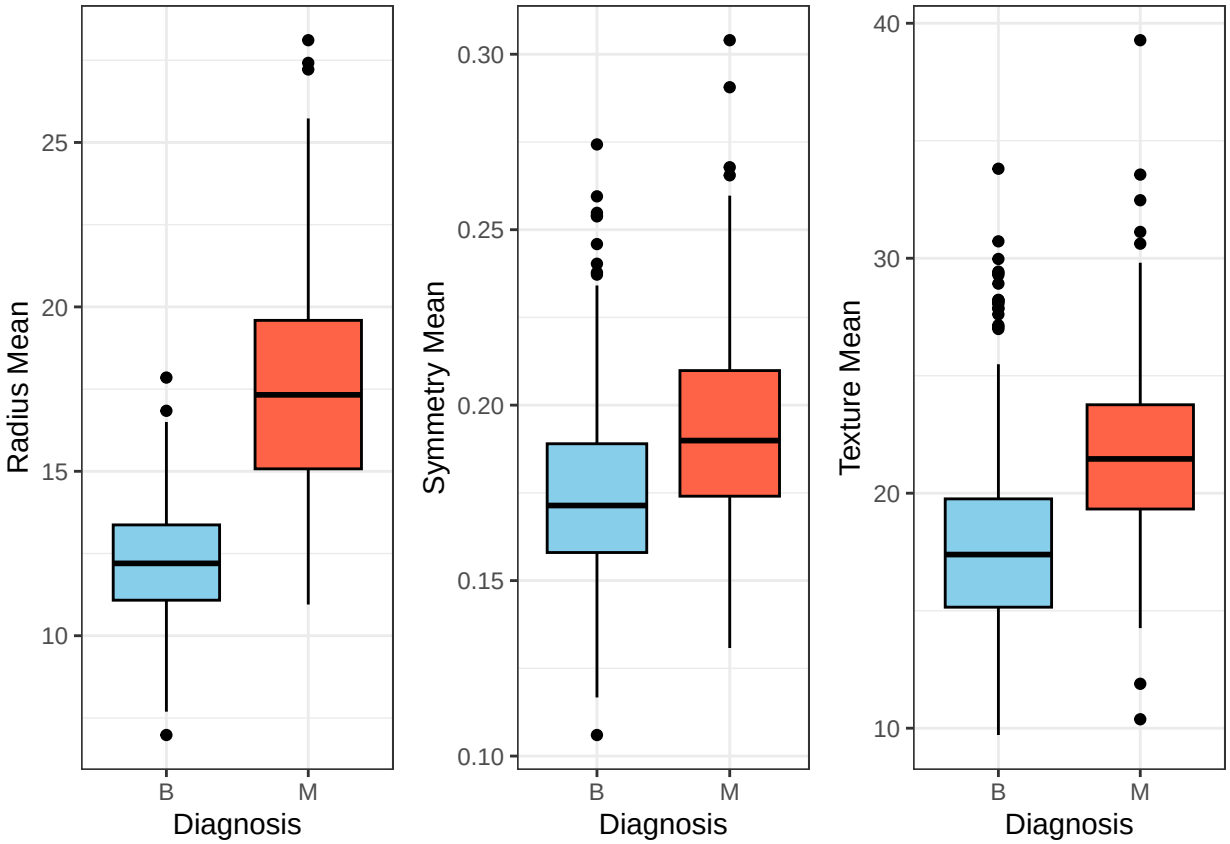
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Radius_mean	0	1	14.13	3.52	6.98	11.70	13.37	15.78	28.11
Symmetry_mean	0	1	0.18	0.03	0.11	0.16	0.18	0.20	0.30
Texture_mean	0	1	19.29	4.30	9.71	16.17	18.84	21.80	39.28

The summary statistics, computed via the `skimr` package, show that `Radius_mean` has a mean value of approximately 14.13 and spans a wide range, from 6.98 to 28.11, indicating a high variability in the dataset. The standard deviation is about 3.52, and 75% of the observations lie below 15.78, suggesting a right-skewed distribution. Similarly, `Texture_mean` varies substantially across patients, with values ranging from 9.71 to 39.28 and a standard deviation of 4.30. In contrast, `Symmetry_mean` has a much smaller scale and lower variability: it ranges from 0.106 to 0.304, with a mean of 0.181 and a standard deviation of 0.027. This limited dispersion may imply that symmetry is a more tightly controlled feature, yet even small changes could be significant in a diagnostic context.

To better understand the association between these predictors and the diagnosis (benign vs malignant), we now proceed with visual exploration via boxplots stratified by diagnostic class. This will allow us to evaluate whether the distribution of each measurement differs systematically between benign and malignant tumors and guide the next steps of our modeling strategy.

```
p1_33 <- ggplot(dtbreast1, aes(x = wdbc.Diagnosis, y = Radius_mean, fill = wdbc.Diagnosis)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("B" = "skyblue", "M" = "tomato")) +
  xlab("Diagnosis") + ylab("Radius Mean") +
  theme_bw() + theme(legend.position = "none")

p2_33 <- ggplot(dtbreast1, aes(x = wdbc.Diagnosis, y = Symmetry_mean, fill = wdbc.Diagnosis)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("B" = "skyblue", "M" = "tomato")) +
  xlab("Diagnosis") + ylab("Symmetry Mean") +
  theme_bw() + theme(legend.position = "none")
```

```r
p3_33 <- ggplot(dtbreast1, aes(x = wdbc.Diagnosis, y = Texture_mean, fill = wdbc.Diagnosis)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("B" = "skyblue", "M" = "tomato")) +
  xlab("Diagnosis") + ylab("Texture Mean") +
  theme_bw() + theme(legend.position = "none")

grid.arrange(p1_33, p2_33, p3_33, ncol = 3)
```



Starting with `Radius_mean`, the separation between classes is particularly marked. The median value for malignant tumors is substantially higher than that for benign ones, and the entire interquartile range for the malignant class lies well above that of the benign class. This suggests that larger average radii are strongly associated with malignancy. Moreover, the distribution for malignant cases shows both a higher median and greater variability, with numerous high outliers. This feature appears to be highly discriminative.

Moving to `Symmetry_mean`, the difference between classes is more modest but still evident. Malignant tumors tend to have slightly higher median symmetry, and their distribution shows greater spread and more extreme outliers. While the overlap between the two classes is more pronounced here, the trend is still consistent: malignancy is associated with greater asymmetry.

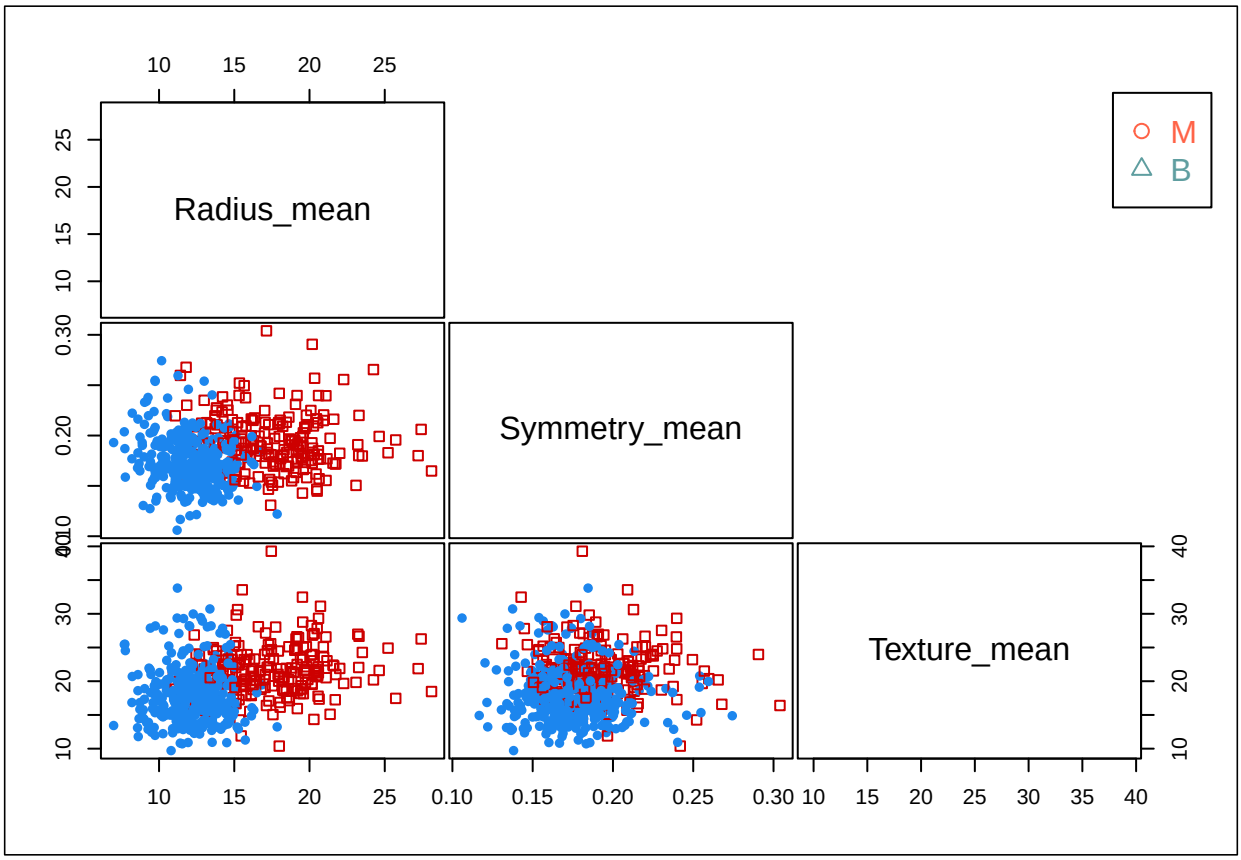
Finally, `Texture_mean` also shows a visible class difference. Malignant tumors present a slightly higher median texture and a wider spread, with a large number of extreme values. While there is substantial overlap in the central part of the distributions, the upper tail is considerably more pronounced for malignant cases, suggesting that high texture values are more typical of malignancies.

In summary, all three variables exhibit distributional shifts across diagnostic categories, with `Radius_mean` standing out as the most effective individual discriminator. These findings reinforce the relevance of these morphological indicators in supporting clinical assessments, and motivate further analysis using multivariate modeling techniques.

To complement the descriptive summaries and boxplots, we investigate the joint behavior of the three variables across diagnoses by constructing a scatterplot matrix and analyzing both raw and partial correlation coefficients. These tools allow us to visually and quantitatively assess associations and potential multicollinearity patterns among predictors.

We start by plotting the scatterplot matrix for `Radius_mean`, `Symmetry_mean`, and `Texture_mean`, with colors indicating the diagnosis (`B` vs `M`). The function `clPairs()` can be used to include group-specific coloring and shaping of the points.

```
clPairs(dtbreast1[, 1:3], dtbreast1$wdbc.Diagnosis, upper.panel = NULL)
clPairsLegend(0.9, 0.9,
              class = unique(dtbreast1$wdbc.Diagnosis),
              col = c("tomato", "cadetblue"),
              pch = c(1, 2))
```

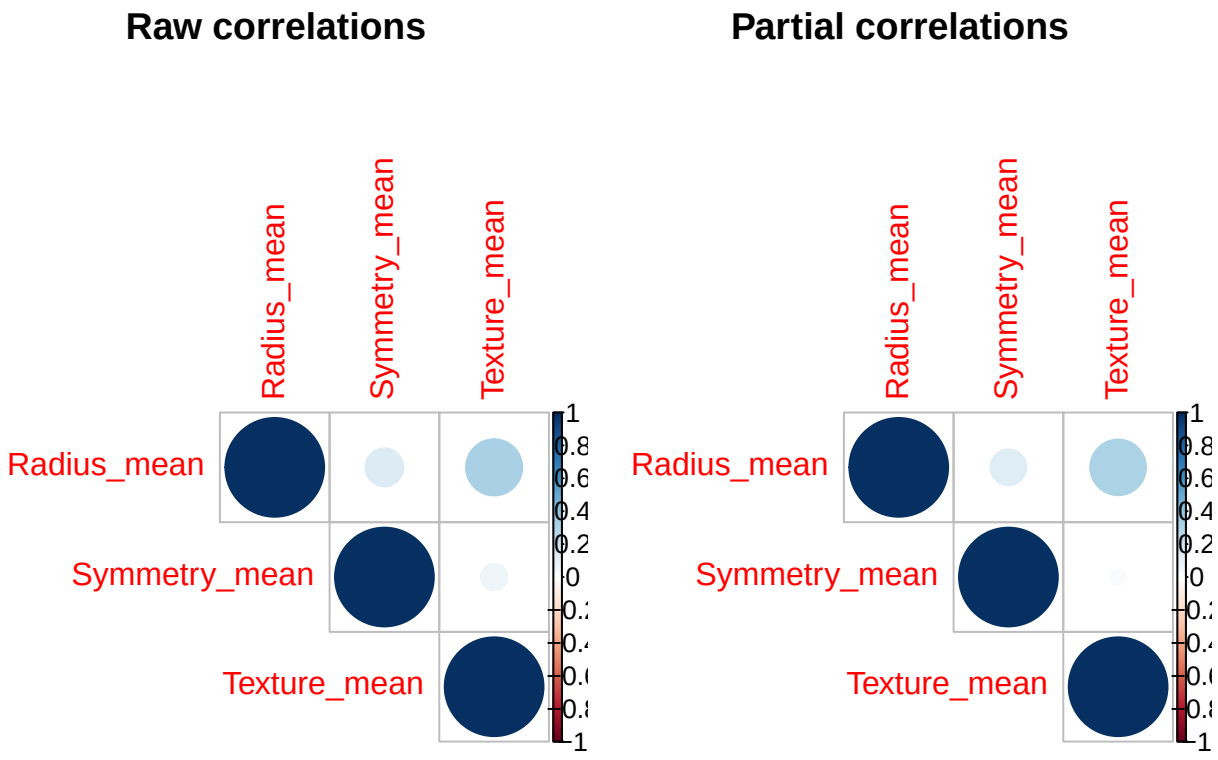


The scatterplot matrix reveals a clear pattern of separation between benign and malignant tumors: most malignant cases (red squares) lie on the upper end of all three dimensions, while benign ones (blue circles) are concentrated in the lower range. However, there is still a notable degree of overlap across classes.

`Radius_mean` appears to be the most discriminative feature, showing a relatively well-separated cloud of red and blue points when paired with both `Symmetry_mean` and `Texture_mean`. In contrast, the association between `Symmetry_mean` and `Texture_mean` is weaker in terms of class discrimination, though some trend is still visible.

```
raw_corr33 <- round(cor(dtbreast1[, 1:3]), 3)
partial_corr33 <- round(parcor(cov(dtbreast1[, 1:3])), 3)
```

```r
par(mfrow = c(1, 2))
corrplot(raw_corr33, type = "upper", title = "Raw correlations", mar = c(0, 0, 2, 0))
corrplot(partial_corr33, type = "upper", title = "Partial correlations", mar = c(0, 0, 2, 0))
```



From the **raw correlations**, we find:

- A high correlation (≈ 0.7) between `Radius_mean` and `Texture_mean`, consistent with the scatterplot.
- A moderate correlation between `Radius_mean` and `Symmetry_mean` (≈ 0.6).
- A weaker correlation (≈ 0.4) between `Symmetry_mean` and `Texture_mean`.

The **partial correlations** tell a different story:

- Once controlling for `Symmetry_mean`, the association between `Radius_mean` and `Texture_mean` remains strong, suggesting a genuine link between these two variables.
- The correlation between `Radius_mean` and `Symmetry_mean` weakens considerably when adjusting for `Texture_mean`, hinting at a mediating effect.
- The partial correlation between `Symmetry_mean` and `Texture_mean` is near zero, indicating that their marginal association may be spurious and explained by their common dependence on `Radius_mean`.

These results suggest that, although all three variables are individually informative, `Radius_mean` and `Texture_mean` carry the bulk of the explanatory signal. This will be especially important in multivariate classification, where collinearity may influence model stability and interpretability.

6.4.0.2 33.2 In order to build a prediction model (learner) to predict a binary outcome, divide data into a reasonable training and test set to evaluate the learner. Choose reasonable sets.

To build a predictive model for binary classification (malignant vs benign), a standard and reasonable approach is to split the dataset into a training set (used to fit the model) and a test set (used to evaluate predictive performance on unseen data). A typical partition is:

- **70% for training**
- **30% for testing**

This ensures the model has sufficient data to learn from, while also preserving a meaningful test set to assess generalization.

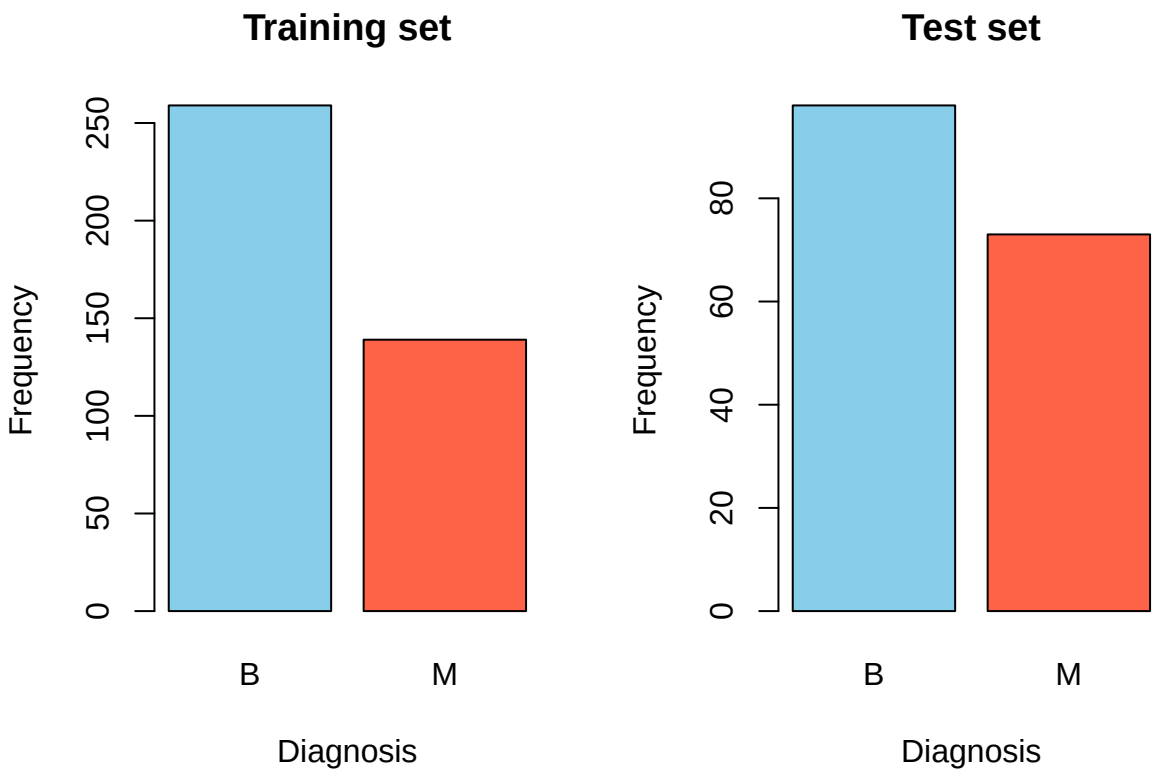
```
set.seed(123)
n33 <- nrow(dtbreast1)
train_index33 <- sample(1:n33, size = 0.7 * n33, replace = FALSE)

X_train33 <- dtbreast1[train_index33, ]
X_test33 <- dtbreast1[-train_index33, ]

# check the class proportions
table(X_train33$wdbc.Diagnosis)
```

```
##
##   B   M
## 259 139
```

```
table(X_test33$wdbc.Diagnosis)
```

```
##
##  B  M
## 98 73
```

```
par(mfrow = c(1,2))

barplot(
  table(X_train33$wdbc.Diagnosis),
  main = "Training set",
  col = c("skyblue", "tomato"),
  xlab = "Diagnosis",
  ylab = "Frequency"
)

barplot(
  table(X_test33$wdbc.Diagnosis),
  main = "Test set",
  col = c("skyblue", "tomato"),
  xlab = "Diagnosis",
  ylab = "Frequency"
)
```



This partition should retain a balanced proportion of benign and malignant diagnoses across the two sets. If large imbalance exists, stratified sampling could be considered. This split will now be used to fit the learner on `X_train33` and evaluate performance on `X_test33`.

6.4.0.3 33.3 Estimate a Gaussian finite mixture model considering discriminant analysis. Report and comment on the results, which type of variance-covariance structure is considered in the estimated model?

We now estimate a Gaussian finite mixture model for discriminant analysis using the training set. We specify the model type as `"EDDA"` (Eigenvalue Decomposition Discriminant Analysis), which assumes **one Gaussian component per class** and allows for various variance-covariance structures. The estimation is performed using the `MclustDA()` function.

```
# RELOAD the full original dataset (replace this with your original object)
load("dtbreast1.Rdata")
dtbreast1$wdbc.Diagnosis <- as.factor(dtbreast1$wdbc.Diagnosis)

# Split again: 70% train, 30% test
set.seed(33)
n <- nrow(dtbreast1)
ind_train <- sample(1:n, size = 0.7 * n, replace = FALSE)
X_train33 <- dtbreast1[ind_train, ]
X_test33 <- dtbreast1[-ind_train, ]

# Check column names
str(X_train33)
```

```
## 'data.frame':    398 obs. of  4 variables:
```

```
##  $ Radius_mean   : num  9.33 15.61 13 10.51 21.56 ...
##  $ Symmetry_mean : num  0.169 0.155 0.235 0.192 0.173 ...
##  $ Texture_mean  : num  21.9 19.4 21.8 20.2 22.4 ...
##  $ wdbc.Diagnosis: Factor w/ 2 levels "B","M": 1 2 2 1 2 1 1 2 2 2 ...
```

```r
# Now select and clean only the numeric features (not the factor)
numeric_cols <- c("Radius_mean", "Symmetry_mean", "Texture_mean")
clean_train33 <- X_train33[complete.cases(X_train33[, numeric_cols]) &
                           sapply(X_train33[, numeric_cols], is.numeric) %>% all(), ]
```

```r
table(clean_train33$wdbc.Diagnosis)
```

```
##
##   B   M
## 250 148
```

```r
library(mclust)
model33 <- MclustDA(data = clean_train33[, c("Radius_mean", "Symmetry_mean", "Texture_mean")],
                    class = clean_train33$wdbc.Diagnosis,
                    modelType = "EDDA")

summary(model33)
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -1266.945 398 13 -2611.714
##
## Classes   n     % Model G
##       B 250 62.81   VVI 1
##       M 148 37.19   VVI 1
##
## Training confusion matrix:
##      Predicted
## Class   B   M
##     B 240  10
##     M  28 120
## Classification error = 0.0955
## Brier score          = 0.0666
```

The estimated model applies Gaussian finite mixture modeling for supervised classification using the EDDA framework, which assumes one Gaussian distribution per class with class-specific variance-covariance structures. In this case, the optimal structure selected by the model is **VVI**, meaning the covariance matrices are diagonal (variables are uncorrelated) but allow for different variances across variables and classes.

The model has been trained on a dataset of 398 observations, with 250 benign (B) and 148 malignant (M) diagnoses. The confusion matrix shows a strong classification performance: only 10 benign cases were misclassified as malignant and 28 malignant cases were misclassified as benign, leading to a **misclassification**

rate of 9.55%. This result is very good for a clinical context, where identifying malignant cases is crucial. Furthermore, the **Brier score is 0.0666**, indicating high probability calibration between predicted class probabilities and actual class labels. The **BIC value of -2611.714** confirms the model's adequacy when considering model complexity and data likelihood.

Overall, the selected model structure VVI balances flexibility and parsimony, and the estimated learner appears well-suited for classifying tumors based on the three input features.

6.4.0.4 33.4 How it is evaluated the accuracy of the provided classification? Report the definitions of sensitivity and specificity and their values on the training set.

The accuracy of the classification provided by the model is evaluated by comparing the **predicted labels** with the **true class labels** in the training set.

From the confusion matrix computed in the previous point, we compute the following metrics:

- **Accuracy**: the proportion of correctly classified observations

$$\text{Accuracy} = \frac{240 + 120}{240 + 10 + 28 + 120} = \frac{360}{398} \approx 0.9045$$

- **Sensitivity (Recall or True Positive Rate)**: the proportion of malignant cases correctly identified

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{120}{120 + 28} \approx 0.811$$

- **Specificity (True Negative Rate)**: the proportion of benign cases correctly identified

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{240}{240 + 10} \approx 0.96$$

These metrics indicate that the model performs particularly well in identifying benign tumors (high specificity), while it still maintains a solid ability to detect malignant tumors (sensitivity around 81%). This balance is particularly important in medical diagnostics, where both false negatives and false positives have serious implications.

6.4.0.5 33.5 Report and comment on the confusion and evaluation metrics for a new test set.

We now evaluate the generalization ability of the discriminant model by testing it on a previously unseen portion of the data. This out-of-sample assessment is critical to understand how the classifier performs in a realistic predictive setting, beyond the training environment.

The `summary()` function from the `mclust` package is applied with the arguments `newdata` and `newclass` to obtain the confusion matrix, the misclassification rate, and the Brier score on the test set.

```
# evaluate model performance on test set
summary(model33,
        newdata = X_test33[, c("Radius_mean", "Symmetry_mean", "Texture_mean")],
        newclass = X_test33$wdbc.Diagnosis)
```

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
##
## EDDA model summary:
##
```

```
##  log-likelihood   n df       BIC
##       -1266.945 398 13 -2611.714
##
## Classes   n     % Model G
##       B 250 62.81   VVI 1
##       M 148 37.19   VVI 1
##
## Training confusion matrix:
##      Predicted
## Class   B   M
##     B 240  10
##     M  28 120
## Classification error = 0.0955
## Brier score          = 0.0666
##
## Test confusion matrix:
##      Predicted
## Class   B   M
##     B 105   2
##     M  12  52
## Classification error = 0.0819
## Brier score          = 0.0623
```

On the test set, the estimated model achieves a **classification error of 8.19%**, slightly lower than the training set error (9.55%), indicating excellent generalization. This suggests that the model is not overfitting and performs reliably when applied to new data.

The **Brier score** on the test set is 0.0623, again very close to the value obtained in the training phase (0.0666). The Brier score assesses the accuracy of probabilistic predictions, and the low value confirms that the predicted class probabilities are well-calibrated.

The confusion matrix can be interpreted as follows:

- **True Negatives (B correctly predicted as B):** 105
- **False Positives (B misclassified as M):** 2
- **False Negatives (M misclassified as B):** 12
- **True Positives (M correctly predicted as M):** 52

We can compute sensitivity and specificity based on the matrix:

- **Sensitivity (True Positive Rate):**

$$\text{Sensitivity} = \frac{52}{52 + 12} = \frac{52}{64} \approx 0.8125$$

- **Specificity (True Negative Rate):**

$$\text{Specificity} = \frac{105}{105 + 2} = \frac{105}{107} \approx 0.9813$$

These results indicate that the model maintains a **very high specificity**, minimizing false positives (benign tumors wrongly classified as malignant), and a **solid sensitivity**, correctly identifying more than 81% of malignant tumors. This balance is particularly valuable in medical contexts, where both missed malignancies and false alarms carry clinical consequences.

In conclusion, the test set evaluation confirms that the estimated EDDA model with VVI structure provides strong and robust predictive performance, with reliable class probability estimates and effective discrimination between benign and malignant cases.

6.4.0.6 33.6 Use the resampling method named V fold cross-validation to evaluate the learner. Which is the estimated standard error of the classification error? And of the Brier score?

To evaluate the **generalization performance** of the learner more robustly, we now apply **V-fold cross-validation** with $V = 10$. This method splits the data into 10 approximately equal folds, trains the model on 9 of them, and evaluates on the remaining one, rotating the hold-out fold each time. The procedure is more stable than a single train-test split and allows us to estimate not only average performance metrics (classification error and Brier score), but also their standard errors.

We employ the `cvMclustDA()` function from the `mclust` package, which performs cross-validation for discriminant analysis models. The argument `modelType = "EDDA"` enforces the same structure used in the previous parts of the exercise.

```
mod33.cv <- MclustDA(data = dtbreast1[, 1:3],
                     class = dtbreast1$wdbc.Diagnosis,
                     modelType = "EDDA")

CV33 <- cvMclustDA(mod33.cv, nfold = 10)
unlist(CV33[c("ce", "se.ce", "brier", "se.brier")])
```

```
##          ce       se.ce       brier    se.brier
## 0.096660808 0.010532861 0.067365429 0.003671989
```

The 10-fold cross-validation confirms the **strong generalization performance** of the learner trained with Gaussian mixture discriminant analysis. The **classification error** is approximately 9.67%, with a **standard error of 1.05%**, indicating that about 1 in 10 tumors is misclassified on average across the folds, with very limited variability. The **Brier score**, equal to 0.0674, confirms that the predicted class probabilities are well-calibrated, and the **standard error of the Brier score** is 0.0037, again showing high stability.

These results are very close to those obtained on the independent test set, reinforcing the robustness of the model. The combination of low error, low Brier score, and small standard errors supports the use of this classifier for medical diagnosis settings, where accuracy and reliability are both crucial.

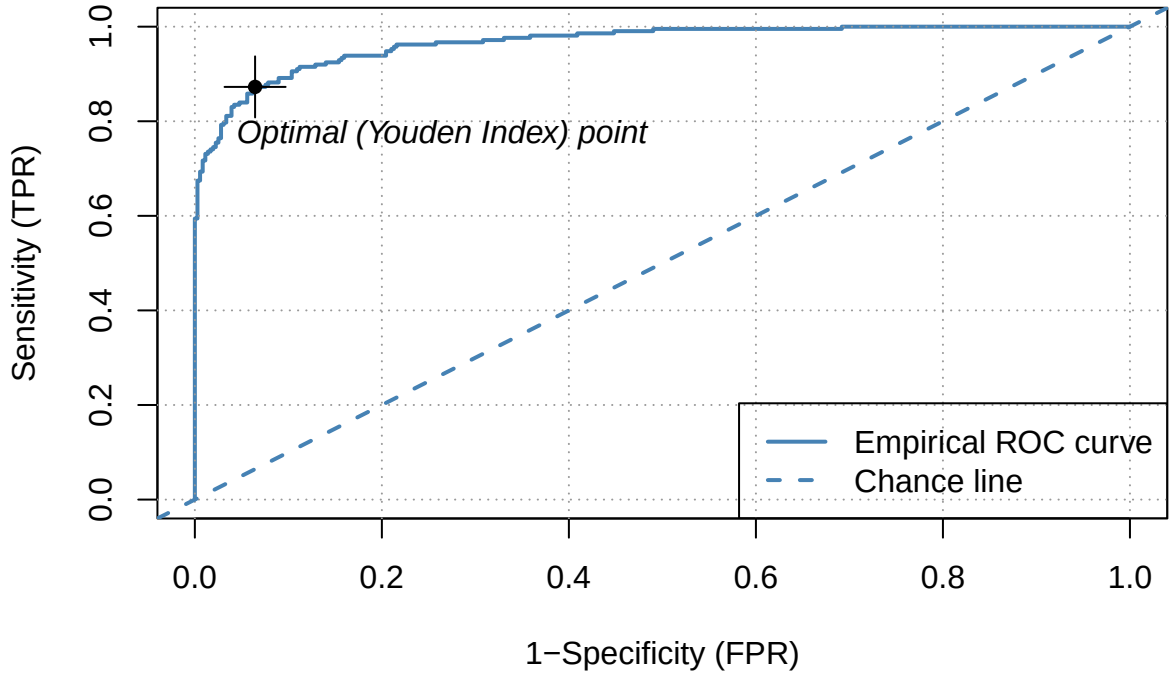
6.4.0.7 33.7 Determine the Receiver Operating Characteristc Curve (ROC) per different values of the cross-classification obtained with different classification probability thresholds. Depict the ROC curve and report and comment on the approximated value of the Area Under the Curve.

To evaluate the performance of the classifier across different decision thresholds, we compute and depict the **Receiver Operating Characteristic (ROC) curve** and report the corresponding **Area Under the Curve (AUC)**. The ROC curve displays the **trade-off between sensitivity and specificity**, while the AUC quantifies the model's ability to distinguish between the two classes.

We use the predicted probabilities of the *malignant* class from the 10-fold cross-validation performed in the previous step:

```
# compute ROC object using the probability of class M (column 2)
ROC.obj33 <- rocit(score = CV33$z[, 2], class = dtbreast1$wdbc.Diagnosis)

plot(ROC.obj33,
     YIndex = TRUE,
     legend = TRUE,
```

```
     legendPos = "bottomright",
     col = "steelblue",
     main = "ROC Curve - Breast Cancer Classification")
```



The ROC curve displayed confirms the excellent predictive ability of the classifier. The curve hugs the top-left corner of the graph, indicating a **high sensitivity across all thresholds** and very few false positives. The optimal cut-off point, highlighted via the **Youden Index**, balances sensitivity and specificity efficiently.

```
ROC.obj33$AUC
```

```
## [1] 0.966043
```

The corresponding **Area Under the Curve (AUC)** is approximately 0.966, a value that is very close to 1. This indicates that the classifier is **highly effective in distinguishing between malignant and benign tumors**. An AUC of this magnitude suggests that, in 96.6% of random pairs where one tumor is malignant and the other benign, the model assigns a higher probability of malignancy to the actual malignant case.

This high AUC validates the robustness of the classifier trained on `Radius_mean`, `Symmetry_mean`, and `Texture_mean`, and its suitability for clinical decision support in breast cancer diagnosis.

6.5 Protein classification with mixture discriminant analysis: variable association, EDDA model fitting, confusion matrix evaluation, cross-validation, and ROC-AUC performance.

An important problem in computational biology is to classify proteins into *functional* (coded as 0 in the data) and *structural* (coded as 1 in the data) classes (`type`) based on their sequence similarities. Protein molecules

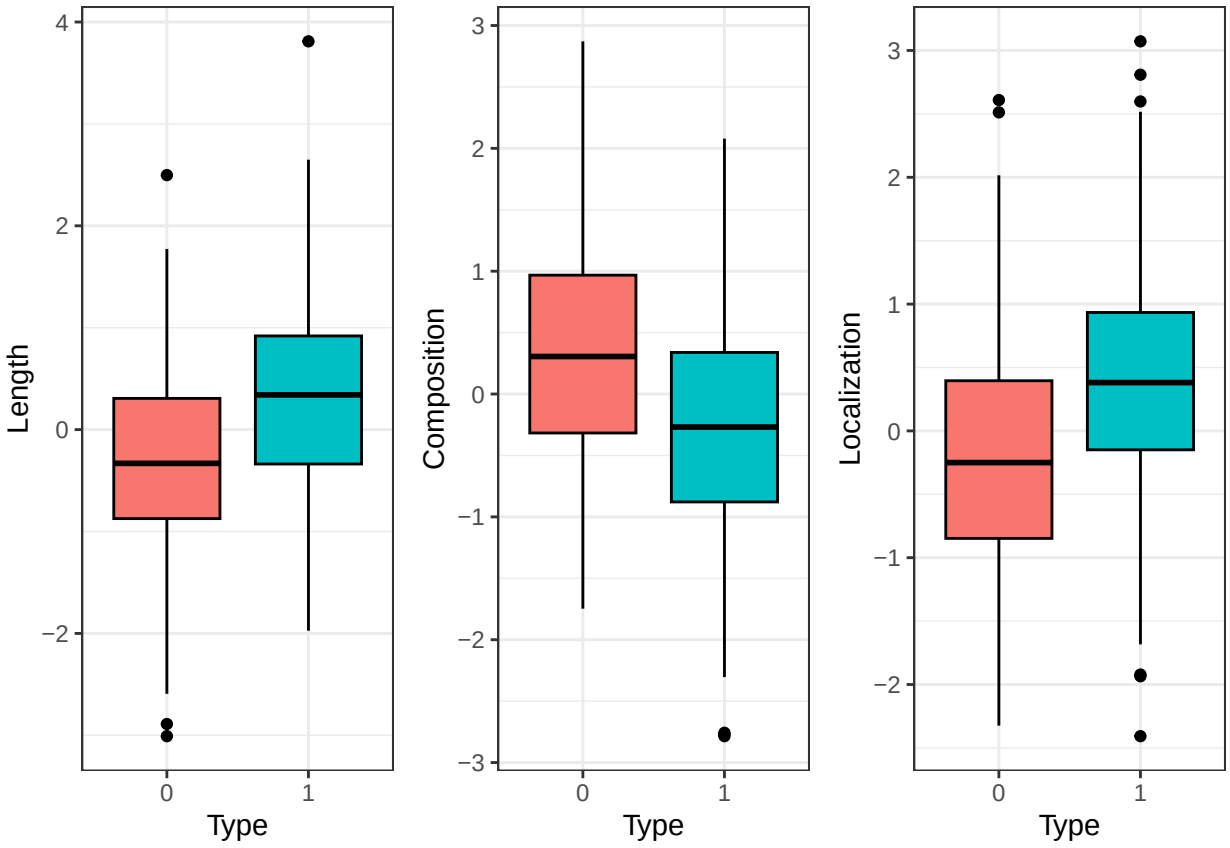
differ for length (`length`), composition (`comp`) and localization (`loc`). Data `dtprotein.Rdata` report the previous information with standardized features values collected on a sample of proteins (simulated data).

Report the analyses made on the previous exercise commenting the results according to the context.

```
load("dtprotein.Rdata")
dtprotein$type = as.factor(dtprotein$type)
```

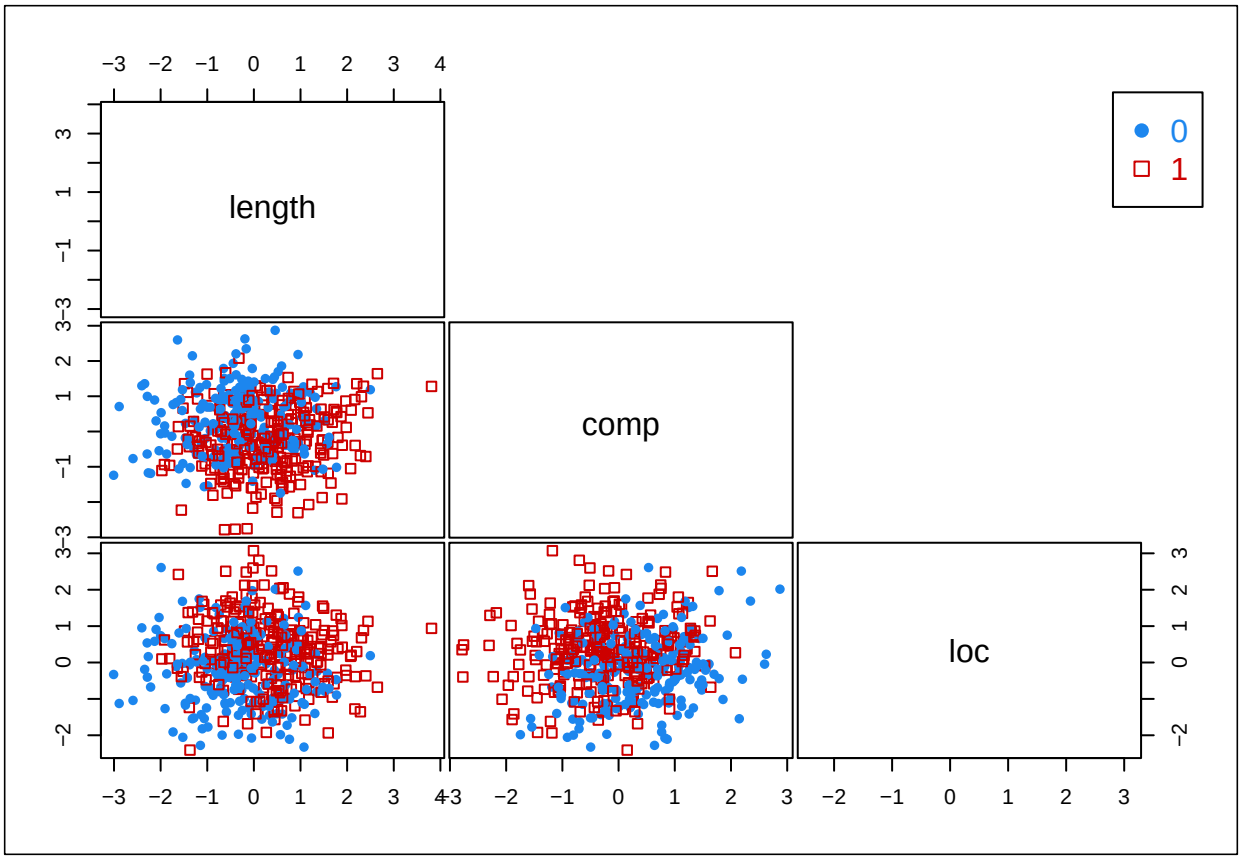
6.5.0.1 34.1 In order to evaluate the observed association of each measurement with the diagnosis, we represent the data through boxplots by protein type (0 = functional; 1 = structural).

```
require(ggplot2)

p1 <- ggplot(dtprotein, aes(type, length, fill = type)) +
  geom_boxplot(col = "black") +
  xlab("Type") + ylab("Length") +
  theme_bw() + theme(legend.position = "none")

p2 <- ggplot(dtprotein, aes(type, comp, fill = type)) +
  geom_boxplot(col = "black") +
  xlab("Type") + ylab("Composition") +
  theme_bw() + theme(legend.position = "none")

p3 <- ggplot(dtprotein, aes(type, loc, fill = type)) +
  geom_boxplot(col = "black") +
  xlab("Type") + ylab("Localization") +
  theme_bw() + theme(legend.position = "none")

require(gridExtra)
gridExtra::grid.arrange(p1, p2, p3, ncol = 3)
```



All three graphs show clear differences between proteins of functional class (0) and structural class (1). In particular, the latter present higher values of length and localization, while lower values of composition. More specifically, it is observed that, for example, approximately 25% of proteins in the functional class have a normalized localization value greater than (approximately) 0.5, compared to 50% of proteins in the structural class. We may also represent the data through a scatterplot matrix with point colors and shapes specific for each protein type.

```
sm34<- clPairs(dtprotein[, 2:4], dtprotein[, 1], upper.panel = NULL)
clPairsLegend(0.9, 0.9,
              class = sm34$class,
              col = sm34$col,
              pch = sm34$pch)
```



The above scatterplots do not show a clear separation between proteins of the two classes. Nonetheless, we can still observe some significant differences in the arrangement of points for the two groups of proteins. Looking at the last column of the plot matrix, for example, we notice that proteins of class functional assume on average lower values for the localization.

6.5.0.2 34.2 We split the data into training (0.7) and test (0.3) sets, sampling without replacement the rows of the original dataset.

```
set.seed(17)
n34 <- nrow(dtprotein)
ind_train <- sample(1:n34, size = n34*0.7, replace = FALSE)
X_train34 <- dtprotein[ind_train, ]
X_test34 <- dtprotein[-ind_train, ]
```

To check if the obtained training and test sets are reasonable, we may check the proportion of proteins of classes functional and structural within the two sets.

```
tab34 <- table(X_train34$type)
cbind(Counts = tab34, Perc = paste0(round(prop.table(tab34)*100, 1), "%"))
```

```
##   Counts Perc
## 0 "157"  "44.9%"
## 1 "193"  "55.1%"
```

```r
tab34 <- table(X_test34$type)
cbind(Counts = tab34, Perc = paste0(round(prop.table(tab34)*100, 1), "%"))
```

```
##   Counts Perc
## 0 "68"   "45.3%"
## 1 "82"   "54.7%"
```

Both the training and the test sets contain around 55% of structural proteins.

6.5.0.3 34.3

We fit a mixture discriminant analysis using eigenvalue decomposition through the `MclustDA` command. The argument `modelType = "EDDA"` specifies that a single component is used for each class, using the same covariance structure.

```r
model34 <- MclustDA(data = X_train34[, 2:4],
                    class = X_train34[, 1],
                    modelType = "EDDA")
summary(model34)
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -1458.827 350  8 -2964.517
##
## Classes   n     % Model G
##       0 157 44.86   EII 1
##       1 193 55.14   EII 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 102  55
##     1  35 158
## Classification error = 0.2571
## Brier score          = 0.1758
```

The estimated model has a covariance structure of type `EII`: the model is spherical (*II), with equal volume (E**). In other words, the three variables are assumed to have the same variability and pairwise covariance equal to zero. The corresponding number of parameters is 7 (3 means in the first components, 3 means in the second component and 1 variance common to all the variables and all the components). To obtain the estimated model parameters, we add the argument `parameters = TRUE`.

```r
summary(model34, parameters = TRUE)$parameters
```

```
## [[1]]
## [[1]]$pro
## [1] 1
```

```
##
## [[1]]$mean
##              [,1]
## length -0.3528607
## comp    0.2103522
## loc    -0.1882053
##
## [[1]]$variance
## , , 1
##
##           length      comp       loc
## length 0.8469374 0.0000000 0.0000000
## comp   0.0000000 0.8469374 0.0000000
## loc    0.0000000 0.0000000 0.8469374
##
##
##
## [[2]]
## [[2]]$pro
## [1] 1
##
## [[2]]$mean
##              [,1]
## length  0.3330638
## comp   -0.2912951
## loc     0.4131361
##
## [[2]]$variance
## , , 1
##
##           length      comp       loc
## length 0.8469374 0.0000000 0.0000000
## comp   0.0000000 0.8469374 0.0000000
## loc    0.0000000 0.0000000 0.8469374
```

Looking at the mean vectors allow us to characterize the two sub-populations from the point of view of the application context. The first group is characterized by a negative average value for variables length and localization, and a positive average value for composition (functional proteins). The second group shows an opposite behavior, with positive values for length and localization and negative a value for composition (structural proteins). The common standard deviation is equal to $\sqrt{0.85} = 0.92$, which is quite high if compared the mean values.

6.5.0.4 34.4 The confusion matrix shows the predicted class of proteins cross-tabulated for their true class. We observe that the majority of proteins is correctly classified by the estimated model 102 of class functional and 158 of class structural. 90 proteins have been assigned to the wrong class (55 proteins of type functional and 35 proteins of type structural).

From the confusion matrix we can compute values of sensitivity and specificity (assuming 1 as the "positive" class):

- sensitivity (true positive rate): $\frac{\text{TP}}{\text{TP+FN}} = \frac{158}{158+35} = 0.82$;
- specificity (true negative rate): $\frac{\text{TN}}{\text{TN+FP}} = \frac{102}{102+55} = 0.65$.

The summary of the estimated model also reports the classification error (misclassification rate), equal to 0.257. It represents the proportion of units that are not correctly classified, showing that around 26% of proteins have been assigned to the wrong class.

Finally the output also provides the Brier score, equal to 0.18. This value, which is not close to 0, shows that the estimated model has a discrepancy between its predictions and the actual protein classes.

6.5.0.5 34.5

We use now the test set to evaluate the prediction performance of the estimated model.

```
summary(model34, newdata = X_test34[, 2:4], newclass = X_test34[, 1])
```

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -1458.827 350  8 -2964.517
##
## Classes   n     % Model G
##       0 157 44.86   EII 1
##       1 193 55.14   EII 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 102  55
##     1  35 158
## Classification error = 0.2571
## Brier score          = 0.1758
##
## Test confusion matrix:
##      Predicted
## Class  0  1
##     0 51 17
##     1 22 60
## Classification error = 0.26
## Brier score          = 0.1686
```

The test confusion matrix shows that the majority of the 98 proteins of class functional are correctly classified. In particular result correctly classified: 51 proteins of class functional (out of 68) and 60 proteins of class structural (over 82).

Both the classification error (0.260) and the brier score (0.169) are very close to the values obtained assessing the predictive performance on the training set. Overall 74% of the units are correctly classified by the estimated model.

6.5.0.6 34.6

To better assess the predictive performance of the estimated model, we employ V-fold Cross Validation ($V = 10$).

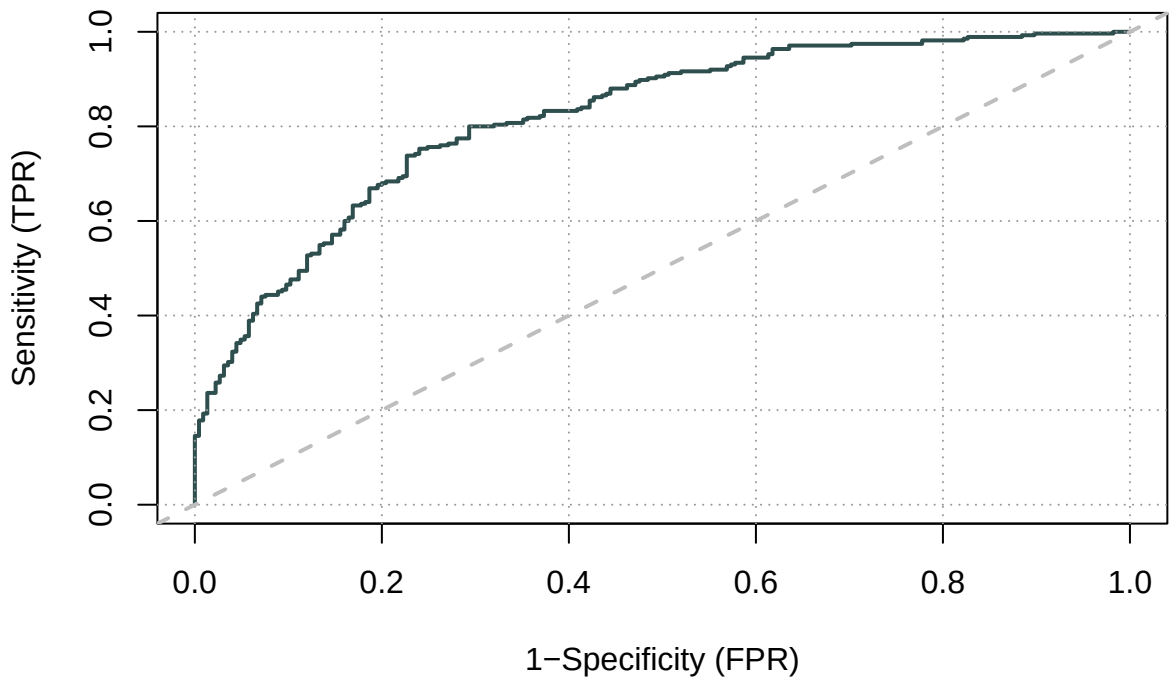
```r
mod34.C <- MclustDA(data = dtprotein[, 2:4],
                    class = dtprotein[, 1],
                    modelType = "EDDA")

CV34 <- cvMclustDA(mod34.C, nfold = 10)
unlist(CV34[c("ce", "se.ce", "brier", "se.brier")])
```

```
##         ce      se.ce      brier   se.brier
## 0.25800000 0.01935555 0.17395603 0.00700368
```

The classification error obtained through the 10-fold Cross-Validation is equal to 0.258; it means that about 75% of units in the dataset are now correctly classified. Note that this value is very close to that obtained assessing the prediction performance on a single test set, but is more robust. The corresponding standard error is equal to 0.019. The Brier score is equal to 0.174, with a standard error of 0.007. Both the classification error and the Brier score are very similar to the two values computed on the training set, showing that the model mainly avoids overfitting.

6.5.0.7 34.7 Finally we represent the ROC and compute the corresponding AUC for different values of the threshold.

```r
ROC.obj34 <- rocit(score = CV34$z[, 2], class = dtprotein$type)
plot(ROC.obj34, YIndex = FALSE, legend = FALSE)
```



```
ROC.obj34$AUC
```

```
## [1] 0.8189899
```

The value of the AUC is equal to 0.82, indicating a good classification performance.