Regression Analysis

Simple Linear Regression Model

In this topic we assume that $\mathbb{D}X < \infty$ and $\mathbb{D}Y < \infty$.

Regression analysis study the form of the relationship between two numerical random variables X and Y. More precisely its aim is by knowing X and the regression model to predict Y.

X is called **independent variable (or predictor) /независима променлива/**.

Y is called **dependent (or outcome) variable /зависима променлива/**.

When there is a single dependent variable and a single independent variable, and the dependence on the coefficients is linear the analysis is called **a simple linear regression analysis /проста линейна регресия/**. More precisely **the simple linear regression model** is

$$Y = \hat{Y} + \varepsilon = \beta_0 + \beta_1 X + \varepsilon, \text{ where}$$

- **ε is the random error term /случайна грешка/.**

$$\varepsilon = Y - \hat{Y} = Y - \beta_0 - \beta_1 X.$$

- β_0, β_1 are unknown coefficients. They will be estimated from the data by using **the method of least squares /метода на най-малките квадрати/** (by minimizing the sum of square errors $\sum_{i=1}^{n} \varepsilon_i^2$).

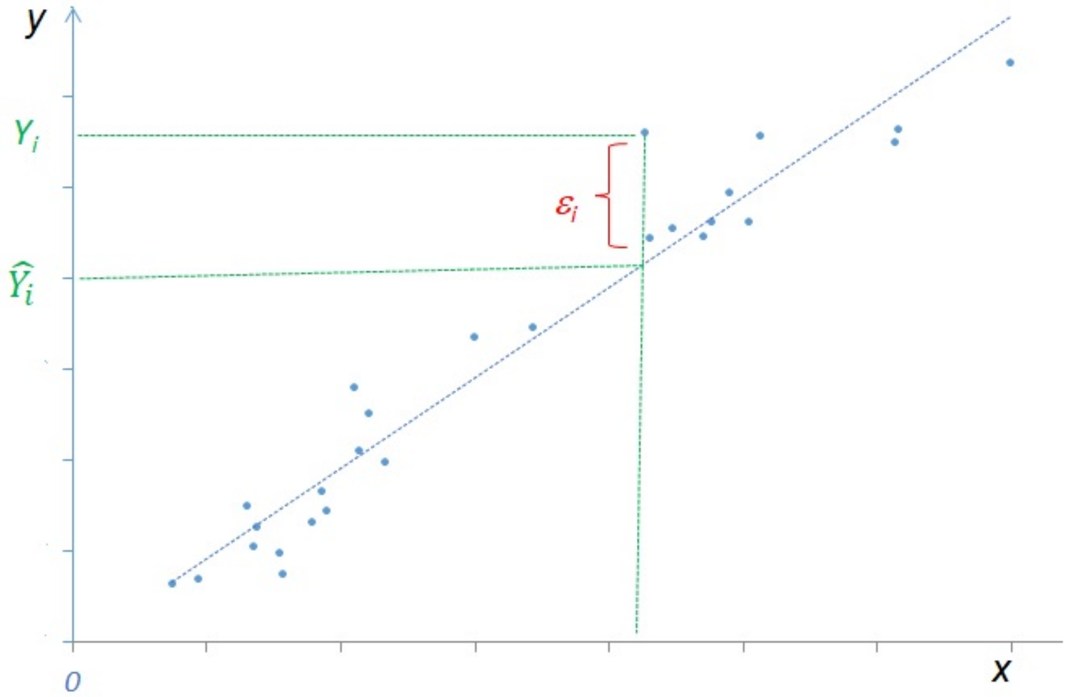
By assumptions:
- $\mathbb{E}\varepsilon = 0$ (and therefore $\mathbb{E}Y = \beta_0 + \beta_1 \mathbb{E}X$)
- $\mathrm{cor}(X, \varepsilon) = 0$ i.e. the independent variable X and the random error term ε are uncorrelated.

Therefore, $\hat{Y}$ and ε are uncorrelated and

$$\mathbb{E}(\varepsilon \,|\, X) = \mathbb{E}\varepsilon = 0,\ \mathbb{D}(\varepsilon \,|\, X) = \mathbb{D}\varepsilon = \sigma_\varepsilon^2.$$
$$\hat{Y} = \mathbb{E}(Y \,|\, X) = \beta_0 + \beta_1 X$$

and the corresponding **simple linear regression equation** (the equation of the corresponding straight line) is as follows:

$$y = \beta_0 + \beta_1 x$$



By the model assumed it is easy to see that:

- $\beta_0 = \mathbb{E}(Y\,|\,X=0)$ is the intercept of the regression line from Oy axis.
- $\beta_1 = \mathbb{E}(Y\,|\,X+1) - \mathbb{E}(Y\,|\,X) = \beta_0 + \beta_1(X+1) - \beta_0 - \beta_1 X$ is the slope - the expected increment of the Y (in its units) when X increases with 1 (in the units of X).

For $\beta_1 > 0$, when X increases Y gets bigger.
For $\beta_1 < 0$, when X increases Y gets smaller.
For $\beta_0 = 0$, X does not influence Y.

When we consider the variances $\mathbb{D}(\hat{Y}) = \mathbb{D}(\beta_0 + \beta_1 X) = \beta_1^2 \mathbb{D}X$,
$\mathbb{D}(Y) = \mathbb{D}(\hat{Y} + \varepsilon) = \mathbb{D}(\hat{Y}) + \mathbb{D}\varepsilon = \beta_1^2 \mathbb{D}X + \sigma_\varepsilon^2$

$$\operatorname{cor}^2(Y,\hat{Y}) = \frac{\operatorname{cov}^2(Y,\hat{Y})}{\mathbb{D}(Y)\mathbb{D}(\hat{Y})} = \frac{\operatorname{cov}^2(Y,\beta_0+\beta_1 X)}{\mathbb{D}Y(\beta_1^2 DX)} = \frac{(\operatorname{cov}(\mathrm{Y},\beta_1\mathrm{X}))^2}{\mathbb{D}Y\beta_1^2\mathbb{D}X} =$$

$$= \frac{(\beta_1 \operatorname{cov}(Y,X))^2}{\mathbb{D}Y\beta_1^2\mathbb{D}X} = \frac{\operatorname{cov}^2(Y,X)}{\mathbb{D}Y\mathbb{D}X} = \operatorname{cor}^2(Y,X)$$

Moreover,

$$\operatorname{cor}^2(Y,\hat{Y}) = \frac{\operatorname{cov}^2(Y,\hat{Y})}{\mathbb{D}Y\mathbb{D}\hat{Y}} = \frac{\operatorname{cov}^2(\hat{Y}+\varepsilon,\hat{Y})}{\mathbb{D}Y\mathbb{D}\hat{Y}} = \frac{(\operatorname{cov}(\hat{Y},\hat{Y}) + \operatorname{cov}(\varepsilon,\hat{Y}))^2}{\mathbb{D}Y\mathbb{D}\hat{Y}} =$$

$$= \frac{(\mathbb{D}\hat{Y}+0)^2}{\mathbb{D}Y\mathbb{D}\hat{Y}} = \frac{\mathbb{D}\hat{Y}}{\mathbb{D}Y} = 1 - \frac{\mathbb{D}\varepsilon}{\mathbb{D}Y} = \beta_1^2\frac{\mathbb{D}X}{\mathbb{D}Y}$$

It can be shown that the minimal value of the mean square error between Y and $\hat{Y}$ can be obtained for

$$\beta_0 = \mathbb{E}X - \beta_1\mathbb{E}X \quad \text{(see the assumptions for } \varepsilon\text{)}$$

$$\beta_1 = \sqrt{\frac{\mathbb{D}Y}{\mathbb{D}X}}\operatorname{cor}(X, Y) = \frac{\operatorname{cor}(X, Y\sqrt{\mathbb{D}X\mathbb{D}Y})}{\mathbb{D}X} = \frac{\operatorname{cov}(X, Y)}{\mathbb{D}X}$$

The corresponding estimators of $\mathbb{E}Y$, $\mathbb{E}X$, $\mathbb{D}Y$, $\mathbb{D}X$, $\operatorname{cov}(X, Y)$ and $\operatorname{cor}(\mathrm{X}, \mathrm{Y})$ are already known. Therefore, we can estimate β_0 and β_1.

When we use these coefficients, the minimal value of the **Residual Standard Error** of the model is (between Y and $\hat{Y}$) is

$$\sigma_\varepsilon = \sqrt{\mathbb{D}\varepsilon} = \sqrt{\mathbb{E}\varepsilon^2} = \sqrt{\mathbb{E}(Y - \hat{Y}^2)} =$$
$$= \sqrt{\mathbb{E}(Y - \beta_0 - \beta_1 X)^2} = \sqrt{\mathbb{D}Y(1 - \operatorname{cor}^2(X, Y))}$$

The coefficient

$$\operatorname{cor}^2(X, Y) = 1 - \frac{\mathbb{D}\varepsilon}{\mathbb{D}Y}$$

(and the corresponding estimator R^2) is called **coefficient of determination / коефициент на определеност/**. And, as far as and

$$\mathbb{D}Y = \mathbb{D}Y\operatorname{cor}^2(X, Y) + \mathbb{D}\varepsilon = \mathbb{D}\hat{Y} + \mathbb{D}\varepsilon$$

$\operatorname{cor}^2(X, Y)$ shows what part of $\mathbb{D}Y$ which is due to regression.

$1 - \operatorname{cor}^2(X, Y)$ is called **coefficient of indetermination /коефициент на неопределеност/**. It shows part of $\mathbb{D}Y$ is due to changes of the error term, i.e. variables that are not considered in the model.

The inequality

$$\mathbb{D}(Y \mid X = x) = \mathbb{D}(\beta_0 + \beta_1 X + \varepsilon \mid X = x) = \sigma_\varepsilon^2 \le \beta_1^2\mathbb{D}X + \sigma_\varepsilon^2 = \mathbb{D}Y$$

means that the information for X can help us to improve the estimation for Y as far as by using X we will obtain shorter confidence intervals for $\mathbb{E}(Y \mid X)$.

Suppose that we have i.i.d. observations on the random vector (X, Y). As far as the last means that

$$Y_i = \hat{Y} + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

we actually assume that

$$\operatorname{cor}(\varepsilon_i, \varepsilon_j) = 0,\ 1 \le i < j \le n$$

and

$$\mathbb{D}\varepsilon_i = \sigma_\varepsilon^2,\ i = 1, 2, \ldots, n$$

The last requirement is called **homoscedasticity /хомоскедастичност/**. If we have different variances of the error terms we speak about **heteroscedasticity / хетероскедастичност/**.

The corresponding Estimator of the **Residual Standard Error (RSE) /Стандартна грешка на остатъците/** is

$$\hat{\sigma}_\varepsilon = RSE = S_\varepsilon = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n-r-1}} = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n-2}},$$

where r is the last subindex of the unknown coefficients β_0, β_1 in the regression line. In this case $r = 1$. Or in the denominator we have the sample size minus the number of the unknown coefficients. Usually S_ε^2 is called **Mean Square Error (MSE)** of the model and we use the following notations

$$SSE = \sum_{i=1}^n \varepsilon_i^2,\ MSE = \frac{SSE}{n-r-1} = RSE^2 = S_\varepsilon^2$$

MSE is unbiased estimator of σ_ε^2.

The most important case of these models is when the errors ε_i are i.i.d $\varepsilon_i \in N(0,\sigma_\varepsilon^2)$. Then,

$$(Y\,|\,X = x) = \beta_0 + \beta_1 X + \varepsilon\,|\,X = x) \in$$
$$\in N\left(\beta_0 + \beta_1 x = \mathbb{E}Y + \operatorname{cor(X, Y)}\sqrt{\frac{\mathbb{D}Y}{\mathbb{D}X}}(x - \mathbb{E}X);\ \sigma_\varepsilon^2 = \mathbb{D}Y(1 - \operatorname{cor}^2(X, Y))\right)$$

and knowing X, β_0 and β_1 we can construct confidence interval for the expected value $\mathbb{E}(Y\,|\,X = x)$

Let us note that another simple linear regression models are:

✓ $Y = \beta_0 + \beta_1 f(X) + \varepsilon$;
✓$g(Y) = \beta_0 + \beta_1 X + \varepsilon$.

We work with them via substitutions.

Let us now explain briefly **the method of least squares /метод на най-малките квадрати/** which is the best way to estimate the coefficients. We are looking for two constants

$$(\beta_0, \beta_1) = \arg\min\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = \arg\min\left(\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\right) = \arg\min\left(\sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2\right)$$

The solution is obtained when we solve the system of equations

$$\begin{cases} \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (b_0 + b_1 X_i - Y_i)^2 & = 0 \\ \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (b_0 + b_1 X_i - Y_i)^2 & = 0 \end{cases}$$

$$\begin{cases} 2\sum_{i=1}^{n} (b_0 + b_1 X_i - Y_i) & = 0 \\ 2\sum_{i=1}^{n} (b_0 + b_1 X_i - Y_i) X_i & = 0 \end{cases}$$

$$\begin{cases} n b_0 + b_1 \sum_{i=1}^{n} X_i & = \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 & = \sum_{i=1}^{n} X_i Y_i \end{cases}$$

$$\begin{cases} b_0 + b_1 \overline{X}_n & = \overline{Y}_n \\ b_0 \overline{X}_n + b_1 \frac{\sum_{i=1}^{n} X_i^2}{n} & = \frac{\sum_{i=1}^{n} X_i Y_i}{n} \end{cases}$$

Intercept:

$$b_0 = \overline{Y}_n - b_1 \overline{X}_n$$

Slope:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_1 - \overline{Y}_n)}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}$$

It can be performed by using the function lm in R.

Example 1.
In order to investigate the dependence of the maximum heart rate of a person from the age, the maximum heart rate and the age of 15 people of different ages are observed. The results are as follows

```
> Age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
> MaxRate <- c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183, 178)
```

a. Build the simple linear regression model.

b. Estimate the coefficients and plot the regression line on the figure with bivariate distribution of the data.
c. Determine the expected maximum heart rate for any of these persons.
d. Determine the expected maximum heart rate for persons at age 30, 40, 50.
e. Determine the errors(residuals).
f. Determine the mean square error of the model and the residual standard error.
g. Compute the coefficient of deteremination.
h. Check if $\mathbb{E}\varepsilon = 0$
i. Check if the errors are normal.

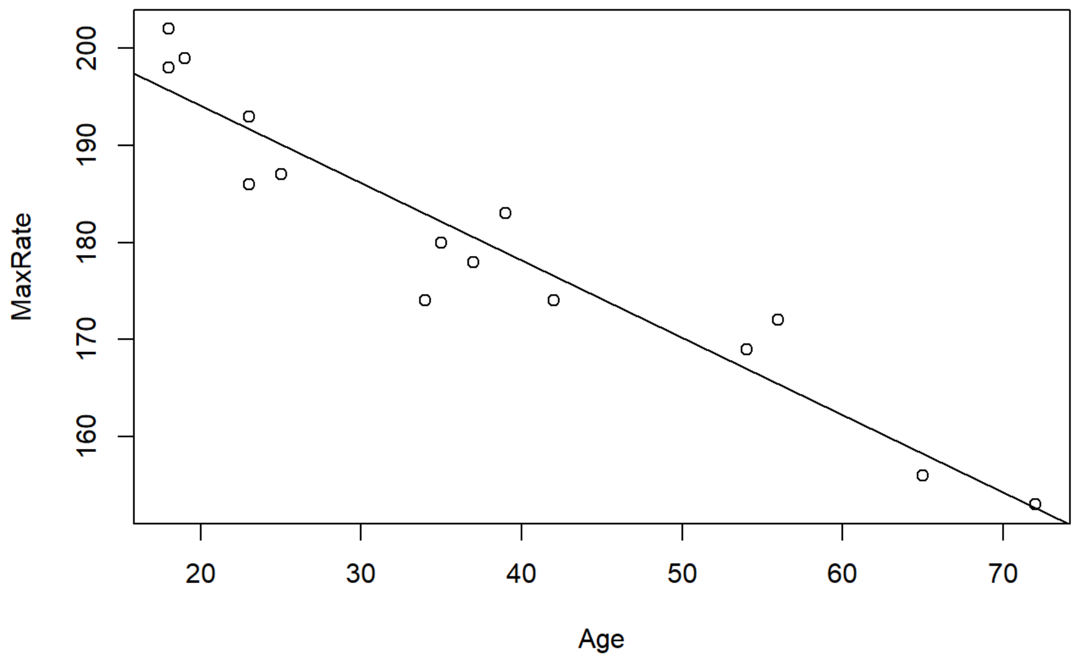
We can use the following functions: lm - linear model plot - plot the data abline - plot the regression line simple.lm - makes everithing required here.

a. The simple linear regression model is

$$Y = \hat{Y} + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

X is age
Y is maximum heart rate

b.
```
> plot(Age, MaxRate)
> abline(lm(MaxRate ~ Age))
```




```
> lm(MaxRate ~ Age)
```

Call:
lm(formula = MaxRate ~ Age)

Coefficients:

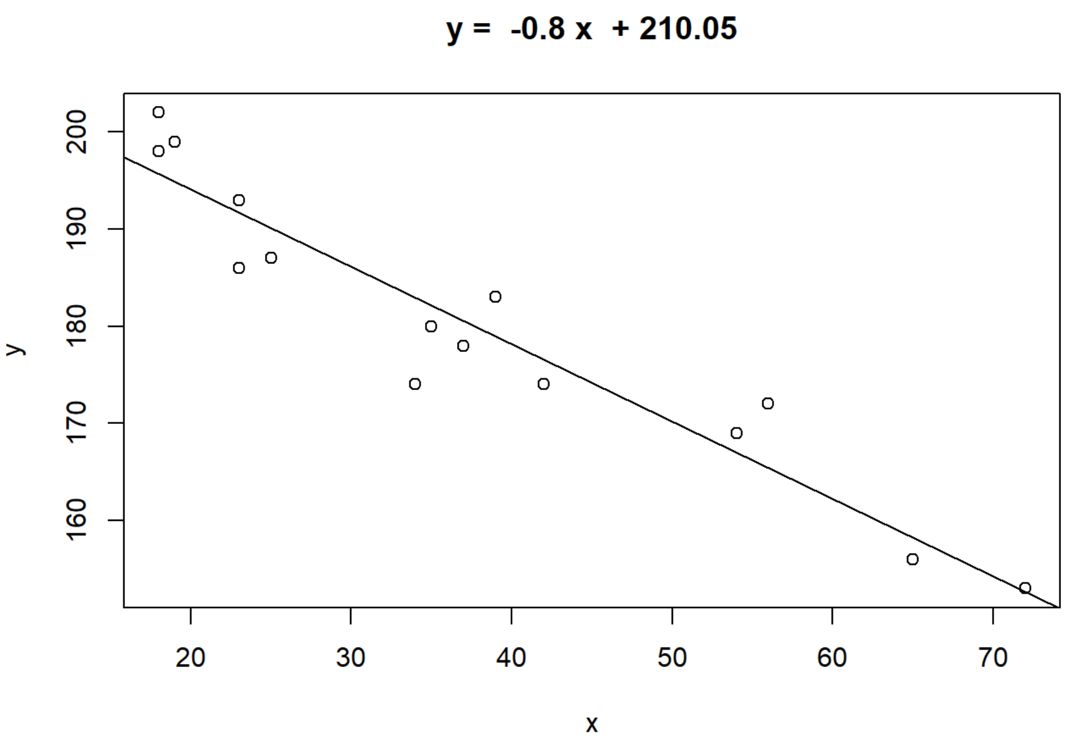
```
(Intercept)        Age
   210.0485     -0.7977
```

Then, $\beta_0 = 210.0485$. The model is

$$Y = 210.0485 - 0.7977X + \varepsilon$$

Or we can use

```
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':

    cancer
> lmResult <- simple.lm(Age, MaxRate)
```



```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.9021 
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
> class(lmResult)
[1] "lm"
> attributes(lmResult)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"         
 [5] "fitted.values" "assign"        "qr"            "df.residual"  
 [9] "xlevels"       "call"          "terms"         "model"        

$class
[1] "lm"
```

The result is of lm type.

First we find the coefficients: b_0, b_1

```
> coef(lmResult)
(Intercept)           x 
210.0484584  -0.7977266 
> lmResult[["coefficients"]]
(Intercept)           x 
210.0484584  -0.7977266 
> lmResult$coefficients
(Intercept)           x 
210.0484584  -0.7977266 
```

or by using the formula

```
> b1 <- sum((Age - mean(Age)) * (MaxRate - mean(MaxRate))) / sum((Age - mean(Age))^2); b1
[1] -0.7977266
> b0 <- mean(MaxRate) - b1 * mean(Age); b0
[1] 210.0485
```

or

```
> b1 <- cov(Age, MaxRate) / var(Age); b1
[1] -0.7977266
> b0 <- mean(MaxRate) - b1 * mean(Age); b0
[1] 210.0485
```

c. Let us now determine the expected maximum heart rate for any of these persons.

```
> predict(lmResult)
       1        2        3        4        5        6        7        8 
195.6894 195.6894 194.8917 191.7007 191.7007 190.1053 182.9258 182.1280 
       9       10       11       12       13       14       15 
180.5326 178.9371 176.5439 166.9712 165.3758 158.1962 152.6121 
```

or by using the formula

```
> yhat <- b0 + b1 * Age; yhat
 [1] 195.6894 191.7007 190.1053 182.1280 158.1962 166.9712 182.9258 165.3758
 [9] 152.6121 194.8917 191.7007 176.5439 195.6894 178.9371 180.5326
```

d. Determine the expected maximum heart rate for persons at age 30, 40, 50.

```
> yhat30 <- b0 + b1 * 30; yhat30
[1] 186.1167
> yhat40 <- b0 + b1 * 40; yhat40
[1] 178.1394
> yhat50 <- b0 + b1 * 50; yhat50
[1] 170.1621
```

e. We can find the errors(residuals): ε_i

```
> resid(lmResult)
         1          2          3          4          5          6          7 
 6.3106197  2.3106197  4.1083463 -5.7007474  1.2992526 -3.1052943 -8.9257552 
         8          9         10         11         12         13         14 
-2.1280287 -2.5325755  4.0628776 -2.5439427  2.0287761  6.6242292 -2.1962317 
        15 
 0.3878543 
> lmResult[["residuals"]]
         1          2          3          4          5          6          7 
 6.3106197  2.3106197  4.1083463 -5.7007474  1.2992526 -3.1052943 -8.9257552 
         8          9         10         11         12         13         14 
-2.1280287 -2.5325755  4.0628776 -2.5439427  2.0287761  6.6242292 -2.1962317 
        15 
 0.3878543 
> lmResult$residuals
         1          2          3          4          5          6          7 
 6.3106197  2.3106197  4.1083463 -5.7007474  1.2992526 -3.1052943 -8.9257552 
         8          9         10         11         12         13         14 
-2.1280287 -2.5325755  4.0628776 -2.5439427  2.0287761  6.6242292 -2.1962317 
        15 
 0.3878543 
```

or by using the formula

```
> e <- MaxRate - yhat; e
 [1]  6.3106197 -5.7007474 -3.1052943 -2.1280287 -2.1962317  2.0287761
 [7] -8.9257552  6.6242292  0.3878543  4.1083463  1.2992526 -2.5439427
[13]  2.3106197  4.0628776 -2.5325755
> summary(resid(lmResult))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-8.9258 -2.5383  0.3879  0.0000  3.1867  6.6242 
```

f. It is time to determine the mean square error of the model.

$$MSE = RSE^2 = S_\varepsilon^2 = \frac{\sum_{i=1}^n (Y_i \hat{-} Y_i)^2}{n-2} = \frac{\sum_{i=1}^n \varepsilon_i^2}{n-2}$$

is an unbiased estimator of σ_ε^2. The denominator $n-2$ comes from the fact that there are two values estimated from the data: β_0 and β_1.

Let us remind that

$$SSE = \sum_{i=1}^n \varepsilon_i^2, MSE = \frac{SSE}{n-r} = \frac{SSE}{n-2}$$

```
> SSE <- sum(e^2); SSE
[1] 272.4312
> n <- length(MaxRate)
> MSE <- SSE / (n - 2); MSE
[1] 20.95625
```

The Residual Standard error is

$$RSE = S_\varepsilon = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = 4.578$$

```
> s <- sqrt(MSE); s
[1] 4.577799
```

or we can extract it via the function summary

```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9021
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

g. Via the function $summary$ we can estimate also the coefficient of deteremination

$$\mathrm{cor}^2(X, Y) = 1 - \frac{\mathbb{E}\varepsilon^2}{\mathbb{D}Y}$$

Note that it is **Adjusted R-squared: 0.9021** as far as MSE has denominator $n-2$ and we cannot cancel it with the denominator $n-1$ of the estimator $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$ of $\mathbb{D}Y$.

The coefficient is close to 100%, therefore, we can say that the independent variable

X - the age is important for the value of the dependent variable

Y - the maximum heart rate. We can determine it also via the formula

```
> Rsquare <- 1 - MSE/var(MaxRate); Rsquare
[1] 0.9021041
```

If we consider these denominators as equal, then we can cancel them and obtain Multiple, R-squared: 0.9091. It does not takes into account that the denominators of the estimators

$$MSE = S_\varepsilon^2 = \frac{SSE}{n-2} \text{ and } S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \text{ are different and computes}$$

$$\text{MultipleR-squared} = 1 - \frac{SSE}{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2} = 0.9090967$$

```
> Rsq<-1 - SSE/sum((MaxRate - mean(MaxRate))^2); Rsq
[1] 0.9090967
```
or
```
> Rsquare <- cov(Age, MaxRate)^2/(var(Age)*var(MaxRate)); Rsquare
[1] 0.9090967
```
or
```
> Rsquare <- cor(Age, MaxRate)^2; Rsquare
[1] 0.9090967
```

h. In order to check if $\mathbb{E}\varepsilon = 0$ we use $t-test$.

$$H_0 : \mathbb{E}\varepsilon = 0$$
$$H_0 : \mathbb{E}\varepsilon \neq 0$$

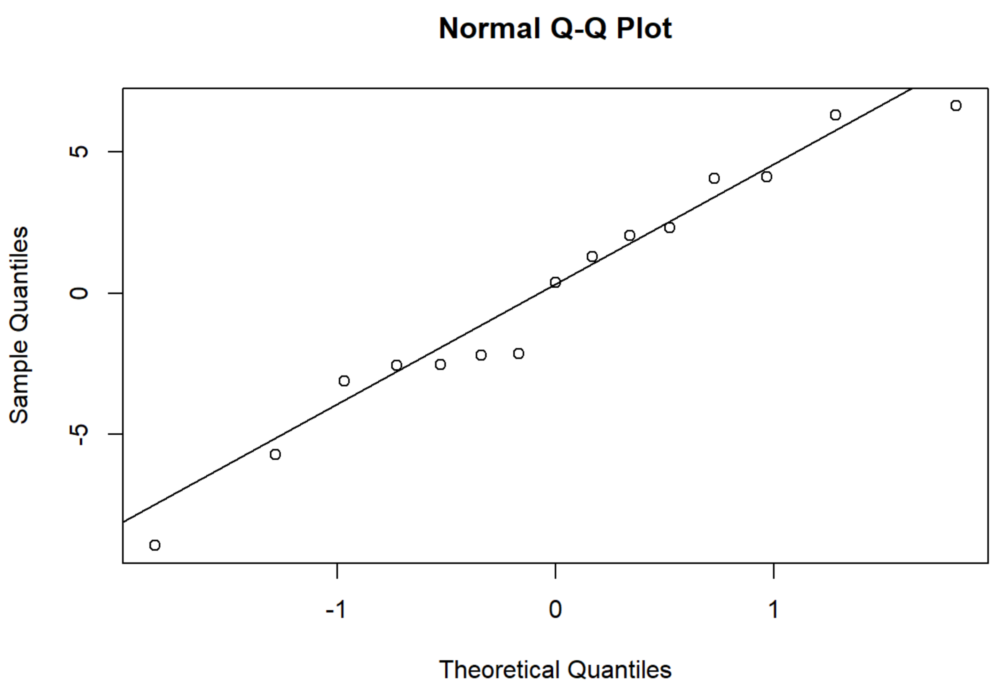
```
> t.test(e, mu = 0)

	One Sample t-test

data:  e
t = -6.6543e-15, df = 14, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.442884  2.442884
sample estimates:
    mean of x
-7.579123e-15
```

$p-value = 1 > 0.05 = \alpha$, so we have no evidence to reject H_0.

i. The next step is to test the assumptions of the model that the residuals are i.i.d. normally distributed $\varepsilon_i \in N(0,\sigma_\varepsilon^2)$

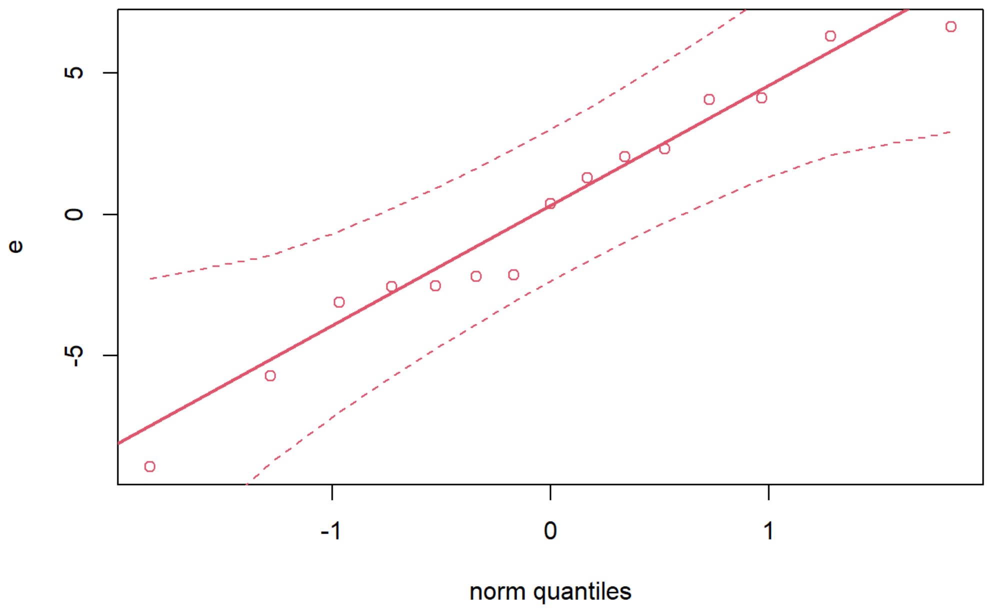
First we make the normal qq-plot

```
> qqnorm(e)
> qqline(e)
```



```
> library(StatDA)
Warning: package 'StatDA' was built under R version 4.0.3
Loading required package: sgeostat
Warning: package 'sgeostat' was built under R version 4.0.3
Registered S3 method overwritten by 'geoR':
  method         from
  plot.variogram sgeostat
```

```
> qqplot.das(e)
```



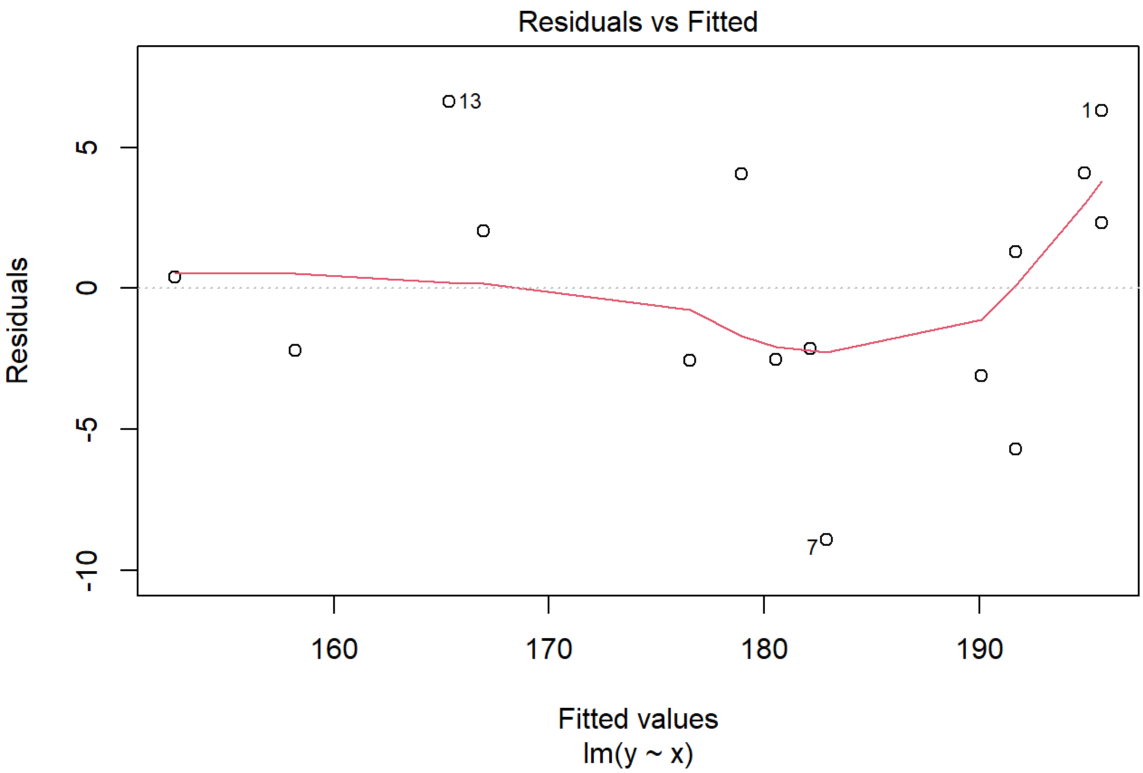
We can perform Shapiro test

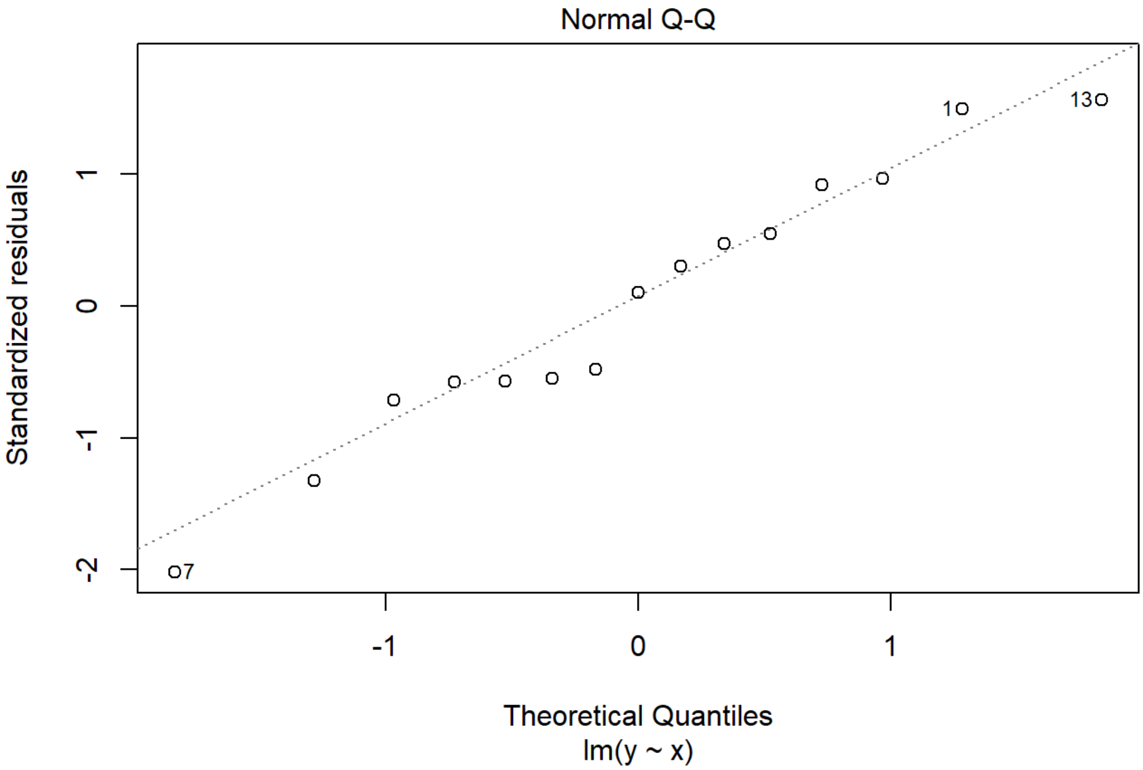
H_0 : ε is normally distributed
H_A : ε is not normally distributed

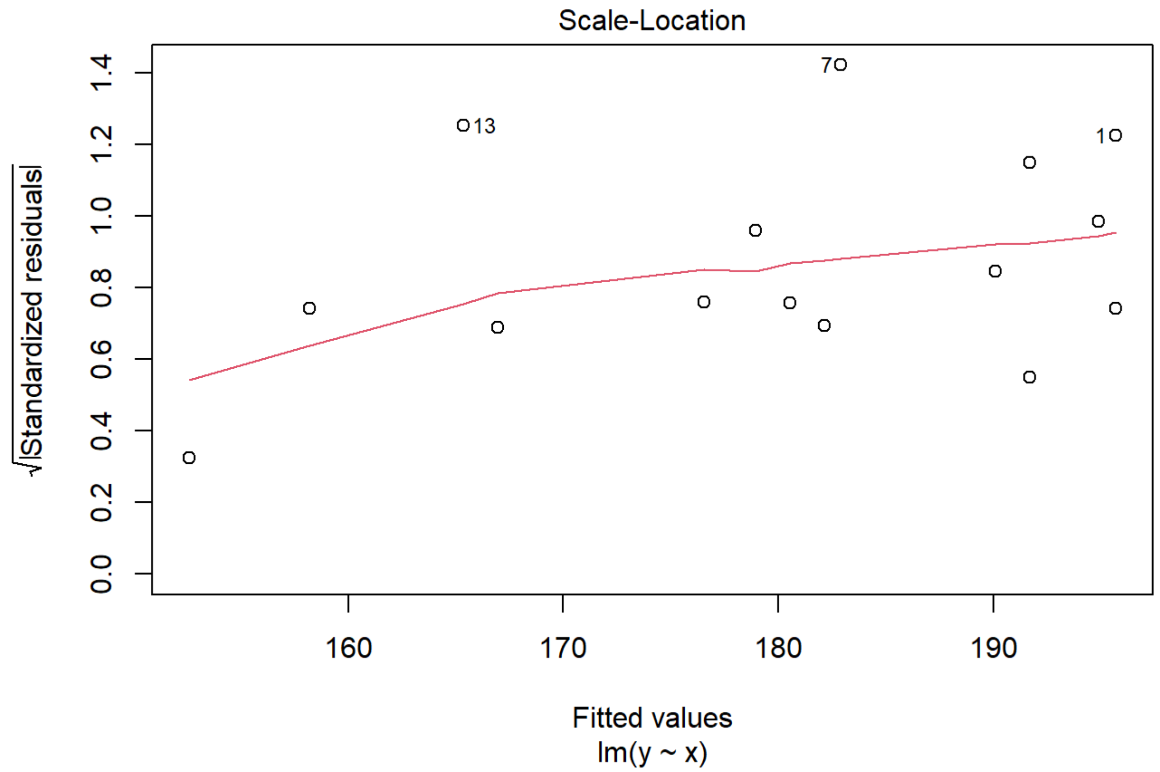
```
> shapiro.test(e)

	Shapiro-Wilk normality test

data:  e
W = 0.96302, p-value = 0.7447
```

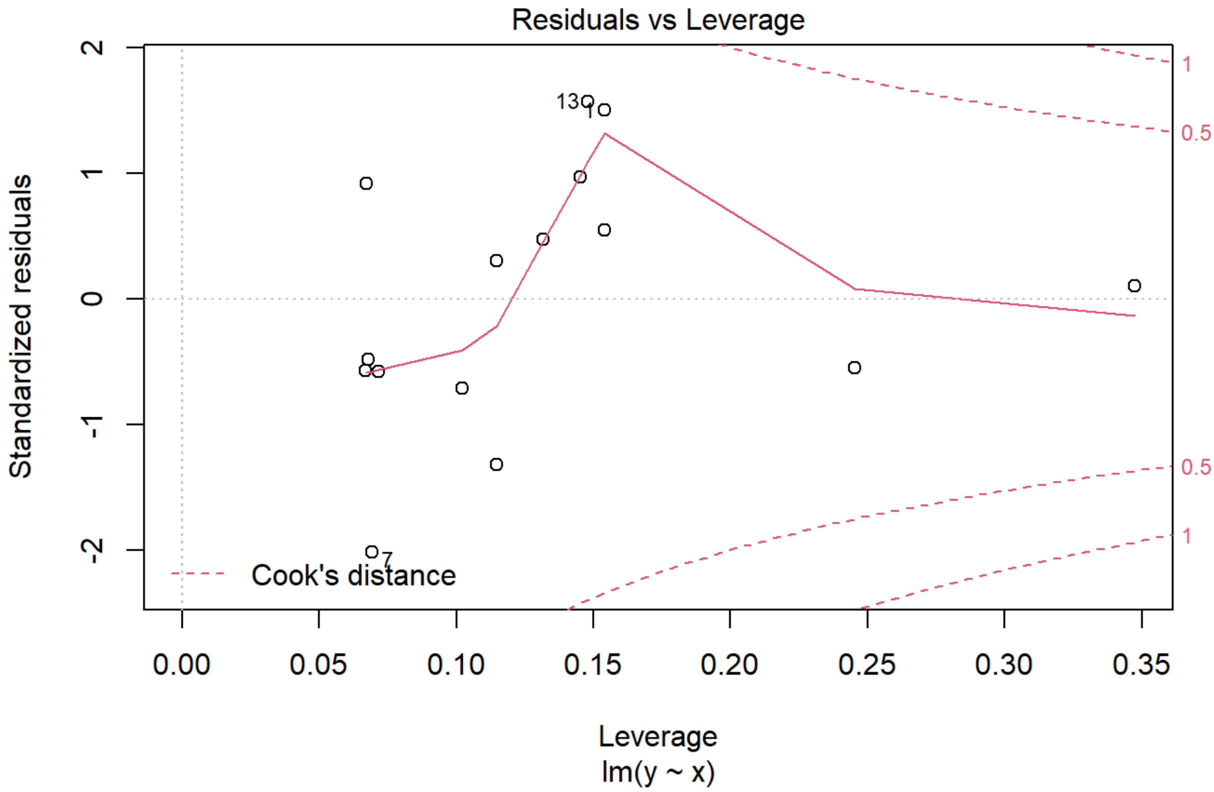
The $p-value = 0.7447 > 0.05 = \alpha$, so we have no evidence to reject H_0.
We can be check all this graphically by

```
> plot(lmResult)
```






Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
7
13
10
160
170
180
190
Fitted values
lm(y ~ x)


Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
13
10
7
Cook's distance
1
0.5
0.5
1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Leverage
lm(y ~ x)

Graphic

- 1 - shows **Residuals vs. fitted** graph. This plots the fitted $\hat{Y}$ against the residuals. Look for spread around the line $y = 0$. This graph help us **check if there is obvious trend for the residuals** or it replaces the test of the hypothesis $H_0 : \mathbb{E}\varepsilon = 0$ that we already did.
- 2 - represent Normal qqplot. This graph help us **check if the residuals are normally distributed.**
- 3 - shows Scale location - square root of the standardized residuals. This graph help us to check if the variance of the error term is a constant, which is the same to observe **homo-** or **heteroscedasticity**. Moreover it helps us to recognize **the largest residuals by looking at the tallest points.**
- 4 - shows **Cook's distance.** This plot help us **identify points which have a lot of influence in the regression line.**

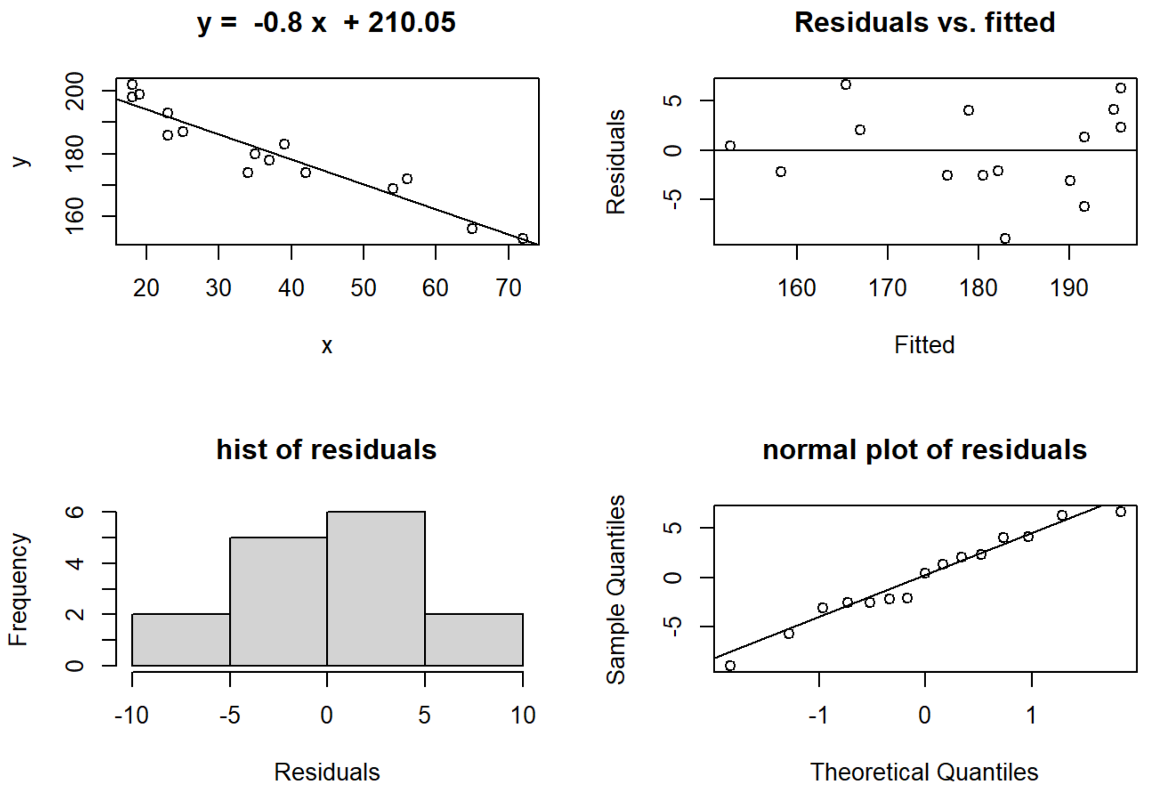
Cook's distance of the j-th observation is

$$D_j = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \hat{Y}_i^{(j)})^2}{2MSE}$$

where $\hat{Y}_i^{(j)}$ is the expected value of Y given X_i when the simple regression model is built up without the j-th observation.

Or we can use the pictures from the simple.lm function

```
> simple.lm(Age, MaxRate, show.residuals = TRUE)
```

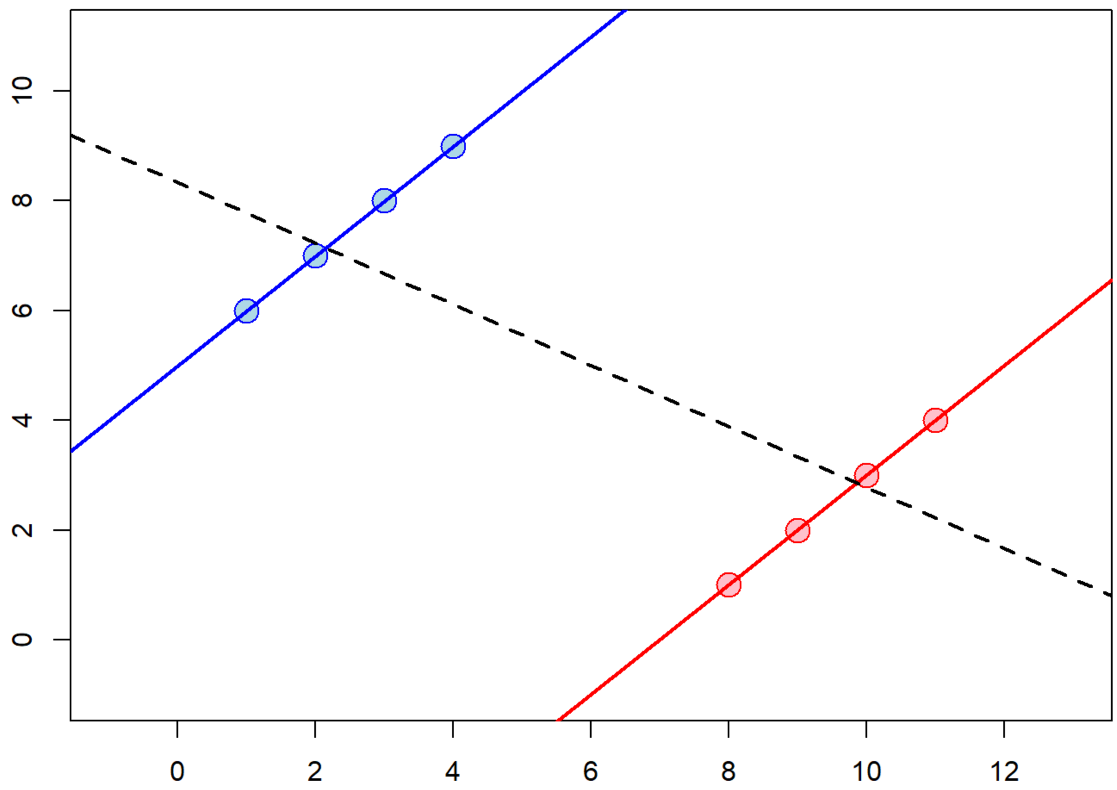


```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
   210.0485      -0.7977
```

Simpson's paradox

When we analyze the results from the regression analysis we have to take in mind the possibility for the following **Simpson's paradox**

```
> x1 <- c(1,2,3,4)
> y1 <- x1 + 5
> x2 <- x1 + 7
> y2 <- x2 - 7
> x <- c(x1,x2)
> y <- c(y1,y2)
> par(mar = c(3,3,0.5,0.5))
> plot(x, y, cex = 2, pch = 21, col = rep(c("blue", "red"), each=4), bg = rep(c("lightblue", "pink"), each=4), xlim = range(x) + c(-2,2), ylim = range(y)+ c(-2,2))
> abline(lm(y1 ~ x1), col="blue", lwd=2)
> abline(lm(y2 ~ x2), col="red", lwd=2)
> abline(lm(y  ~ x), lwd=2, lty=2)
```



In such cases it is better to divide the population in subgroups and then to perform regression analysis in any of these groups.

Confidence intervals for $\mathbb{E}(Y \mid X = x)$ and $(Y \mid X = x)$

The regression line is used to predict the value of Y for a given X, or the average value of Y for a given X and we would like to know how accurate this prediction is. **Confidence interval** do these.

The estimator of the mean value of Y given $X = X_i$ has a standard error of

$$SE(\hat{Y}_i) = S_\varepsilon \sqrt{\frac{1}{n} + \frac{(X_i - \overline{X}_n)^2}{\sum_{j=1}^{n} (X_j - \overline{X}_n)^2}}$$

Therefore, the confidence interval is

$$[\hat{Y}_i - t_{1-\frac{\alpha}{2};n-r} \times SE(\hat{Y}_i);\ \ \hat{Y}_i + t_{1-\frac{\alpha}{2};n-r} \times SE(\hat{Y}_i)]$$
$$[\beta_0 + \beta_1 X_i - t_{1-\frac{\alpha}{2};n-r} \times SE(\hat{Y}_i);\ \beta_0 + \beta_1 X_i + t_{1-\frac{\alpha}{2};n-r} \times SE(\hat{Y}_i)]$$

Example 2

Compute and plot 90% confidence intervals for $\mathbb{E}(Y \mid X = X_i)$ in the previous example.

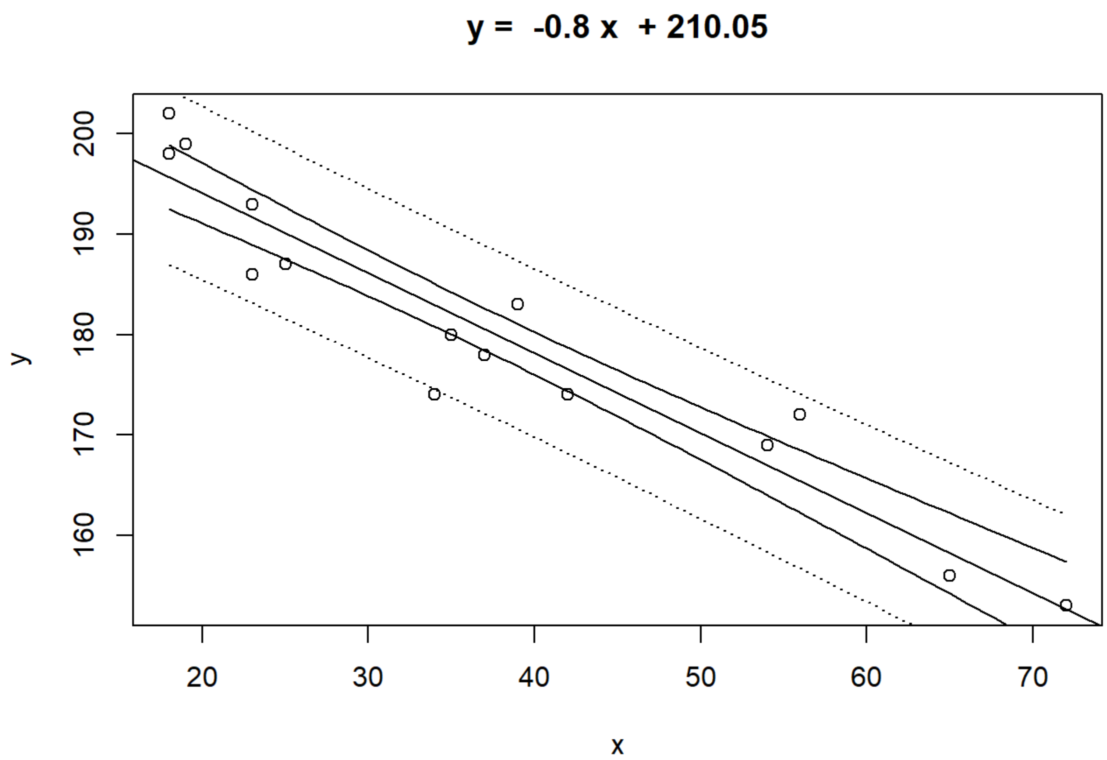
Solution:

The function predict computes the estimators for $\mathbb{E}(Y \mid X = X_i)$ (the fitted values) and the corresponding confidence intervals.

```
> pr = predict(lmResult, interval = "confidence", level = 0.90)
> head(pr)
       fit      lwr      upr
1 195.6894 192.5083 198.8705
2 195.6894 192.5083 198.8705
3 194.8917 191.8028 197.9805
4 191.7007 188.9557 194.4458
5 191.7007 188.9557 194.4458
6 190.1053 187.5137 192.6969
```

The function simple.lm(show.ci = TRUE, ...) plots these confidence intervals via solid lines. The following script produces a graph with $90\,\%$ confidence bands drawn.

> simple.lm(Age, MaxRate, show.ci = TRUE, conf.level = 0.90)



Call:
lm(formula = y ~ x)

Coefficients:

(Intercept)	x
210.0485	-0.7977

Dashed lines are confidence intervals for unique values $Y \mid X = X_i$ based on the fact that if the assumptions of the model are satisfied $\varepsilon_i \in N(0, \sigma_\varepsilon^2)$. As far as the confidence intervals for the next value of the error term is

$$\left[\bar{\varepsilon}_n - t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1+\frac{1}{n}};\ \bar{\varepsilon}_n + t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1+\frac{1}{n}}\right]$$

if the condition $\mathbb{E}\varepsilon = 0$ is satisfied, then there is no statistically significant difference between $\bar{\varepsilon}_n$ and 0, therefore, we can use

$$\left[{}_n - t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1+\frac{1}{n}};\ t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1+\frac{1}{n}}\right]$$

and the confidence interval for the values of $(Y \mid X = X_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$ are

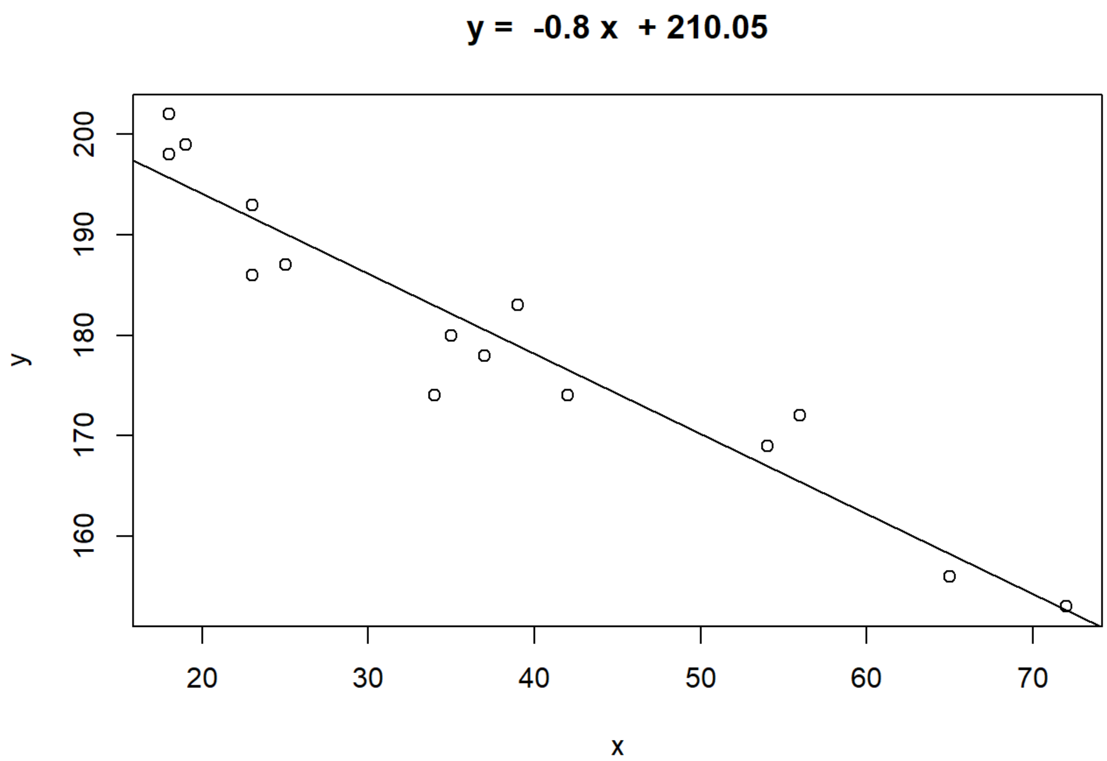
$$\left[\beta_0 + \beta_1 X_i - t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1 + \frac{1}{n}};\ \beta_0 + \beta_1 X_i + t_{1-\frac{\alpha}{2};n-1} S_\varepsilon \sqrt{1 + \frac{1}{n}}\right]$$

Example 3

In the previous example determine $90\,\%$ confidence intervals for the mean of the maximum heart rate for persons at age 30, 40, 50.

Solution:

```
> library(UsingR)
> lmResult <- simple.lm(Age, MaxRate)
```

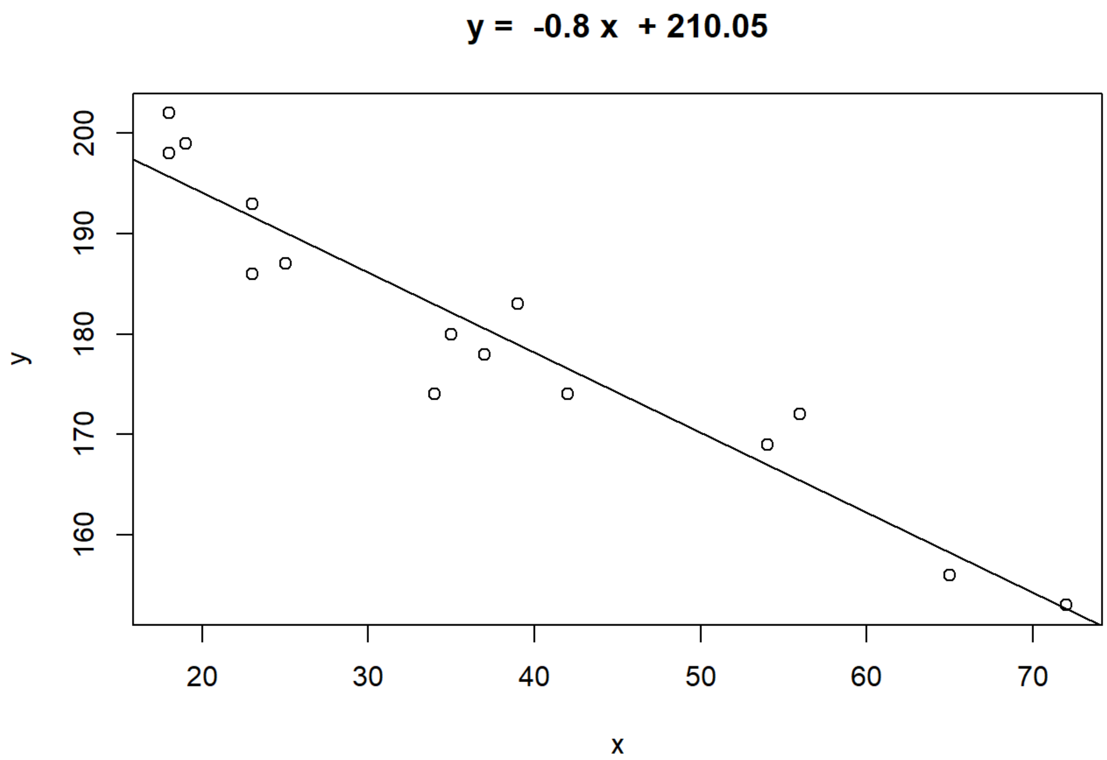



```
> e<-resid(lmResult)
> SSE <- sum(e^2); SSE
[1] 272.4312
> n <- length(MaxRate)
> MSE <- SSE / (n - 2); MSE
[1] 20.95625
> Seps<-sqrt(MSE)
> ci30<-yhat30 + c(-1,1)*Seps*sqrt(1/n+(30-mean(Age))/sum((Age-mean(Age))^2)); ci30
[1] 184.9500 187.2834
> ci40<-yhat40 + c(-1,1)*Seps*sqrt(1/n+(40-mean(Age))/sum((Age-mean(Age))^2)); ci40
[1] 176.9519 179.3269
> ci50<-yhat50 + c(-1,1)*Seps*sqrt(1/n+(50-mean(Age))/sum((Age-mean(Age))^2)); ci50
[1] 168.9542 171.3701
```

Example 4

In the previous example determine $90\,\%$ confidence intervals for the next observed maximum heart rate for persons at age $30, 40, 50$.

```
> library(UsingR)
> lmResult <- simple.lm(Age, MaxRate)
```




```
> e<-resid(lmResult)
> SSE <- sum(e^2); SSE
[1] 272.4312
> n <- length(MaxRate)
> MSE <- SSE / (n - 2); MSE
[1] 20.95625
> Seps<-sqrt(MSE)
> ci30<-yhat30 + c(-1,1)*Seps*sqrt(1/n+1); ci30
[1] 181.3887 190.8446
> ci40<-yhat40 + c(-1,1)*Seps*sqrt(1/n+1); ci40
[1] 173.4115 182.8673
> ci50<-yhat50 + c(-1,1)*Seps*sqrt(1/n+1); ci50
[1] 165.4342 174.8901
```

When compare the results from this and the previous task we see that the confidence interval for unique values are wider than those for the corresponding means.

Statistical inference related with simple linear regression models

Confidence intervals for $\mathbb{E}\beta_1$ and hypothesis testing related with the slope β_1 of the regression line

If we consider the unbiased estimator $\hat{\beta}_1$ of β_1 as a random variable and if the assumptions of the model are satisfied (inclusively the requirement for the normality of the residual term), then we can compute confidence intervals for $\mathbb{E}\hat{\beta}_1 = \beta_1$ and we can test the hypothesis if β_1 is equal to a given constant.

The most frequently we test if $\beta_1 = 0$ which means that the independent variable X has no statistically significant influence on Y. Or this could be one of the ways to say that the model is not adequate.

The standard error of the unbiased estimator $\hat{\beta}_1$ of β_1 is given by

$$SE(\beta_1) := \frac{S_\varepsilon}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\beta_1)} \in t(n-2)$$

Therefore, the corresponding confidence interval for β_1 is

$$[\hat{\beta}_1 - t_{1-\frac{\alpha}{2};n-2}SE(\beta_1);\ \hat{\beta}_1 + t_{1-\frac{\alpha}{2};n-2}SE(\beta_1)]$$

And for $b_1 =$ const we can test

$H_0 : \beta_1 = b_1$
$H_A : \beta_1 \neq b_1$

Given α the critical area is

$$W_\alpha = \left\{ \frac{|\hat{\beta}_1 - b_1|}{SE(\beta_1)} \geq t_{1-\frac{\alpha}{2};n-2} \right\}$$

Example 5
In the previous example

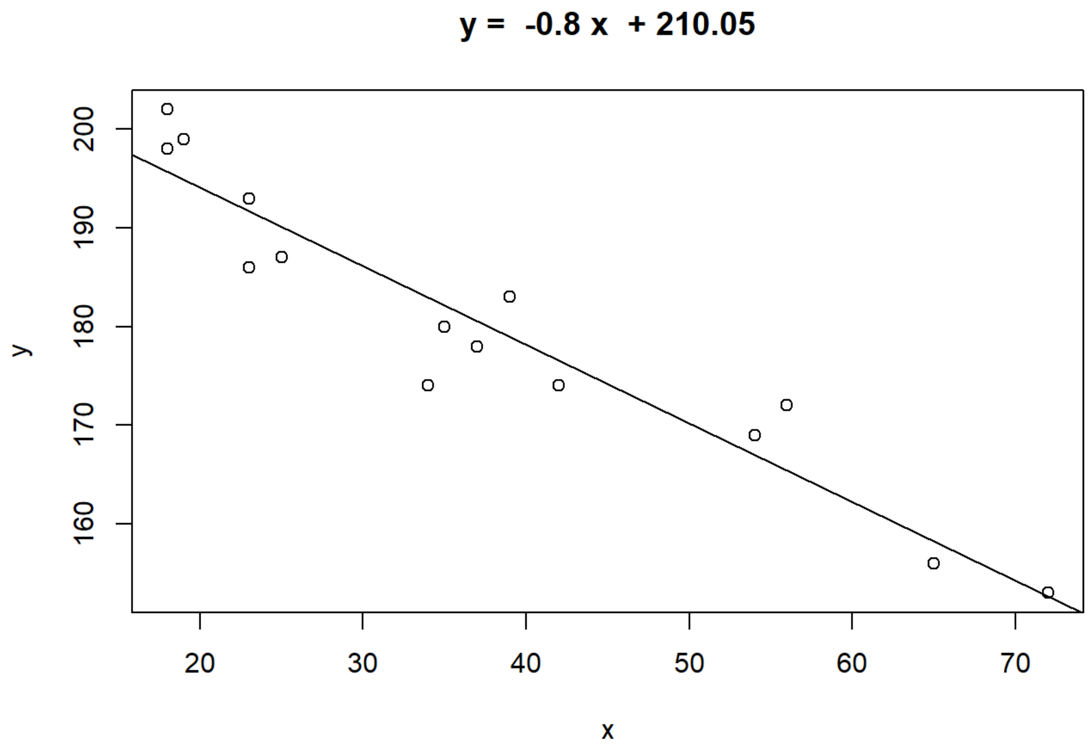
a. construct confidence interval for the parameter β_1.
b. Test the hypothesis that it is equal to -1.
c. Test the hypothesis that it is equal to 0.
d. We compute the required confidence interval via the following function which computes confidence intervals given the corresponding statistics $bhat$ computed

from the data, the corresponding quantile t and the corresponding SE

```
> myCI = function(bhat, SE, t) {
+ bhat + c(-1,1)*SE*t
+ }
```

In this case first we have to compute

```
> library(UsingR)
> lmResult <- simple.lm(Age, MaxRate)
```




```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.9021
```

```
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
> e <- resid(lmResult)
> n<-length(e)
> beta1hat <- (coef(lmResult))[['x']]; beta1hat
[1] -0.7977266
> Seps <- sqrt(sum(e^2)/(n-2))
> SEbeta1 <- Seps / sqrt(sum((Age - mean(Age))^2)); SEbeta1
[1] 0.06996281
> alpha<-0.05
> t <-qt(1-alpha/2, n - 2, lower.tail = TRUE)
> myCI(beta1hat, SEbeta1,t)
[1] -0.9488720 -0.6465811
```

As far as -1 is not in this confidence interval we can guess that the following H_0 will be rejected, however let us see.

e. We test

$$H_0 : \beta_1 = -1$$
$$H_A : \beta_1 \neq -1$$

```
> const <- -1
> temp <- abs(beta1hat-const)/SEbeta1; temp
[1] 2.891157
> pvalue<-2*pt(temp, n - 2, lower.tail = FALSE); pvalue
[1] 0.01262031
```

The $p-value = 0.01262031 < 0.05 = \alpha$, so it is unlikely for this data the slop to be -1 and we reject H_0.

f. It will automatically do a hypothesis test for

$H_0 : \beta_1 = 0$ there is no statistically significant dependence between X and Y, which means that there is no slope in the regression line, or which is the same X and Y are linearly uncorrelated.

```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared: 0.9091,   Adjusted R-squared: 0.9021
F-statistic: 130 on 1 and 13 DF, p-value: 3.848e-08

$H_A : \beta_1 \neq 0$
The $p-value < 3.85e-08 < 0.05 = \alpha$, therefore, we reject H_0. The linear dependence between X and Y is statistically significant.

Confidence intervals for $\mathbb{E}\beta_0$ and hypothesis testing related with the intercept β_0 of the regression line on Oy.

The **standard error** of the unbiased estimator $\hat{\beta}_0$ of β is given by

$$SE(\beta_0) := S_\varepsilon \sqrt{\frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \overline{X}_n)^2}} = S_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{X}_n^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}}$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\beta_0)} \in t(n-2)$$

Therefore, the corresponding confidence interval for β_0 is

$$[\hat{\beta}_0 - t_{1-\frac{\alpha}{2};n-2} SE(\beta_0); \hat{\beta}_0 + t_{1-\frac{\alpha}{2};n-2} SE(\beta_0)]$$

And for $b_0 = const$ we can test

$H_0 : \beta_0 = b_0$
$H_A : \beta_0 \neq b_0$

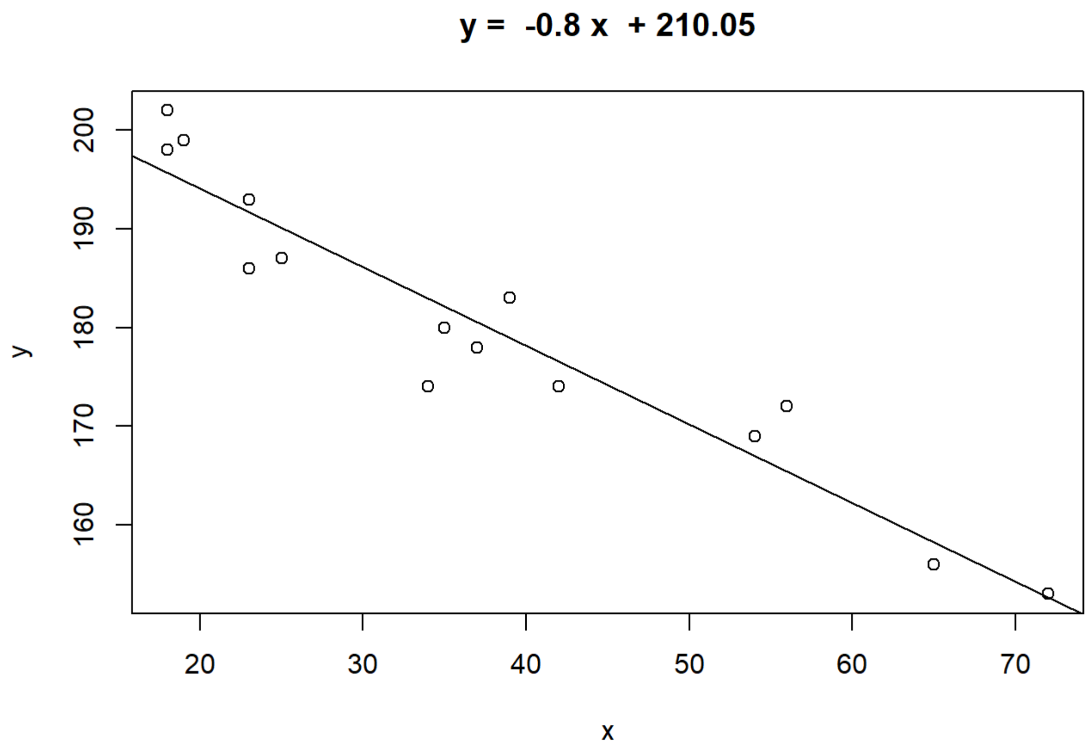
Given α the critical area is

$$W_\alpha = \left\{ \frac{|\hat{\beta}_0 - b_0|}{SE(\beta_0)} \geq t_{1-\frac{\alpha}{2};n-2} \right\}$$

Example 6
In the previous example

a. construct confidence interval for the parameter β_0
b. Test the hypothesis that the regression line goes trough the coordinate origin.
c. Test the hypothesis that it is equal to 220.
d. In order to compute the required confidence interval we are going to use again our function $myCI$. In this case

```
> library(UsingR)
> lmResult <- simple.lm(Age, MaxRate)
```




```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.9021 
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
> beta0hat <- (coef(lmResult))[['(Intercept)']];  beta0hat
[1] 210.0485
> SEbeta0 <- Seps * sqrt(sum(Age^2)/(n*sum((Age - mean(Age))^2))); SEbeta0
[1] 2.866939
> myCI(beta0hat, SEbeta0, t)
[1] 203.8548 216.2421
```

As far as 0 is not in this confidence interval we can guess that the following H_0 will be rejected, however let us see.

e. We test

$H_0 : \beta_0 = 0$ which means that there is no intercept of Oy in the regression line.
$H_A : \beta_0 \neq 0$

```
> const <- 0
> temp <- abs(beta0hat-const)/SEbeta0; temp
[1] 73.26576
> pvalue<-2*pt(temp, n - 2, lower.tail = FALSE); pvalue
[1] 2.124074e-18
```

See also the outputs of summary(lmResult).

The $p-value = 2.124074e-18 < 0.05 = \alpha$, so it is unlikely for this data the intercept to be 0 and we reject H_0.

c. As far as 220 is outside the built confidence interval we can guess that we will reject the next H_0. Now let us automatically test for

$H_0 : \beta_0 = 220$, which means that there is no statistically significant difference between the intercept and 220.

$H_A : \beta_0 \neq 220$

```
> SEbeta0 <- Seps * sqrt(sum(Age^2) / (n * sum((Age - mean(Age))^2))); SEbeta0
[1] 2.866939
> temp <- abs(beta0hat - 220) / SEbeta0; temp
[1] 3.471138
> pvalue<-2*pt(temp, n - 2, lower.tail = FALSE); pvalue
[1] 0.004136843
```

The $p-value = 0.004136843 < 0.05 = \alpha$, so we reject the value H_0. The difference between β_1 and 220 is statistically significant.

Tests for adequacy

Tests for adequacy check if the independent variable X has no statistically significant influence on

H_0 : The model is not adequate. The linear dependence between X and Y is not statistically significant. I.e. the slope $\beta_1 = 0$.

H_A : The model is adequate. The linear dependence between X and Y is statistically significant. I.e. the slope $\beta_1 \neq 0$.

As you can see for this model the test for adequacy is equivalent to the one for $H_0 : \beta_1 = 0$.

However more generally it is an F-test and given $\alpha > 0$ the critical area is

$$W_\alpha = \left\{ \frac{\frac{SS(\hat{Y})}{r}}{SSE} n - r - 1 \geq x_{1-\alpha, F(r; n-r-1)} \right\}$$

Here we have used that

$$\left(\frac{\frac{SS(\hat{Y})}{r}}{\frac{SSE}{n-r-1}} | H_0 \right) \in F(r; n-r-1), SS(\hat{Y}) := \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}_n)^2, \overline{Y}_n = \overline{\hat{Y}}_n$$

When we divide the numerator and the denumerator to

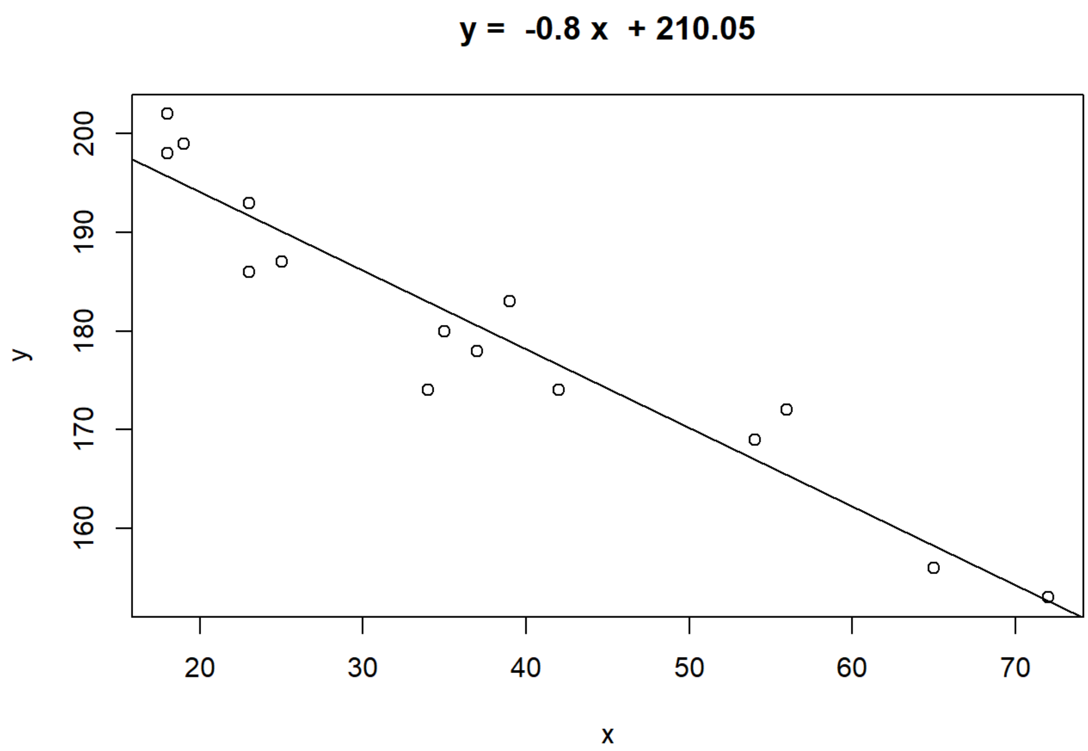
$$SS(Y) := \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$$

we obtain

$$\frac{\frac{SS(\hat{Y})}{r}}{\frac{SSE}{n-r-1}} = \frac{\frac{R^2}{r}}{\frac{1-R^2}{n-r-1}}$$

Example 7
In the previous example test the simple linear regression model for adequacy.

```
> library(UsingR)
> lmResult <- simple.lm(Age, MaxRate)
```




```
> summary(lmResult)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9021 
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

Here $F-statistic : 130$ is the empirical value of $\frac{\frac{SS(\hat{Y})}{r}}{\frac{SSE}{n-r-1}}$. We use the p-value of the F-statistics $p-value = 3.848e-08 < 0.05 = \alpha$, therefore, we reject H_0. The model is adequate. The linear dependence between X and Y is statistically significant.

Second way to make the same is:

```
> beta0hat <- (coef(lmResult))[['(Intercept)']];  beta0hat
[1] 210.0485
> beta1hat <- (coef(lmResult))[['x']];  beta1hat
[1] -0.7977266
> r<-1
> yhat<-beta0hat+beta1hat*Age
> SSE <- sum(e^2); SSE
[1] 272.4312
> n <- length(MaxRate)
> SSYhat<-sum((yhat-mean(yhat))^2); SSYhat
[1] 2724.502
> Femp <- (SSYhat/r) / (SSE/(n-r-1)) ; Femp
[1] 130.0091
> Fquantile<-qf(1-alpha, df1=r, df2=n - r - 1);Fquantile
[1] 4.667193
> pvalue<-pf(Femp, df1=r, df2=n - r - 1, lower.tail = FALSE);pvalue
[1] 3.847987e-08
```

The $p-value = 3.847987e-08 < 0.05 = \alpha$, therefore, we reject H_0. The model is adequate. The linear dependence between X and Y is statistically significant.

Third way to make the same.

It is faster to use the function anova. Its names comes from **Analysis of Variances / Дисперсионния анализ/**

```
> anova(lmResult)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x          1 2724.50 2724.50  130.01 3.848e-08 ***
Residuals 13  272.43   20.96                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here $F-statistic = 130.01$ is the empirical value of $\frac{\frac{SSY(\hat{Y})}{r}}{\frac{SSE}{n-r-1}}$. We use the p-value of the F-statistics $p-value = 3.848e-08 < 0.05 = \alpha$, therefore, we reject H_0. The model is adequate. The linear dependence between X and Y is statistically significant.