Financial Mathematics and Statistics

Msc Finance (2022-2023)

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

This document contains the statements but also the corrections of the exercises treated during the different Lab sessions.

The exercises covered in the document range from elementary reminders in probability, through the construction of confidence intervals, hypothesis tests on parameters such as the mean or the proportion. This last notion is extended to tests of comparisons of means between several populations with the one or two factor ANOVA.

The last part of the course is devoted to simple and multiple linear models for the prediction of a real numerical variable. A small digression on logistic regression is also discussed.

Contents

1 Lab 1

1.1 Managing Ashland Multicomm Services

In this first part, the technical services department of AMS worked on a project to improve the service and the quality of access to Internet of these customers, for that they carried out measurements as for the speed of loading and have for objective a value of 1.0 in term of speed of loading. Their study is based on data from the previous year and they show that the data, which represent the loading speed, are distributed according to a normal law of mean $\mu = 1.005$ and standard deviation $\sigma = 0.10$. The loading speed is considered acceptable if the measure is between 0.95 and 1.05.

1. We assume that the distribution of upload speed remains the same as it was last year.

   I remind you that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X-\mu}{\sigma}$ is normally distributed with mean and variance respectively equal to 0 and 1, *i.e.* $Z \sim \mathcal{N}(0,1)$.

   The exercise can be done using the Z-table or the function `NORM.INV` of Excel

   - $\mathbb{P}[X < 1] = 0.4801$
   - $\mathbb{P}[0.95 < X < 1.0] = \mathbb{P}[X < 1.0] - \mathbb{P}[X < 0.95] = 0.1889$
   - $\mathbb{P}[1 < X < 1.05] = \mathbb{P}[X < 1.05] - \mathbb{P}[X < 1] = 0.1936$
   - $\mathbb{P}[(X < 0.95) \cup (X > 1.05)] = 1 - \mathbb{P}[0.95 < X < 1.05] = 0.6175$

2. We now want to reduce the probability that the upload speed is below 1.0. We wonder if it is better to improve the mean upload speed or to reduce the standard deviation to the upload speed. To answer to this question, we will have to compare the probability that the upload speed takes values below 1 and select the process for which this probability is minimized

   - If we increase the value of μ to 1.05, we have:

$$
\begin{aligned}
\mathbb{P}[X < 1] &= \mathbb{P}\left[\frac{X-\mu}{\sigma} < \frac{1-1.05}{0.1}\right], \\
&= \mathbb{P}\left[Z < -0.5\right], \\
&\quad \downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > 0.5\right],
\end{aligned}
$$

$$
\begin{aligned}
& \quad \downarrow \text{ complementary events} \\
&= 1 - \mathbb{P}\left[Z < 0.5\right], \\
\mathbb{P}[X < 1] &= 0.3085.
\end{aligned}
$$

- In the case we decrease the value of σ to 0.075, we have:

$$
\begin{aligned}
\mathbb{P}[X < 1] &= \mathbb{P}\left[\frac{X-\mu}{\sigma} < \frac{1-1.005}{0.075}\right], \\
&= \mathbb{P}\left[Z < -0.067\right], \\
& \quad \downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > -0.067\right], \\
& \quad \downarrow \text{ complementary events} \\
&= 1 - \mathbb{P}\left[Z < 0.067\right], \\
\mathbb{P}[X < 1] &= 0.4734.
\end{aligned}
$$

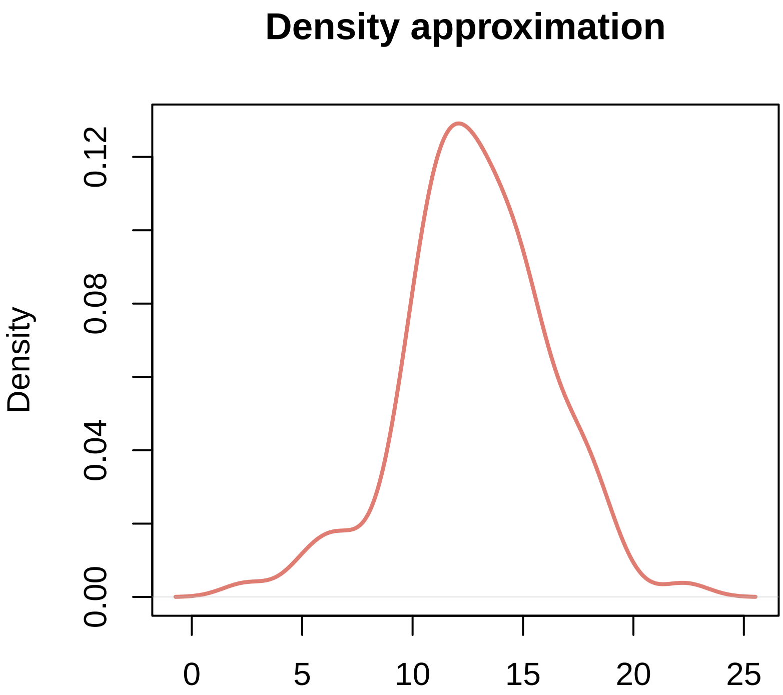
It it is then preferable to increase the mean upload speed to improve user experience.

1.2 Digital Case

The aim of this exercise is to study the veracity of a report on a study conducted in a company.

Our data have the following distribution

```r
# Import the data
data = read.csv("data/DownloadTimes.csv", header = FALSE)
# We plot the density
d <- density(data$V1)
plot(d, lwd=2,col = "#DF7D72", xlab="", main = "Density approximation")
```

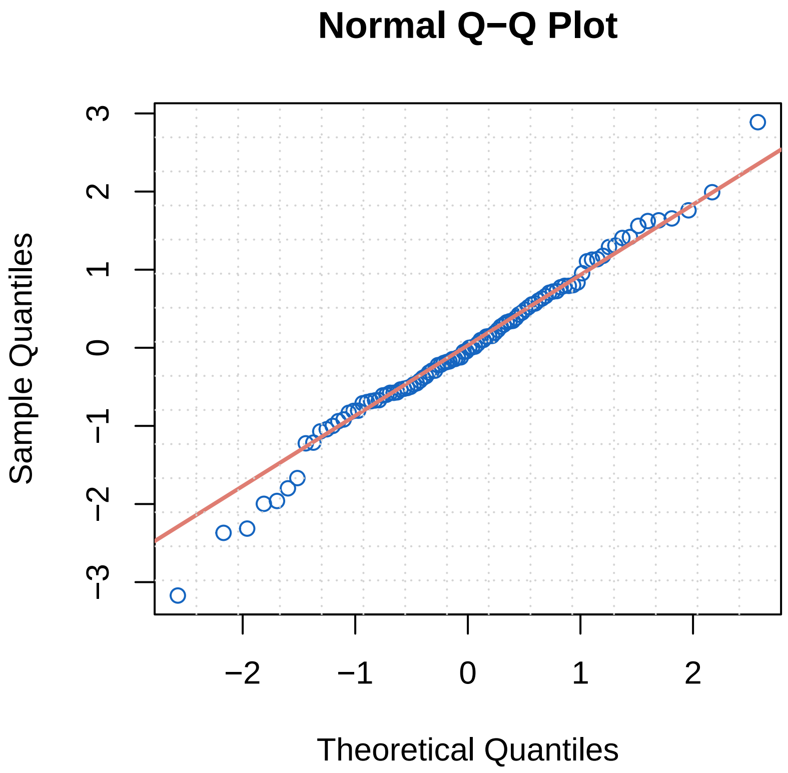


1. We start by looking if the data are normally distributed. To do this we will start by plotting the QQ-plot (also called the *Normal Probability Plot*, see Section 6.3 of your reference *Pearson*) of our data, this is to plot the empirical quantiles of the data vs. the empirical quantiles of the centered reduced normal distribution.

   If the points are aligned on a straight line, then we can say that our data are distributed according to a normal distribution. (See Excel). Attention, the only criterion $mean = median = mode$ is not sufficient to say that data are from a normal distribution !

   Your graph shall have the following form:

```
# We compute the Empirical Quantiles
data_plot = (data$V1 - mean(data$V1))/sd(data$V1)
# Plot the normal probability plot
qqnorm(data_plot, col = "#1565C0")
qqline(data_plot, col = "#DF7D72", lwd = 2)
grid(nx = 15, ny = 15, col = "lightgray", lty = "dotted",
    lwd = par("lwd"), equilogs = TRUE)
```



2. We will now check if the conclusion made by the experts are correct or not using the data (see the associated Excel File)

-
  ```
  # Computation of the mean
  mean_sample = mean(data$V1)
  mean_sample

  ## [1] 12.8596

  # Computation of the median
  median_sample = median(data$V1)
  median_sample

  ## [1] 12.785
  ```

  The mean value is accurate. On the other hand, we note that the median is slightly lower than the mean, however, this difference can only be explained by our sample size. Further analysis will be required to see if this difference is statistically different.

- This assertion is false, as a reminder the probability that a continuous law takes an exact value is zero. It would have been better to study the probability that a download time is less than a given value.
- The assertion is not good, removing extreme values contributes to strongly reduce the mean of our sample but especially its variance, so in general, we will find less data in the interval :

  $$\bar{x} - 3s; \bar{x} + 3s,$$

  where $\bar{x}$ stands for the empirical mean (the mean evaluated on the sample) and s for the standard deviation evaluated on the sample.

  The only thing we can say then is that the probability of observing loading times of less than 22.7 seconds will then increase.
- A quick analysis of our Excel spreadsheet shows us that this assertion seems to be true for the observed sample. On the other hand, be careful with the interpretation of this result. The data allowing us to say that the probability that we observe a loading time higher than the value of 17.06 is indeed 0.1, but this does not mean that the phenomenon will be observed in 10% of the cases, it is an *asymptotic* result.

  ```
  quantile(data$V1,0.9)

  ##    90%
  ## 17.096
  ```

  It means that 90% of our sample values are lower or equal than 17.096.
- We have to check if 99% of the data lies in the interval

  $$\bar{x} - 3s; \bar{x} + 3s,$$

  to say that the process follows the Six Sigma benchmark for industrial quality. Thus, we only to see if 99% of the data belongs in this interval.

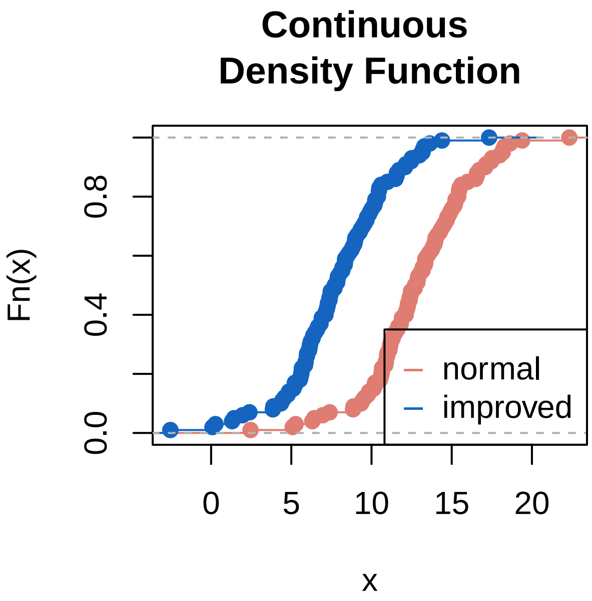
  ```
  # Lower bound
  L = mean(data$V1)-3*sd(data$V1)
  L

  ## [1] 3.021766

  # Upper bound
  U = mean(data$V1)+3*sd(data$V1)
  U
  ```

```
## [1] 22.69743

# Proportion of the data in the interval
Rate = mean( ((data$V1 >=L)&&(data$V1 <= U)))
Rate

## [1] 1
```

So the conditions are met with this sample, but it quite normal since the data are normally distributed.

3. Improving the loading time means reducing the average loading time. If we reason about the empirical distribution function, this will have the effect of increasing the observed probabilities for a given value of loading time. This can be seen by using the 15 seconds example used earlier.

```
data_normal = data$V1
data_improved = data$V1 - 5
plot(ecdf(data_normal),col = "#DF7D72",main="Continuous
Density Function", xlim=c(min(data_normal,data_improved)-0.1,
max(data_normal,data_improved)+0.1))
lines(ecdf(data_improved),col="#1565C0")
legend("bottomright",pch='_',c("normal","improved"),
col=c("#DF7D72","#1565C0"))
```



1.3 Nestlé Purina Petcare: part I

We consider the production of 85g cans. The process includes: filling in the cans, sealing, weighing and disposal if the weight is less than 85g (production waste), sorting, labeling. Even if the automatons are of very good quality, it is impossible to obtain cans weighing exactly 85g. There is a standard deviation in packing the cans related to the technical processes, which follows a normal distribution of the standard deviation equal to 0.8g.

1. We aim to determinate the mean value of packing the cans that has to be used so that the waste related to the weight of package is less than 2%.

   Let us denote as X the random variable used to describe the weight of a can. The technical processes follows a normal distribution $\mathcal{N}(\mu, 0.8^2)$ and we want to determine μ such that :

   $$P[X \leq 85] \leq 0.2.$$

   As usual we have to consider the standardization to solve the problem. It is also the only for us to ensure that the value μ appears.

   $$P[X \leq 85] = P\left[\frac{X-\mu}{0.8} \leq \frac{85-\mu}{0.8}\right] = P\left[Z \leq \frac{85-\mu}{0.8}\right].$$

   We now want to determine a value z such that

   $$P\left[Z \leq z\right] \leq 0.2.$$

   Note that is the same problem as finding z such that $P\left[Z \leq z\right] = 0.2$. Once the value z is found we have to use that fact that

   $$z = \frac{85-\mu}{0.8}$$

   to have the value of the mean. Using the z-table we found that z shall be equal to -2.055. Then μ is equal to

   $$\mu = 85 - 0.8 \times z = 85 + 0.8 \times 2.055 = 86.65\text{g}.$$

We are now interested in the production of 100g packages. Once the mixture has been prepared, it is packaged by the automatons. The process includes: packing, sealing, weighing and disposal if the weight is less than 98.5g or bigger than 101.5g, sorting, labeling. The automatons are the same as those used for the 85g packages. The standard deviation in packing, related to the technical processes, follows, also for the 100g packages, a normal distribution of the standard error equal to 0.8g.

2. Knowing that the automatons are set to a mean $\mu = 100$g, we are interested in the percentage of packages that are then discarded as waste, statistically speaking.

   To answer this question we will apply the same process as before, we consider $X \sim \mathcal{N}(100, 0.8^2)$ and compute the following probabilities. Then the probability of being a reject is the sum of these two probabilities.

$$P[X \leq 98.5] \quad \text{and} \quad P[X \leq 101.5].$$

$$\begin{aligned}
\mathbb{P}[X < 98.5] &= \mathbb{P}\left[\frac{X-100}{0.8} < \frac{98.5-100}{0.8}\right], \\
&= \mathbb{P}\left[Z < -1.875\right], \\
&\quad \downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > 1.875\right], \\
&\quad \downarrow \text{ complementary events} \\
&= 1 - \mathbb{P}\left[Z < 1.875\right], \\
\mathbb{P}[X < 98.5] &= 0.03.
\end{aligned}$$

   We also have $P[X \leq 101.5] = 0.03$ because the gaussian is centered at the position 100. Thus the probability of being a reject is equal to 0.06.

3. Would it be appropriate to set the machines to a mean $\mu = 100.5$?

   The answer is clearly no! And the justification is simply based on the fact that the value of a gaussian are centered around the mean, *i.e.* most of them are concentrated around the mean. Furthermore, the set of acceptable cans is also centered at the position 100, so if if we shift the distribution, we will necessarily increase the probability of being a reject for a can.
   This fact can also be shown by computing the probability.

1.4 Nestlé Purina Petcare: part II

To check the adjustment of the machine, it proceeds to a random sampling of cans on the production line. The company wants to check whether the packing mean is always equal to 86.65g.

For that reason, the company has sampled today 100 cans. The mean of that sample of 100 cans is 86.5g.

1. We first aim to give a confidence interval of the mean with a confidence rate equal to $1 - \alpha = 95\%$ in order the check the good behaviour of the process.
   We will see later another tool which will provide us and the answer.

   We still know that the filling process is normally distributed and we consider X the random variable which represents the weight of a can, *i.e.* $X \sim N(\mu, 0.8^2)$.

   We want a confidence interval **on the mean and we know the standard deviation of the distribution. Thus the later will be based on the Z-distribution.** Thus the confidence interval will have the following form:

   $$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

   where n is the sample size, $\bar{x}$ is the mean value on the sample, σ is the standard deviation of the distribution and $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of a centered and reduced normal distribution. It remains to compute the bound using the information, $\alpha = 0.05$, thus $z_{0.975} = 1.96$

   ```
   n = 100
   z = 1.96
   x = 86.5
   sigma = 0.8
   # Lower Bound
   x-z*sigma/sqrt(n)

   ## [1] 86.3432

   # Upper Bound
   x+z*sigma/sqrt(n)

   ## [1] 86.6568
   ```

2. Can we conclude there is a miss-adjustment of the machines with a 5% error risk?

   The value $\mu = 86.65$ is in the previous interval, so we cannot say that the machine is miss-adjusted.

3. Let us now analyze the two previous result when we increase the sample size to $n = 1,000$.

   We apply the same process as the one used in question 1.. We simply replace $n = 100$ by $n = 1,000$.

```
n = 1000
z = 1.96
x = 86.5
sigma = 0.8
# Lower Bound
x-z*sigma/sqrt(n)

## [1] 86.45042

# Upper Bound
x+z*sigma/sqrt(n)

## [1] 86.54958
```

This time, we can conclude there is a miss-adjustment of the machines with a 5% error risk.

1.5 IT Consulting Firm

An information technology (IT) consulting firm specializing in health care solutions wants to study communication deficiencies in the health care industry. A random sample of 70 health care clinicians reveals the following:

- Time wasted in a day due to outdated communication technologies:

$$\bar{x} = 45 \text{ minutes} \quad \text{et} \quad s = 10 \text{ minutes}.$$

- Thirty-six health care clinicians cite inefficiency of pagers as the reason for the wasted time, *i.e.* $\bar{p} = \dfrac{36}{70} = \dfrac{18}{35}$

1. We first want to build a confidence interval to estimate for the population mean time wasted in a day due to outdated communication technologies. The rate of confidence equal uses for this interval is equal to $1 - \alpha = 0.99$, *i.e.* $\alpha = 0.01$.

   First we have to identify in which situation we are. **Here we want an estimation of a mean when the standard deviation at the population scale is unknown**, that is why the value s is provided.
   Our confidence interval will be based on the *Student* distribution and will have the following form

$$\left[\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ;\ \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right],$$

where n is the sample size and $t_{1-\alpha/2,n-1}$ denotes the quantile of order $1-\alpha/2$ of student law where the number of degree of freedom is equal to $n-1$.

In our context, our interval can be rewritten as:

$$\left[45 - t_{0.995,69}\frac{10}{\sqrt{70}}\ ;\ 45 + t_{0.995,69}\frac{10}{\sqrt{70}}\right],$$

and $t_{0.995,69} \simeq 2.66$ using the t-table. It remains to compute the values.

```
x = 45
s = 10
n=70
t = 2.66
# Lower Bound
x-t*s/sqrt(n)

## [1] 41.82069

# Upper Bound
x+t*s/sqrt(n)

## [1] 48.17931
```

2. We want now to apply the same process to construct a 95% confidence interval estimate for the population proportion of health care clinicians who cite inefficiency of pagers as the reason for the wasted time.

   We want to estimate proportion. Thus our confidence interval will be based on the center and reduced normal distribution and will have the following form:

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\ ;\ \bar{p} + \bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right],$$

   where n is the sample size and $z_{1-\alpha/2}$ denotes the quantile of order $1-\alpha/2$ of center and reduced normal law.

In our context, $1-\alpha = 0.95$, *i.e.* $\alpha = 0.05$ and $\bar{p} = \frac{18}{35}$ So our interval can be rewritten as:

$$\left[\frac{18}{35} - z_{0.975}\sqrt{\frac{\frac{18}{35}\left(1-\frac{18}{35}\right)}{70}}\;;\;\frac{18}{35} + z_{0.975}\sqrt{\frac{\frac{18}{35}\left(1-\frac{18}{35}\right)}{70}}\right],$$

and $z_{0.975} \simeq 1.96$ using the z-table. It remains to compute the values.

```
p = 18/35
n=70
z = 1.96
# Lower Bound
p-z*sqrt(p*(1-p)/n)

## [1] 0.3972011

# Upper Bound
p+z*sqrt(p*(1-p)/n)

## [1] 0.6313703
```

2 Lab 2

2.1 Managing Ashland MultiComm Services

The technical operations department wants to ensure that the mean target upload speed for all Internet service subscribers is at least 0.97 on a standard scale in which the target value is 1.0. Each day, upload speed was measured 50 times, with the following results

```
# Import the data
data = read.csv("data/AMS9.csv", sep=";")
data = data$Upload.Speed
```

1. Compute the sample statistics and determine whether there is evidence that the population mean upload speed is less than 0.97.

   For this first question we will do the following test (with an error rate of $\alpha = 0.05$)

   $$H_0 : \mu = 0.97 \text{ vs. } H_1 : \mu < 0.97$$

   In other words, we aim to test the assumption that the mean upload speed if greater than 0.97 and we hope that there is nothing in our data that proof the contrary.

   Since we do not know the variance of the population, the statistical value we use is based on the *Student* distribution:

   $$t_{\text{stat}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},$$

   where μ refers to the tested value.

   ```
   # Compute the quantities
   mu = 0.97
   x_bar = mean(data)
   x_bar

   ## [1] 0.95872

   s = sd(data)
   s
   ```

```
## [1] 0.1559678

n = length(data)
n

## [1] 50

t_stat = (x_bar-mu)/(s/sqrt(n))
t_stat

## [1] -0.5113981
```

We can now compare this value to a critical one or compute the p-value and compare this value with α. Here, we perform a lower t-test so the p-value is defined as $\mathbb{P}(Z < z_{\text{stat}})$

```
# Computation of the p-value
p_value = pt(t_stat,n-1)
p_value

## [1] 0.3056846

# Critical value
t_crit = qt(0.05,n-1)
t_crit

## [1] -1.676551

t_stat<t_crit

## [1] FALSE
```

There is no evidence that the mean upload speed is lower than 0.97.

2. Write a memo to management that summarizes your conclusions.

   According to the previous question there is no evidence that the mean upload speed is lower than 0.97. However, if we apply the same process considering $\mu = 1.0$, we have:

```
# Statistical value
mu = 1
t_stat = (x_bar-mu)/(s/sqrt(n))
t_stat

## [1] -1.871499

# Computation of the p-value
p_value = pt(t_stat,n-1)
p_value

## [1] 0.033626
```

There is a statistical evidence that the mean upload speed is lower than 1.0.

2.2 Coffee Shop

The owner of a specialty coffee shop wants to study coffee purchasing habits of customers at her shop. She selects a random sample of 60 customers during a certain week, with the following results:

- The amount spent was $\bar{x} = 7.25$ and $s = 1.75$.
- Thirty-one customers say they *"definitely will"* recommend the specialty coffee shop to family and friends.

1. At the $\alpha = 0.05$ level of significance, is there evidence that the population mean amount spent was different from 6.50$?

   We apply the same process as in the previous question and formulate the following test:

   $$H_0 : \mu = 6.5 \text{ vs. } H_1 : \mu \neq 6.5$$

   Since we do not know the variance of the population, the statistical value we use is based on the *Student* distribution:

   $$t_{\text{stat}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},$$

   where μ refers to the tested value.

```
# Compute the quantities
mu = 6.5
x_bar = 7.25
s = 1.75
n = 60

t_stat = (x_bar-mu)/(s/sqrt(n))
t_stat

## [1] 3.3197
```

We can now compare this value to a critical one or compute the p-value and compare this value with α. Here, we perform a two-sided t-test so the p-value is defined as $2(1-\mathbb{P}(T<|t_{\text{stat}}|))$

```
    # Critical value
    t_crit = qt(0.975,n-1)
    t_crit

## [1] 2.000995

    abs(t_stat)>t_crit

## [1] TRUE
```

There is a statistical evidence that the population mean is different from 6.50$

2. Determine the p-value

```
# Computation of the p-value
p_value = 2*(1-pt(abs(t_stat),n-1))
p_value

## [1] 0.001548771
```

3. At the 0.05 level of significance, is there evidence that more than 50% of all the customers say they *"definitely will"* recom- mend the specialty coffee shop to family and friends?

We will this time perform a test on the proportion and make the following assumptions

$$H_0 : p = 0.5 \text{ vs. } H_1 : p \geq 0.5$$

The statistical value we use is based on the Z distribution:

$$z_{\text{stat}} = \frac{\bar{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}},$$

where p refers to the tested value and $\bar{p}$ to the observed proportion on the sample.

```r
# Compute the quantities
p = 0.5
p_bar = 31/60
n = 60

z_stat = (p_bar-p)/(sqrt(p*(1-p)/n))
z_stat

## [1] 0.2581989
```

We can now compare this value to a critical one or compute the p-value and compare this value with α. Here, we perform an upper z-test so the p-value is defined as $\mathbb{P}(Z > z_{\text{stat}})$

```r
# Computation of the p-value
p_value = 1-pnorm(z_stat)
p_value

## [1] 0.3981267

# Critical value
z_crit = qnorm(0.95)
z_crit

## [1] 1.644854

z_stat>z_crit

## [1] FALSE
```

There is no evidence that more than 50% of all the customers say they *"definitely will"* recommend the specialty coffee shop to family and friends.

4. What is your answer to 1. if the sample mean equals 6.25$?

   We will perform the same computation and just change the values

```
# Compute the quantities
mu = 6.5
x_bar = 6.25
s = 1.75
n = 60

t_stat = (x_bar-mu)/(s/sqrt(n))
t_stat

## [1] -1.106567

# Compute the p-value
p_value = 2*(1-pt(abs(t_stat),n-1))
p_value

## [1] 0.2729728
```

   This time, we do not reject H_0 and we cannot say that the mean spent different from 6.50$.

5. What is your answer to 3. if 39 customers say they "definitely will" recommend the specialty coffee shop to family and friends?

   We will apply the same process as in question 3.

```
# Compute the quantities
p = 0.5
p_bar = 39/60
n = 60

z_stat = (p_bar-p)/(sqrt(p*(1-p)/n))
z_stat

## [1] 2.32379
```

```
# Computation of the p-value
p_value = 1-pnorm(z_stat)
p_value

## [1] 0.01006838

# Critical value
z_crit = qnorm(0.95)
z_crit

## [1] 1.644854

z_stat>z_crit

## [1] TRUE
```

There is an evidence that more than 50% of all the customers say they *“definitely will”* recommend the specialty coffee shop to family and friends.

2.3 Managing Ashland Multicomm Services

AMS communicates with customers who subscribe to telecommunications services through a special secured email system that sends messages about service changes, new features, and billing information to in-home digital set- top boxes for later display. To enhance customer service, the operations department established the business objective of reducing the amount of time to fully update each subscriber’s set of messages. The department selected two candidate messaging systems and conducted an experiment in which 30 randomly chosen cable subscribers were assigned one of the two systems (15 assigned to each system). Update times were measured.

```
# Import the data
data = read.csv("data/AMS10.csv", sep=";")
user1 = data$U1
user2 = data$U2
```

1. Analyze the data and write a report to the computer operations department that indicates your findings.

The aim of this first exercise is to compare the mean of update times of the groups of users.

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_0 \neq \mu_1.$$

Before applying the test to compare the means, we first need to test if the two variances are equal or not using the test of Fisher. We first need to compute the variance of the two groups.

```r
# Compute variances
v1 = var(user1)
v1

## [1] 0.08896857

v2 = var(user2)
v2

## [1] 0.1272457

v1>v2

## [1] FALSE
```

Here the variance of group 2 is greater than the variance of group 1. Let us check that this difference is significant using:

$$F_{\text{stat}} = \frac{s_2^2}{s_1^2}.$$

```r
# Proceed to the test
n1 = length(user1)
n2 = length(user2)

F_stat = v2/v1
F_stat

## [1] 1.430232

# Compute the p-value
p_value = 1-pf(F_stat,n2-1, n1-1)
p_value
```

```
## [1] 0.2559432
```

We cannot say that variances are different. Thus, to compare the means, we will use the following statistical quantity which follows a t-distribution:

$$t_{\text{stat}} = \frac{\bar{x_1} - \bar{x_2}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

where s is defined by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 2)s_2^2}{n_1 + n_2 - 2}.$$

It remains to perform the test.

```
# Compute the quantities
x1_bar = mean(user1)
x2_bar = mean(user2)
s = sqrt( ((n1-1)*v1 + (n2-1)*v2)/(n1+n2-2))
s

## [1] 0.3287965

# Compute the statistical value
t_stat = (x1_bar-x2_bar)/(s*sqrt((1/n1 + 1/n2)))
t_stat

## [1] -3.315023

# Compute the p-value
p_value = 2*(1-pt(abs(t_stat),n1+n2-2))
p_value

## [1] 0.002540922
```

We can then say that the two mean values are statistically different.

2. Suppose that instead of the research design described in the case, there were only 15 subscribers sampled, and the update process for each subscriber email was measured for each of the two messaging systems.

   This time, we suppose that there is dependence between the two populations because we are dealing with *repeated measures.*
   In other words, if we aim to prove a significant difference between these two series of values, we have to consider a new sample $D = (d_1, d_2, \ldots, d_n)$ where for each index i we have:

   $$d_i = x_i^{(1)} - x_i^{(2)}.$$

   ```
   # Compute the new sample
   D = user1-user2
   ```

   Then, we formulate the following test

   $$H_0 : \mu_D = 0 \text{ vs. } H_1 : \mu_D \neq 0.$$

   The procedure is then the same as the one which consists in testing the mean on a sample. It is based on t-distribution. Let us do it.

   ```
   # Compute the statistical values
   mu_D = 0
   x_bar_D = mean(D)
   s_D = sd(D)
   n = length(D)
   t_stat = (x_bar_D - mu_D)/(s_D/sqrt(n))
   t_stat

   ## [1] -3.596283

   # Compute the critical value
   t_crit = qt(0.975,n-1)
   t_crit

   ## [1] 2.144787

   abs(t_stat)>t_crit

   ## [1] TRUE
   ```

```
# Compute the p-value
p_value = 2*(1-pt(abs(t_stat),n-1))
p_value

## [1] 0.002919939
```

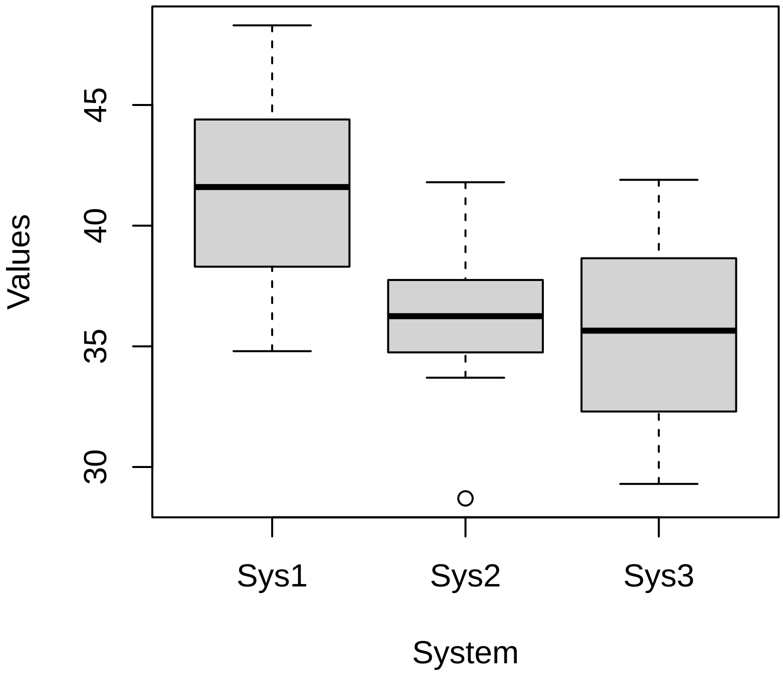
We can still say that there is significant difference between the two mean values.

2.4 A Study of Different Systems

2.4.1 Phase 1: Anova with one factor

The computer operations department had a business objective of reducing the amount of time to fully update each subscriber's set of messages in a special secured email system. An experiment was conducted in which 24 subscribers were selected and three different messaging systems were used. Eight subscribers were assigned to each system, and the update times were measured

```
# Import the data
data = read.csv("data/AMS11.1.csv", sep=";")
boxplot(Values~System,data)
```



Graphically, we can already see a significant difference between the different groups. We now propose to show this by statistical reasoning

Analyze the data and write a report to the computer operations department that indicates your findings.

We are going to see if there is a statistical difference between the three groups using an ANOVA. I will simply use the appropriate function to get the results.

```
anova <- aov(Values~System,data)
summary(anova)

##             Df Sum Sq Mean Sq F value Pr(>F)
## System       2  173.3   86.67   5.135 0.0153 *
## Residuals   21  354.4   16.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As it is shown by the previous computations, the system has a significant impact on the the observed values.

Because the samples have the same size in each group, there is no need to check the variance of each group (using the *Levene Test* or Bartlett one), but we are going to do it in order to be sure that this assumption is checked.

```
library(car)
leveneTest(Values~System,data)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.3278 0.7241
##       21
```

We can compare the results with test of Bartlett

```
bartlett.test(Values~System,data)

##
##  Bartlett test of homogeneity of variances
##
## data:  Values by System
## Bartlett's K-squared = 0.10335, df = 2, p-value = 0.9496
```

And the conclusions are the same, the variances are not significantly different.

We can now go further with the Tukey Cramer Procedure

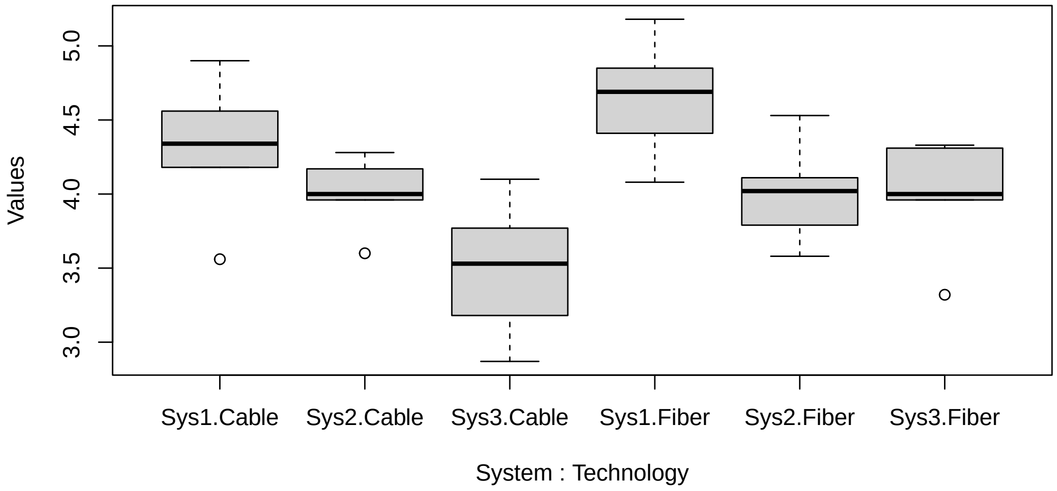
```
TukeyHSD(anova)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Values ~ System, data = data)
##
```

```
## $System
##                diff        lwr        upr     p adj
## Sys2-Sys1 -5.4625 -10.639823 -0.2851771 0.0374606
## Sys3-Sys1 -5.9125 -11.089823 -0.7351771 0.0234325
## Sys3-Sys2 -0.4500  -5.627323  4.7273229 0.9739203
```

2.4.2 Phase 2: Anova with two factors

After analyzing the data, the computer operations department team decided to also study the effect of the connection media used (cable or fiber). The team designed a study in which a total of 30 subscribers were chosen. The subscribers were randomly assigned to one of the three messaging systems so that there were five subscribers in each of the six combinations of the two factors messaging system and media used. Measurements were taken on the updated time.

```
# Import the data
data = read.csv("data/AMS11.2.csv", sep=";")
boxplot(Values~System+Technology,data)
```



Graphically, we can already see a significant difference between the different groups. We now propose to show this by statistical reasoning

1. Completely analyze these data and write a report to the team that indicates the importance of each of the two factors and/ or the interaction between them on the update time.

   We are going to see if there is a statistical difference between the three groups using an ANOVA. I will simply use the appropriate function to get the results.

```
anova <- aov(Values~System*Technology,data)
summary(anova)

##                   Df Sum Sq Mean Sq F value  Pr(>F)
## System             2  2.793  1.3963   8.217 0.00191 **
## Technology         1  0.577  0.5769   3.395 0.07780 .
## System:Technology  2  0.312  0.1561   0.918 0.41270
## Residuals         24  4.078  0.1699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   As it is shown by the previous computations, the system and the Technology have a significant impact on the the observed values. However, the interaction between the two variables has no significant impact on the data.
   By removing the interaction term from the ANOVA, we still observe the same results.

```
anova <- aov(Values~System+Technology,data)
summary(anova)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## System       2  2.793  1.3963   8.268 0.00166 **
## Technology   1  0.577  0.5769   3.416 0.07598 .
## Residuals   26  4.391  0.1689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   Because the samples have the same size in each group, there is no need to check the variance of each group (using the *Levene Test* or Bartlett one).

3 Lab 3

3.1 Volatility of a stock

The volatility of a stock is often measured by its beta value. You can estimate the beta value of a stock by developing a simple linear regression model, using the percentage weekly change in the stock as the dependent variable and the percentage weekly change in a market index as the independent variable. The $S\&P$ 500 Index is a common index to use. For example, if you wanted to estimate the beta value for Disney, a market model:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where Y represents the % of weekly change in company, X represents the % of weekly change in $S\&P$ 500 Index.

The least-squares regression estimate of the slope β_1 is the estimate of the beta value for the studied company. A stock with a beta value of 1.0 tends to move the same as the overall market. A stock with a beta value of 1.5 tends to move 50% more than the overall market, and a stock with a beta value of 0.6 tends to move only 60% as much as the overall market. Stocks with negative beta values tend to move in the opposite direction of the overall market. The following table gives some beta values for some widely held stocks as of July $11, 2015$:

Company	β_1
Apple	1.13
Disney	1.42
American Eagles Mines	-0.64
Marriott	1.5
Microsoft	0.78
Procter & Gamble	1.0

1. For each of the six companies, interpret the beta value.

   (a) The stock of Apple increases by a ratio of 113% with the overall market. So this market is more volatile than the overall mean.

   (b) The stock of Disney increases faster with a ratio of 42% with the overall market. So this market is more volatile than the overall mean.

   (c) The stock of American Eagles Mines decreases by a ratio of 46% with the overall market, so the trends of this stock moves in the opposite direction of the overall market. Maybe it means that this company has a lot of opponents in the global market and the increase of stock of all them results in a reduction for this particular company.

(d) The stock of Marriott increased by a ratio of 50% with the overall market. So this market is more volatile than the overall mean.

(e) Microsoft's stock moves 78% as much as the overall market and is considered less volatile as the market.

(f) Proctor Gamble's stock moves as much as the overall market and is as volatile as the market.

2. How can investors use the beta value as a guide for investing?

   It is enough to choose the company with the highest beta value. And this β value can also be associated to the volatility of the stock so as a measure of the risk.

3.2 Stock Prices

Refer to the discussion of beta values and market models. The S&P 500 Index tracks the overall movement of the stock market by considering the stock prices of 500 large corporations.
The following variables are included:

- WEEK—Week ending on date given
- SP—Weekly closing value for the SP 500 Index
- GE—Weekly closing stock price for General Electric
- DISCA—Weekly closing stock price for Discovery Communications
- GOOG—Weekly closing stock price for Google

```r
# Import the data
data = read.csv("data/StockPrices2014.csv", sep=";")
```

1. Estimate the market model for GE. (Hint: Use the percentage change in the S&P 500 Index as the independent variable and the percentage change in GE's stock price as the dependent variable.)

```
mymodel = lm(GE~SP,data)
summary(mymodel)

##
## Call:
## lm(formula = GE ~ SP, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70136 -0.44498 -0.02585  0.54620  1.60960
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.397113   2.268937  10.312 4.48e-14 ***
## SP           0.001351   0.001172   1.152    0.255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7034 on 51 degrees of freedom
## Multiple R-squared:  0.02536,Adjusted R-squared:  0.00625
## F-statistic: 1.327 on 1 and 51 DF,  p-value: 0.2547

coeff <- mymodel$coefficients
coeff

##  (Intercept)           SP
## 23.397112983  0.001350512
```

2. Interpret the beta value for GE.

   The estimated values of β_1 is close to 0 and the R^2 is really low, which mean that the relation between the two variables is really low. Note that same conclusion can be made using the p-value associated to the slope of our linear model (> 0.05) which means that the slope is not significantly different from 0. We cannot say that is reasonable to estimate the stock prices of GE using the S&P 500 Index. Other factors shall be tested and/or taken into account.

3. Repeat the two previous questions for Discovery Communications.

```
mymodel = lm(DISCA~SP,data)
summary(mymodel)

##
## Call:
## lm(formula = DISCA ~ SP, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.868 -2.146 -0.266  1.472  6.083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.23380    9.32778   8.923 5.41e-12 ***
## SP          -0.02264    0.00482  -4.697 2.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.892 on 51 degrees of freedom
## Multiple R-squared:  0.302,Adjusted R-squared:  0.2883
## F-statistic: 22.07 on 1 and 51 DF,  p-value: 2.028e-05

coeff <- mymodel$coefficients
coeff

## (Intercept)          SP
## 83.23380070 -0.02263982
```

For Discovery Communications, we can see that β_1 is still low but negative this time, which means that stock prices of DISCA are decreasing when the value of the Index tracks increases.
Furthermore, this behavior is significant since the associated p-value of β_1 is less than 0.055.
We can say that it has a sense to try to estimate the stock price of this company using this index for this particular company.
However, the R^2 remains low with a value of 0.28 which means that other information should be taken into account to have a better estimation. It can also means that the relation between stock price and the Index is not necessarily linear.

4. Repeat the two previous questions for Google.

```
mymodel = lm(Google~SP,data)
summary(mymodel)

##
## Call:
## lm(formula = Google ~ SP, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.725 -17.919   1.279  21.241  42.350
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 655.70137   84.68993   7.742 3.68e-10 ***
## SP           -0.04609    0.04376  -1.053    0.297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.26 on 51 degrees of freedom
## Multiple R-squared:  0.02129,Adjusted R-squared:  0.002095
## F-statistic: 1.109 on 1 and 51 DF,  p-value: 0.2972

coeff <- mymodel$coefficients
coeff

##  (Intercept)          SP
## 655.70136557  -0.04608582
```

When it comes to Google, we can make the same conclusion as the one drawn for GE. Note that, this time, the coefficients is negative while it was positive for GE.

5. Write a brief summary of your findings.

We can note that for two of the three studied company, trying to estimate the stock price with a linear model using only the given index is not significant; both slopes were not significantly different from 0 (because the p-values were lower than 0.05). However, note that for the tree studied companies, the linear model using the S&P 500 Index was not a good candidate to explain the stock prices for the different company.
To have a better prediction or estimation of the stock price, we can:

- try a non linear model

- use other variables

3.3 Ashland MultiComm Services

To ensure that as many trial subscriptions to the *3-For-All* service as possible are converted to regular subscriptions, the marketing department works closely with the customer support department to accomplish a smooth initial process for the trial subscription customers.

To assist in this effort, the marketing department needs to accurately forecast the monthly total of new regular subscriptions.

A team consisting of managers from the marketing and customer support departments was convened to develop a better method of forecasting new subscriptions. Previously, after examining new subscription data for the prior three months, a group of three managers would develop a subjective forecast of the number of new subscriptions.

Livia Salvador, who was recently hired by the company to provide expertise in quantitative forecasting methods, suggested that the department look for factors that might help in predicting new subscriptions. Members of the team found that the forecasts in the past year had been particularly inaccurate because in some months, much more time was spent on telemarketing than in other months.

Livia collected data for the number of new subscriptions and hours spent on telemarketing for each month for the past two years.

```r
# Import the data
data = read.csv2("data/AMS13.csv")
```

1. What criticism can you make concerning the method of forecasting that involved taking the new subscriptions data for the prior three months as the basis for future projections?

   The problem with a model that only takes into account the previous three months, in this case, is that it will not take into account any seasonal effects that may occur during the year.

   These are phenomena that are interesting to take into account in the modeling. For example, we know that the number of registrations in a gym increases strongly after the summer vacations or after the Christmas period. Phenomena that will not be studied if we learn our model on a much too restricted time space

2. What factors other than number of telemarketing hours spent might be useful in predicting the number of new subscriptions? Explain.

   As we have said, this can also depends on the period of the year if we take the example of our fitness club. The different offers that are made, the potential gifts for new subscribers.
   Furthermore, rather than the number of hours spent in telemarketing, we can also use the investment in different marketing strategies to attract new subscribers as an independent variable.

3. (a) Analyze the data and develop a regression model to predict the number of new subscriptions for a month, based on the number of hours spent on telemarketing for new subscriptions.

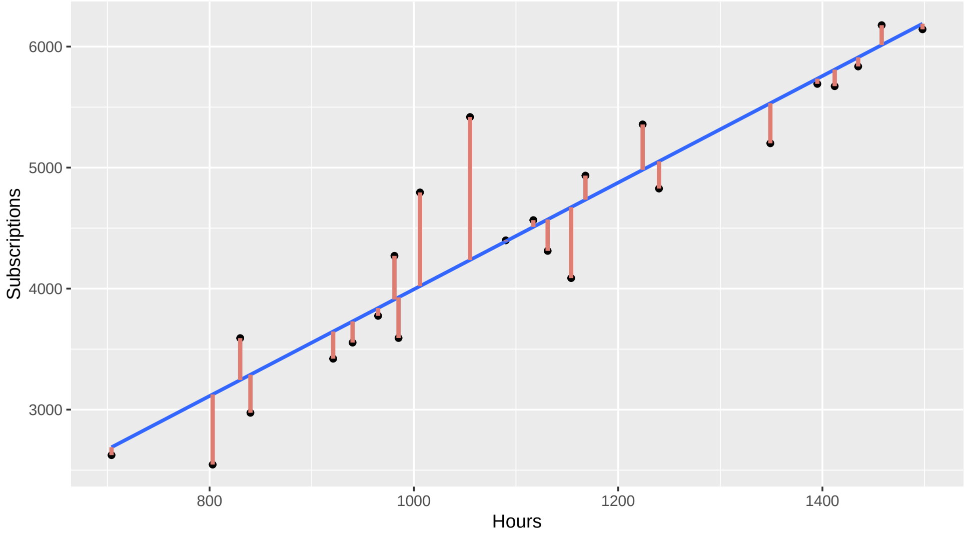
   Our model takes the following form

   $$Y = \beta_0 + \beta_1 X,$$

   where Y is the number of new subscriptions and X represents the number of hours.

```
mymodel = lm(Subscriptions~Hours,data)
summary(mymodel)

##
## Call:
## lm(formula = Subscriptions ~ Hours, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -585.95 -233.67  -65.11  172.85 1181.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -413.8204   427.0878  -0.969    0.343
## Hours          4.4080     0.3765  11.709 6.36e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 409.8 on 22 degrees of freedom
## Multiple R-squared:  0.8617,Adjusted R-squared:  0.8554
## F-statistic: 137.1 on 1 and 22 DF,  p-value: 6.358e-11
```

```
coeff <- mymodel$coefficients

# Graphical Representation of the Model and Residuals

library(ggplot2)
ggplot(data, aes(x=Hours, y=Subscriptions)) +
geom_point() +
geom_smooth(method=lm,se=FALSE) +
geom_segment(aes(x = Hours, y = Subscriptions, xend = Hours,
                 yend = coeff[1] + coeff[2]*Hours,
                 col = "Residuals"),
             col = "#DF7D72", lwd= 1.2, data = data)
```



(b) If you expect to spend 1,200 hours on telemarketing per month, estimate the number of new subscriptions for the month. Indicate the assumptions on which this prediction is based. Do you think these assumptions are valid? Explain.

We have to apply the previous model using the X value 1,200. We have found

$$\beta_= - 413.82 \quad \text{and} \quad \beta_1 = 4.41.$$

Thus this estimated value of new subscriptions is equal to

```
# Estimated number of new subscriptions
coeff[1] +  coeff[2]*1200
```

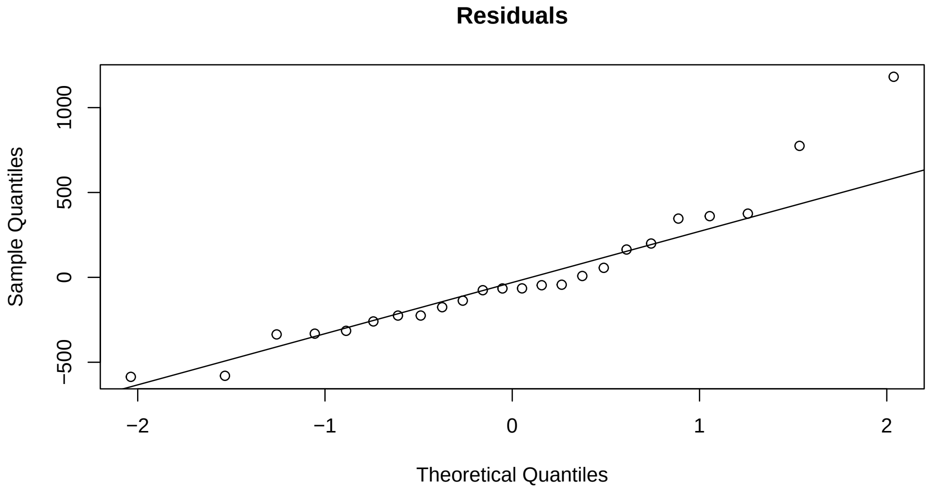
```
## (Intercept)
##     4875.72
```

This estimation is based on the fact the number of new subscriptions only depends on the number hours spent by the telemarketing team. This model is maybe to simple to be true or realistic.
We can also cite other assumptions that required to apply such model, as the fact that our data are normally distributed.

We will try to check the different assumptions:

```
# Quantile-quantile plot
qqnorm(mymodel$residuals, main="Residuals")
qqline(mymodel$residuals)
```

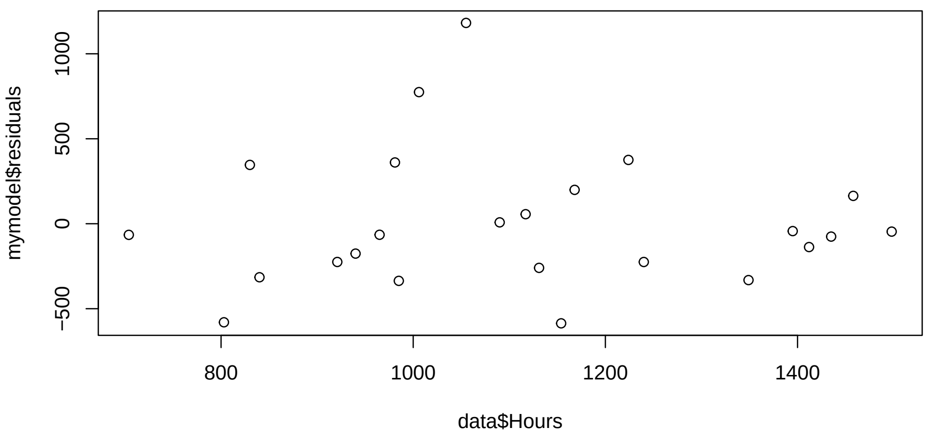
Normality of the residuals



Except for two examples, the residuals can be considered as normally distributed, but it is sure that a statistical test will reject the normality assumption.

```
        plot(data$Hours,mymodel$residuals)
```

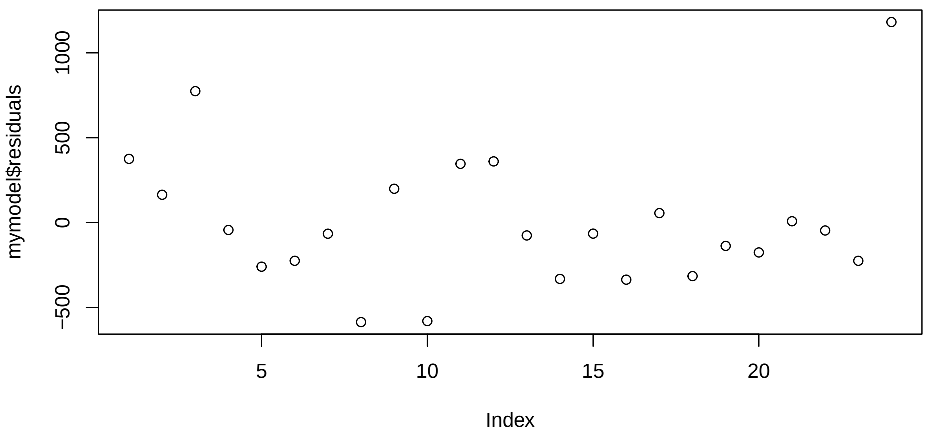
Homoscedasticity



Here we can say that the residuals have equal variance if they are normally distributed.

```r
plot(mymodel$residuals)
```

Independence



The residuals can be considered are also independent.

(c) What would be the danger of predicting the number of new subscriptions for a month in which 2,000 hours were spent on telemarketing?

Hypothetically, this value would be equal to

```
coeff[1] +  coeff[2]*2000

## (Intercept)
##    8402.081
```

The observed values for the construction of the model are in the interval $[1200, 1400]$ and the value for which we want to make an estimate is outside the observed values.
We therefore have no guarantee that our linear model is always valid for values outside the range of observed values for its construction.

3.4 Digital Case

Leasing agents from the Triangle Mall Management Corporation have suggested that Sunflowers consider several locations in some of Triangle's newly renovated lifestyle malls that cater to shoppers with higher-than-mean disposable income. Although the locations are smaller than the typical Sunflowers location, the leasing agents argue that higher-than-mean disposable income in the surrounding community is a better predictor of higher sales than profiled customers. The leasing agents maintain that sample data from 14 Sunflowers stores prove that this is true.

Let us perform our linear regression

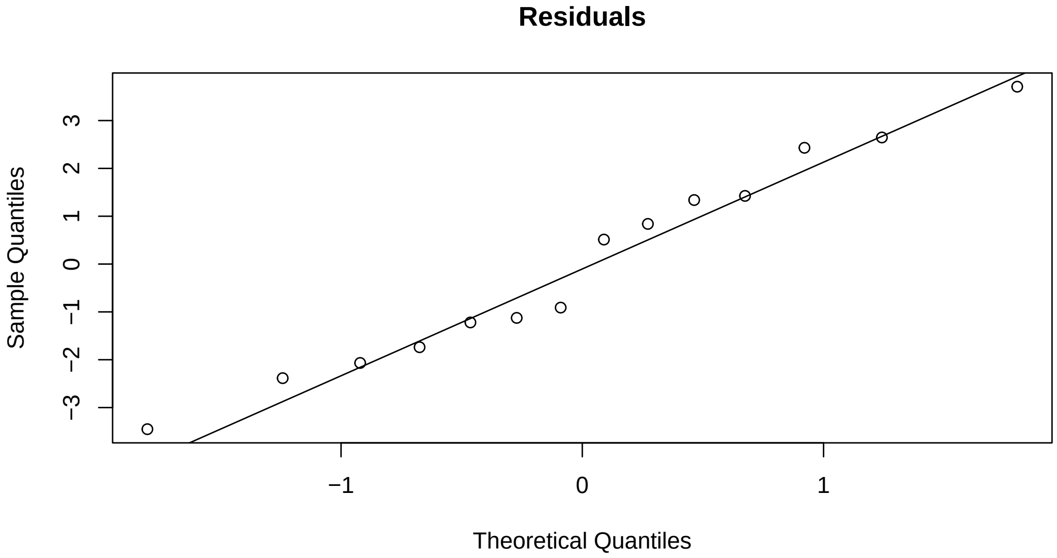
```
# Import the data
data = read.csv("data/TriangleProposal.csv", sep=";")

# Estimate the model
mymodel = lm(Annuale_Sales~Disposable,data)
summary(mymodel)

##
## Call:
## lm(formula = Annuale_Sales ~ Disposable, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4518 -1.6089 -0.1991  1.4032  3.7079
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.94122    2.37999  -0.816  0.43060
## Disposable   0.19295    0.05711   3.379  0.00548 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.235 on 12 degrees of freedom
## Multiple R-squared:  0.4875,Adjusted R-squared:  0.4448
## F-statistic: 11.42 on 1 and 12 DF,  p-value: 0.005481
```

We will try to check the different assumptions:

```r
# Quantile-quantile plot
qqnorm(mymodel$residuals, main="Residuals")
qqline(mymodel$residuals)
```

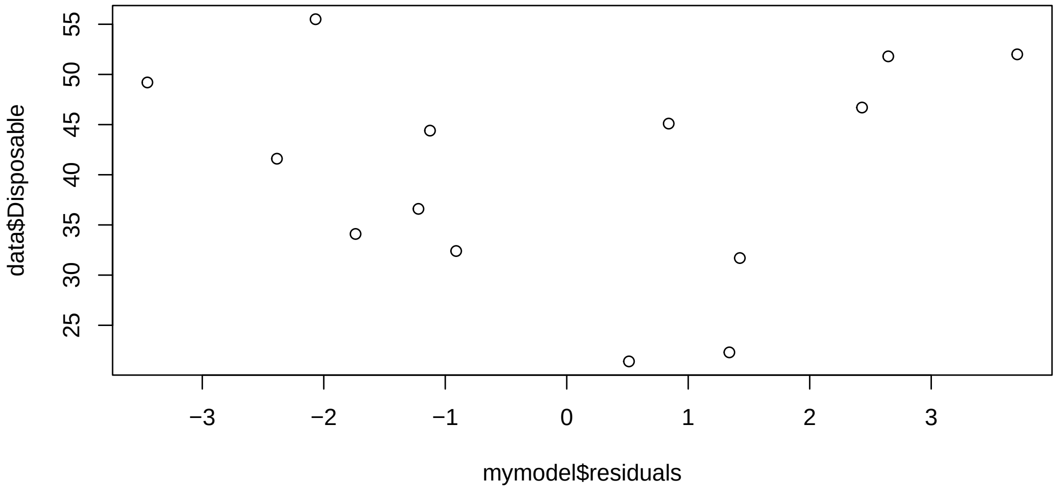
Normality of the residuals



The residuals can be considered as normally distributed.

```
plot(mymodel$residuals,data$Disposable)
```

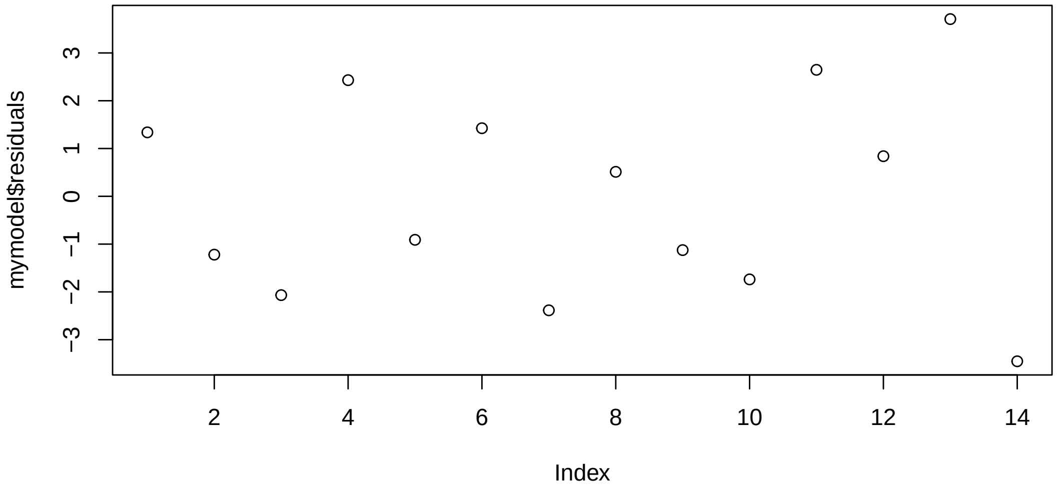
Homoscedasticity



The residuals have equal variance.

```
plot(mymodel$residuals)
```

Independence



The residuals can be considered are also independent.

1. Should mean disposable income be used to predict sales based on the sample of 14 Sunflowers stores?

   The results show that we can use the mean disposable to predict sales, due to the correlation between the two variables. However, this correlation is not very strong (look at the R^2 value). In their report, they also indicate that they want to make new predictions of their sales where the mean disposable is not in the range of current observed values. This point cannot be done, and we cannot be sure that the results are trustworthy. they shall extend their study first in areas of "exceptional affluence".

2. Should the management of Sunflowers accept the claims of Triangle's leasing agents? Why or why not?

   According to the report and the regression analysis, clearly not! The correlation between the two studied variables is clearly to weak and maybe other factors shall be taken into account for their prediction sales as the store size for instance of their location.

3. Is it possible that the mean disposable income of the surrounding area is not an important factor in leasing new locations? Explain.

That might be a possibility since the correlation between the two studied variables is weak. However, opening a new retail location would be based on a number of factors (some of these factors such as competitive retail analysis, demographic and geographic profiles, regional economic analysis, and sales potential forecast analysis, store size and the locations).

4. Are there any other factors not mentioned by the leasing agents that might be relevant to the store leasing decision?

   It is also important to study if the proposed products are well suited for the population in the area.

4 Lab 4

4.1 Checking your Understanding

1. What is the difference between R^2 and the adjusted R^2?

   The R^2 measures the proportion of variance in the data that is explained by the model, i.e. by the different independent variables. However, this criterion does not allow to compare models with a different number of variables in an equal way. The adjusted R^2 modifies the value of the R^2 taking into account the number of variables used to build the model.

   $$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

   where p is the number of dependent variables in the model.

2. How does the interpretation of the regression coefficients differ in multiple and simple linear regression?

   In simple linear regression, the coefficient β_1 allows us to directly translate the link between the independent variable and the dependent variable, it thus informs us about the slope of the line which links these two variables, from a geometrical point of view.

   In a multiple regression, this interpretation remains the same provided that the values of the other variables are fixed. In other words, the value β_i gives information on the impact of the independent variable X_i when all the variables X_j with $j \neq i$ are fixed on the same level.
   Otherwise, in general, it is important to take into account all the variables to determine the final impact on

3. How does testing the significance of the entire multiple regression model differ from testing the contribution of each independent variable?

   In multiple linear regression, when you want to test the significance of the model:

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   we formulate the following assumptions:

   $$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p \quad \text{v.s.} \quad H_1 : \exists k, \text{ s.t.} \beta_k \neq 0.$$

The test is then based on the **Fisher Distribution** using the following statistic:

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}},$$

where SSR denotes the *part of the variation* explained by the regression and SSE denotes the *part of the variation* not explained by the model, *i.e.* the residual variation.

If you want to test the significance of a given coefficient β_p, you formulate the following assumptions:

$$H_0 : \beta_k = 0 \quad \text{v.s.} \quad H_1 : \beta_k \neq 0.$$

The test is then based on the **Student Distribution** using the following statistic:

$$T_k = \frac{\hat{\beta}_k - \beta_k}{S_{\hat{\beta}_k}}.$$

4. How do the coefficients of partial determination differ from the coefficient of multiple determination?

   R^2 can be interpreted as the percentage of variance in the dependent variable that can be explained by the predictors.
   The coefficient of multiple determination (the multiple R^2) is defined as the squared root of the R^2, it represents the correlation between the **dependent** variable and the **linear combination of the independent variable**, *i.e.* it can be seen as the correlation between Y and the predicted values $\hat{Y}$.

5. Why and how do you use dummy variables?

   We are using dummy variables when we have to deal with categorical variables in our (multiple) linear model.

   When we have a categorical variable with only two modalities, for instance "M" and "F", we choose one category as a reference, for instance "M" and we replace this value by 0 and the second category will take the value 1

   When we have more than two modalities, let us say $k > 2$ we have to create $k-1$ variables which will be used to denote if an instance belong to one of the k groups, the last group, which is not represented, is considered as the baseline to which the other situations are compared.

6. How can you evaluate whether the slope of the dependent variable with an independent variable is the same for each level of the dummy variable?

   We have to see if the interaction term is significant or not. First you can check if the associated parameter is significant or not (based on the T-test). Then you can use the partial F-test to whether adding this new term can significantly improve the model's prediction.

7. Under what circumstances do you include an interaction term in a regression model?

   The interaction term is included when you are dealing with categorical variables in your model. It will give you the opportunity the see if the slope is the same between the two studied groups in your linear regression.

8. When a dummy variable is included in a regression model that has one numerical independent variable, what assumption do you need to make concerning the slope between the dependent variable, Y, and the numerical independent variable, X?

   We need to check if the slope is significant or not?

9. When do you use logistic regression?

   The logistic regression is used when the dependent variable Y is no more **a numerical variable but a categorical** one.
   In linear regression, we assume the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon_i,$$

   where Y is the dependent and real random variable, β_j are the parameters of the model, X_j are the independent random variables and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ represent the noise in the data. When it comes to the logistic regression (binary logistic regression), our dependent variable can take two values, *i.e.* $Y \in \{\pm 1\}$, and the model is expressed as follows

$$\mathbb{P}[Y = 1 \mid X] = (1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p))),$$

   and:

$$Y = \begin{cases} 1 & \text{if} \quad \mathbb{P}[Y = 1 \mid X] > 0.5, \\ 0 & \text{otherwise} \end{cases}.$$

In other words, we are directly dealing with the value of Y, but we are trying to estimate the probability that our data belongs to the class 1, *i.e.* to a group of reference.

10. What is the difference between least squares regression and logistic regression.

Some details we provided in the previous question, but we will now focus on the quantity we aim to minimize.
In least sqare regression, we minimiser the *Mean square Error* (MSE), *i.e.* the mean quadratic distance between the **predicted value** $\hat{y}$ by the model and the **observed one** y:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 .$$

In logistic regression it has no more sense since the dependent variable can only take two values. So instead of minimizing the previous error, we rather maximize the likelihood of our data (more precisely, the log-likelihood):

$$\sum_{i=1}^{m} y_i \ln\left(\frac{1}{1+\exp\left(-(\langle \boldsymbol{\beta}, \mathbf{x}_i\rangle)\right)}\right) + (1-y_i)\ln\left(1 - \frac{1}{1+\exp\left(-(\langle \boldsymbol{\beta}, \mathbf{x}_i\rangle)\right)}\right).$$

4.2 Managing Ashland MultiComm Services

In its continuing study of the 3-For-All subscription solicitation process, a marketing department team wants to test the effects of two types of structured sales presentations (personal formal and personal informal) and the number of hours spent on telemarketing on the number of new subscriptions. The staff has recorded these data for the past 24 weeks.

Analyze these data and develop a multiple regression model to predict the number of new subscriptions for a week, based on the number of hours spent on telemarketing and the sales presentation type. Write a report, giving detailed findings concerning the regression model used.

First of all, we are going to deal with the categorical variable and we are transforming this variable into a numerical one. to do this, we set the value 1 when the level is equal to *Formal*, and 0 otherwise.

```
# Import the data
data = read.csv("data/AMS_14.csv", sep=";")

# Binarization of the categorical variable
data$Formal = ifelse(data$Presentation == "Formal", 1, 0)
data = data[,c("New.Subscriptions","Hours","Formal")]
```

We can now build our model and try to analyze it:

```
# Multiple Linear Model
mymodel = lm(New.Subscriptions~.,data)
summary(mymodel)

##
## Call:
## lm(formula = New.Subscriptions ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -111.22  -50.50  -13.69   51.31  155.18
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -277.0791    88.9423  -3.115  0.00524 **
## Hours          4.8646     0.3119  15.597 5.06e-13 ***
## Formal        96.3074    32.4011   2.972  0.00727 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.8 on 21 degrees of freedom
## Multiple R-squared:  0.9214,Adjusted R-squared:  0.9139
## F-statistic: 123.1 on 2 and 21 DF,  p-value: 2.529e-12
```

This first output tries to provide first results on the linear which contains 2 dependent variables, a numerical one and a categorical one. Let us make some remarks and conclusions on this first analysis:

1. The intercept of the mode, β_0, is significantly different from 0.
2. The parameter β_1 associated to the variable *Hours* is also different from 0, which means that the variable *Hours* is correlated or explains a part of the variance in

the data.
The model also tells us that each hour of work would bring in about 4 *to* 5 *new registrants.*

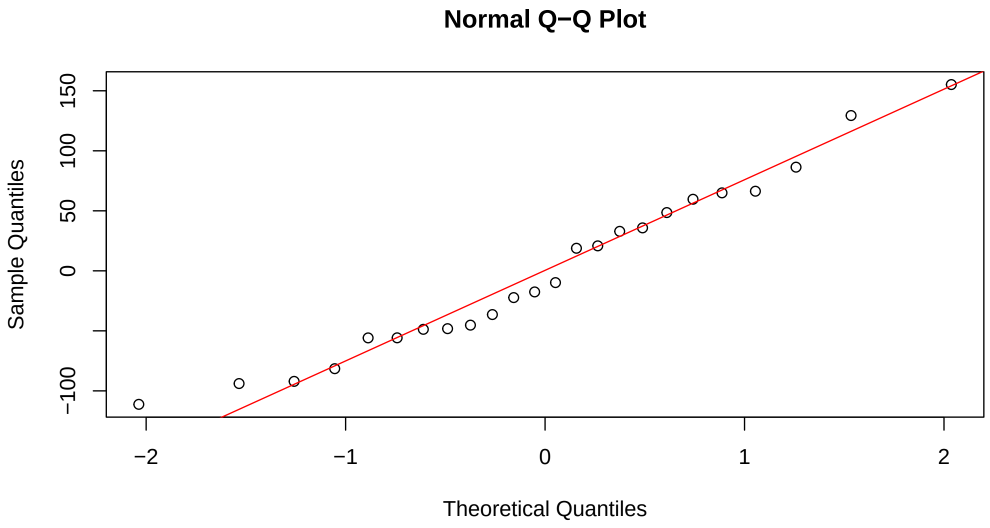
3. The parameter β_2 is also different from 0 so we can draw the same conclusion when it comes to the categorical variable.
   In this case, we can say that formal presentations have a significant impact on the number of new registrants, we can even say that a formal presentation can generate 100 *new registrants.*

4. Overall, our model is statistically significant as the p-value of $2.53e^{-12}$ is showing to us, *i.e.* all the parameters are not simultaneously equal to 0. This is consistent with the fact that the R^2 of the model is very high, so the model is able to explain much of the variance present in the data.

Before making the predictions, we are going to plot the residuals in order to check the following assumptions:

- Normality

```r
qqnorm(mymodel$residuals)
qqline(mymodel$residuals, col ="red")
```



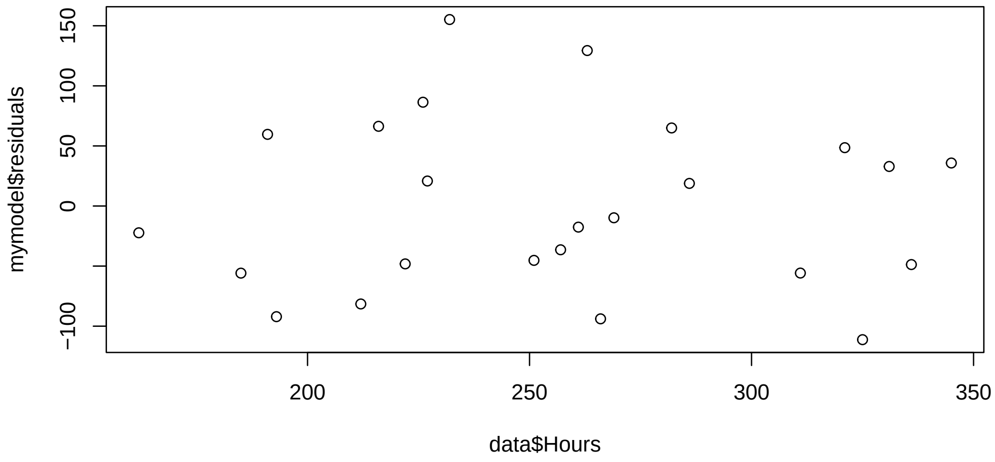
```r
# For the sake of completeness
shapiro.test(mymodel$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mymodel$residuals
## W = 0.96391, p-value = 0.5217
```

Residuals are normally distributed as shown by the graph and the appropriated statistical test.

- Homoscedasticity

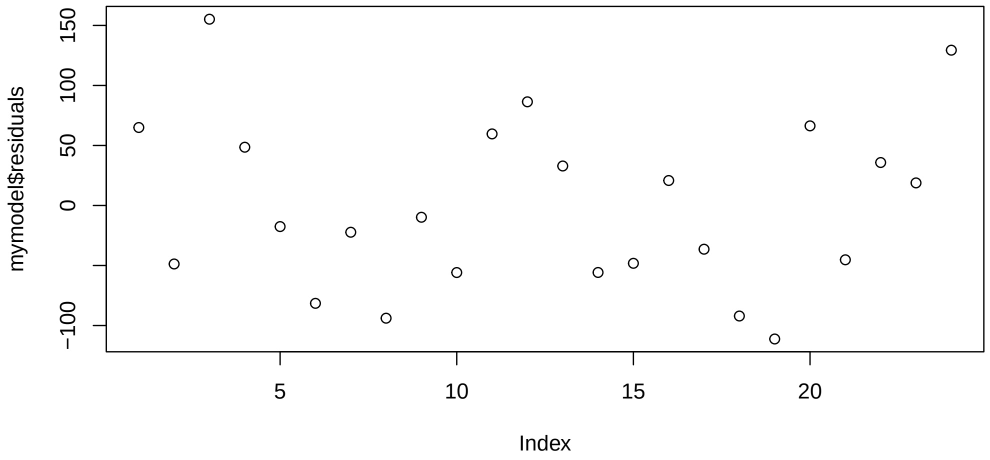
```
plot(data$Hours,mymodel$residuals)
```



The assumption of homoscedasticity is fulfilled.

- Independency

```
plot(mymodel$residuals)
```



The residuals are independent, and we are going to see that is no positive auto-correlation.

```
durbinWatsonTest(mymodel)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.0955251       1.63056   0.386
##  Alternative hypothesis: rho != 0
```

There is no positive auto-correlation.

There is no need, in this case to compute any partial R^2 or contribution of the the independent variables.

We will focus on the study of an interaction term, so we will consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where X_1 represents the numerical variable ("*Hours*") and X_2 represents the dummy variable ("*Formal*").
First, let us perform the linear regression

```
# We will create the new variable
#data$Interaction = data$Hours*data$Formal

# Multiple Linear Model with Interaction Term
mymodel_interaction = lm(New.Subscriptions~ Hours*Formal ,data)
summary(mymodel_interaction)


##
## Call:
## lm(formula = New.Subscriptions ~ Hours * Formal, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.769  -46.089   -9.757   40.759  152.857
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -341.0273   132.1268  -2.581   0.0178 *
## Hours           5.0972     0.4727  10.784 8.76e-10 ***
## Formal        206.2773   169.3167   1.218   0.2373
## Hours:Formal   -0.4210     0.6358  -0.662   0.5155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.82 on 20 degrees of freedom
## Multiple R-squared:  0.9231,Adjusted R-squared:  0.9115
## F-statistic: 79.99 on 3 and 20 DF,  p-value: 2.591e-11
```

Here it seems that the interaction term is not significant so there s no need to study its contribution.