CHAPTER 10. SIMPLE REGRESSION ANALYSIS

1. Regression analysis
2. Scatter plot
3. Correlation coefficient
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination

Chapter 10: Assignments

1. **The number of hours of study of the students of a course and the final grade of the students (out of 100), is shown in the table. Calculate the correlation coefficient and determine whether the correlation is significant. Obtain the regression line.**

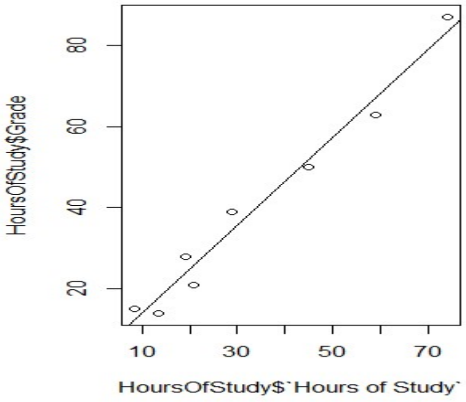
Hours_of_study	Grade
74	87
59	63
45	50
29	39
20.8	21
19.1	28
13.4	14
8.5	15

Solution:

A. First use the function in R.
- In this case we can see that the correlation result we get is equal to 0,9877

B. Compare the p-value from cor.test and conclude whether it is significant.
- In this case we can see that we have got a p-value which is 4,537e-06 which 0, 00004.. and that is lower than the confidence level which is 95%=0,05.
- Therefore, we can say, that it is significant because of the lower level of the p-value.

C. Construct a scatterplot graph and draw the regression line.
- In this case, we will use the scatterplot by using the command function plot and thereafter we will use abline. Here the graph with the regression line will be as follows:



D. Conclude
- Therefore, we can in the end conclude that the graph is positive strong because of the correlations coefficient and thereafter we can see that the correlation is not significant. And the regression line becomes as follows.

2. **A researcher carried out an experiment to investigate the relationship between alcohol consumption and blood concentration. The experiment included 5 participants. These were the results:**

Participant number of glasses parts per 1000

	Alcohol consumption,	Blood alcohol concentration,
1	1	10
2	2	8
3	3	12
4	4	16
5	5	20

a. **Is there a significant relationship between the alcohol consumption, and the concentration of alcohol in blood?**

Solution:

1. Start by using the cor.test function to find the p-value.

- We can in this case, see that the p-value has been found to become p-value=0,02776 and thereafter we can compare it with 95%. Here we can see that there is a significant relationship between the alcohol consumption and the concentration of alcohol in blood.

b. **What is the equation of the regression line?**

Solution:

1. Use the scatterplot function to draw the graph.

- Here in this case, we can see that our graph becomes as followed.

2. Use the lm-function and thereafter abline-function.

- We have afterwards use the abline function and therefore we get the following result of regression line of the graph.

c. **What is the % of the variance in blood alcohol concentration that can be explained by the alcohol consumption?**

Solution:

1. **Look at the correlation's coefficient from the graph.**

- Here it can be seen that the correlations coefficient is equal to 0,9877 which is very close to become a very precise linear line. But when we are asked about the percentage then we can see that the variance of blood alcohol becomes 98,77%.

d. **We want to predict the blood alcohol concentration of a person that has consumed 4.2 glasses. What is the predicted value of blood alcohol concentration and the prediction interval?**

Solution:

1. **Use the predict function.**

- In this case we can see that the following picture shown below is where we define a function so we can predict something. In this case we can see that we have written Regression_filename and then we have defined the data.frame with the axis of the Number of Glass which has been defined to be 4,2. Along with that the interval has also been added, where it is referred to predict.

2. **Write the Interval.**

- From the function we have got the following interval:
- 16,56 < 8,47 < 24,64

3. **Conclude the prediction of 4.2 glasses with help of function.**

- In this case we can see, that after having consumed 4,2 glasses we can afterwards predict that we have got a value of 16,56 with the interval of: 16,56 < 8,47 < 24,64

3. Businesses often use linear regression to understand the relationship between advertising spending and revenue. For example, they might fit a simple linear regression model using advertising spending as the predictor variable and revenue as the response variable. Calculate the value of the correlation coefficient between advertising spending and revenue based on the following data. What is the predicted revenue if a business spends 50 million DKK?

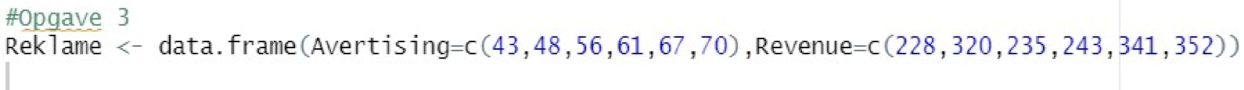
DKK

Business	Advertising spending, in million	Revenue, in million DKK
1	43	228
2	48	320
3	56	235
4	61	243
5	67	341
6	70	352

Solution:

1. **Start by construction a dataframe, to become independent from Excel 😊**

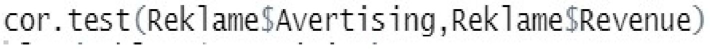
- We will start by using the keyword of the function data.frame from Lesson 7.
- We will write the keyword and inside the parentheses the axis are defined as shown on the following example.



```
#Opgave 3
Reklame <- data.frame(Avertising=c(43,48,56,61,67,70),Revenue=c(228,320,235,243,341,352))
```

2. **Use the correlation coefficient test function.**

- The following example shows how the function has been written.

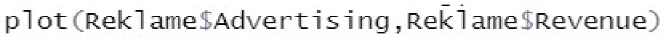


```
cor.test(Reklame$Avertising,Reklame$Revenue)
```

- In this case we can see that the following function gives us the following answer: 0,5912

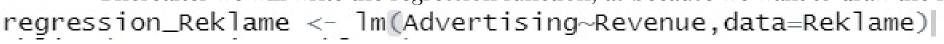
3. **Draw the scatterplot and draw the regression line.**

- In this case we can see that we have written the following plot-function.



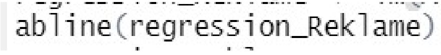
```
plot(Reklame$Advertising,Reklame$Revenue)
```

- Thereafter we will write the regression function, as because we want to draw the regressionsline with lm.



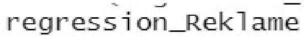
```
regression_Reklame <- lm(Advertising~Revenue,data=Reklame)
```

- Now we want to print the regression line's function.



```
abline(regression_Reklame)
```

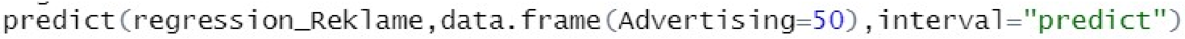
- Now, we will write the regression_filenname.



```
regression_Reklame
```

4. **Use the predict function and create the prediction.**

- The following function, shows how the prediction function is being used.



```
predict(regression_Reklame,data.frame(Advertising=50),interval="predict")
```

5. In a time, series, we find a significant relationship between the increase in the number of people who are exercising in Denmark and the increase in the number of people who are committing crimes in US. Comment on whether there is causation.

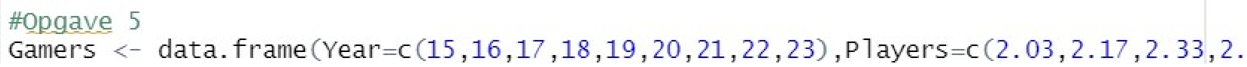
6. Available videogaming statistics have estimated that there are 3.1 billion gamers across the globe. The number of gamers from 2015 to 2023 is shown in the table. What will be the number of gamers across the globe in the year 2040?

Year	Global_players
2015	2.03
2016	2.17
2017	2.33
2018	2.49
2019	2.64
2020	2.81
2021	2.96
2022	3.09
2023	3.22

Solution:

1.Use the Dataframe to construct the table for the Gamers.

- Looking at the picture, we have constructed a data frame, where we afterwards will follow the same steps which we did in the previous tasks.



```
#Opgave 5
Gamers <- data.frame(Year=c(15,16,17,18,19,20,21,22,23),Players=c(2.03,2.17,2.33,2.
```

2. Use the correlations coefficient function and compare with the p-value.

- We have used the correlations coefficient function, and we have afterwards found the correlations coefficient for the gamers and also the p-value. We can see by comparing both values together, that the p-valuer is significantly very low, which tells us that there is a strong correlation between Years and Gamers.

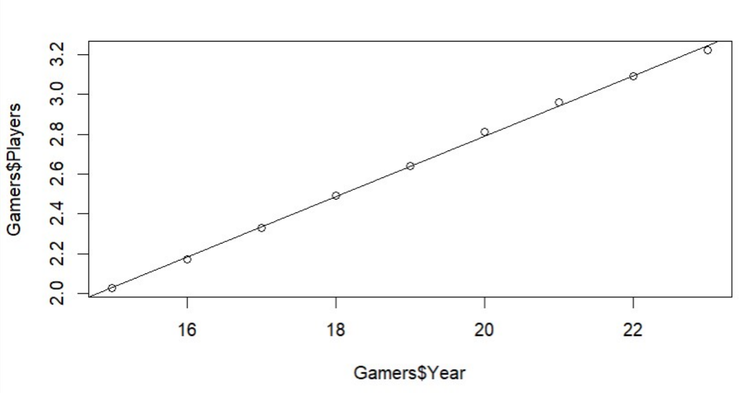
3. Use the plot function to draw the graph.

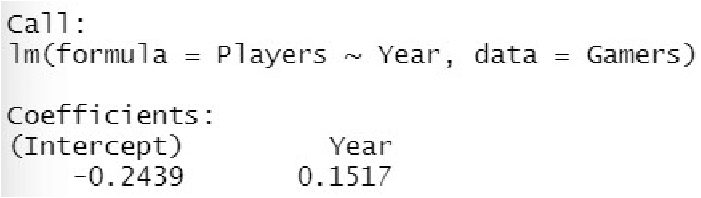
4. Use the regression function to draw the regressionsline.

5. Use the abline function to print the values.

6. Use the regression_Gamers to print the values.

- We have constructed graph as followed:





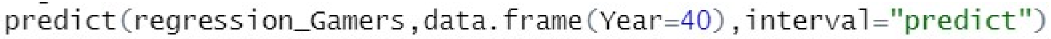
```
Call:
lm(formula = Players ~ Year, data = Gamers)

Coefficients:
(Intercept)         Year
    -0.2439       0.1517
```

- Our function becomes as followed:

F(x)=-0,24+0,15*years.

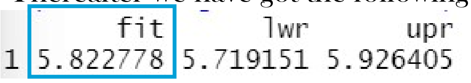
7.Use the predict-function and conclude number of players across the globe in 2040.

- As shown on the picture, we can see that we've written the prediction function.



```
predict(regression_Gamers,data.frame(Year=40),interval="predict")
```

-Thereafter we have got the following values.



```
       fit      lwr      upr
1 5.822778 5.719151 5.926405
```

-We can in this case see, that the value from the predict-function shows us the exact value for the year 2040, which marked with blue.

8.Conclude and Predict

-After getting the values from the predict function, we can say that in the year 2040, there would be approximately 5,88 billion gamers across the globe. This value fits quite good inside the given interval from the prediction function which is (5,71-5,92).

-But we cannot conclude this as our exact answer, because there is an uncertainty whether the graph continues linearly after 2022 or not?