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

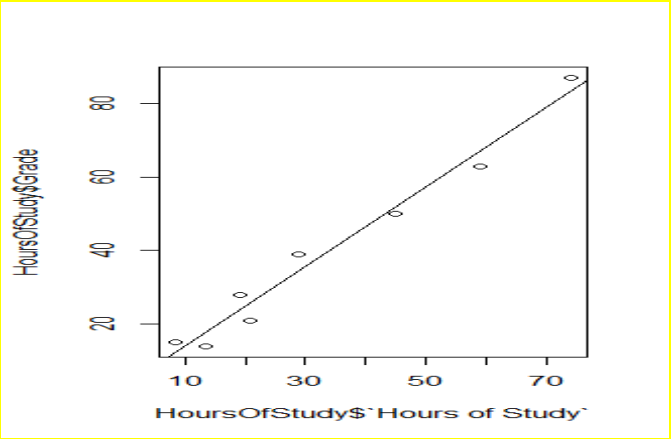
a. Start med, at opskrive de to kolonner i to sektioner i data.frame.

```
"Opgave 1"
Studer <- data.frame(HRS=c(74,59,45,29,20.8,19.1,13.4,8.5),Grade=c(87,63,50,39,21,28,14,15))
cor.test(Studer$HRS,Studer$Grade)
plot(Studer$HRS,Studer$Grade)
regression_Studer <- lm(HRS~Grade,data=Studer)
abline(regression_Studer)
regression_Studer
#Vi kan se, at vi har en signifikant stærk korrelationsværdi
```

b. Derefter brug korrelationstesten, og plotgrafens kommando.

```
        Pearson's product-moment correlation

data:  Studer$HRS and Studer$Grade
t = 15.515, df = 6, p-value = 4.537e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9313803 0.9978695
sample estimates:
      cor
0.9877649
```

c. Lav regressionsanalyse, hvor du efterfølgende danner tendenslinje gennem abline.



d. Derefter udskriv regressionsanalysen og dermed anvend predictions-funktion hvis nødvendigt.

```
> regression_Studer

Call:
lm(formula = HRS ~ Grade, data = Studer)

Coefficients:
(Intercept)        Grade
    -1.9943       0.8983
```

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	Alcohol Consumption, number of glasses	Blood alcohol concentration parts per 1000
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?
b. What is the equation of the regression line?
c. What is the % of the variance in blood alcohol concentration that can be explained by the alcohol consumption?
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?

A. Derefter brug korrelationstesten, og plotgrafens kommando.
B. Brug denne kommando til at kigge på p-værdien og derved sammenligne det med 0,05=95%.

```
"Opgave 2"
Drikke <- data.frame(Number=c(1,2,3,4,5),Blood=c(10,8,12,16,20))
cor.test(Drikke$Number,Drikke$Blood)
plot(Drikke$Number,Drikke$Blood)
regression_Drikke <- lm(Blood~Number,data=Drikke)
abline(regression_Drikke)
regression_Drikke
predict(regression_Drikke,data.frame(Number=4.2),interval="predict")
#A.) Vores forhold er signifikant mellem antal af glas og blodkoncentrationen.
#B.) I tilfældet, kan det ses at vi har funktionen 4,80*x+2,80
#C.) Vi kigger på korrelationskoefficenten og opløfter i anden!
# Svaret giver os 0,844 = 94,4%
#D.) Vi kan se, at vores svar bliver 16,56 ved Blodsukkeret!
```

```
> cor.test(Drikke$Number,Drikke$Blood)

        Pearson's product-moment correlation

data:  Drikke$Number and Drikke$Blood
t = 4.0415, df = 3, p-value = 0.02726
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1950535 0.9947434
sample estimates:
      cor
0.919145
```

- Vi kan i tilfældet se, at vores p-værdi er under 0,05=95%, som betyder at den er signifikant.

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

- For at kunne løse denne opgave, skal vi kigge på billedet foroven især under korrelationskoefficenten. Her skal vi bare kvadrere selve korrelationskoefficenten og derved får vi svaret.

$$0{,}919^2 \approx 0{,}844561$$

- Vi kan dermed se, at selve variansen af blodkoncentrationen er 84,4%.

E. Derefter udskriv regressionsanalysen og dermed anvend predictions-funktion hvis nødvendigt.

```
> regression_Drikke

Call:
lm(formula = Blood ~ Number, data = Drikke)

Coefficients:
(Intercept)       Number
        4.8          2.8

> predict(regression_Drikke,data.frame(Number=4.2),interval="predict")
    fit      lwr      upr
1 16.56 8.476837 24.64316
>
```

- Kigger du på selve Intercept-værdien, så er det svarende til vores a-værdi. Hvorimod Number er svarende til vores b-værdi. Derfor kan opstille følgende formel.
- Teknisk set besvarer vi del B, her men formlen er: $y = 4{,}8 + 2{,}8 \cdot x$

- Hvorimod hvis vi indsætter værdien 4,2 genstande på x's plads, så får vi følgende resultat som står i vores fit, under predict.
- Du kan se, at selve fit passer inde i den øvre grænse og nedre grænse ☺

a. 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?

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

A. Derefter brug korrelationstesten, og plotgrafens kommando.

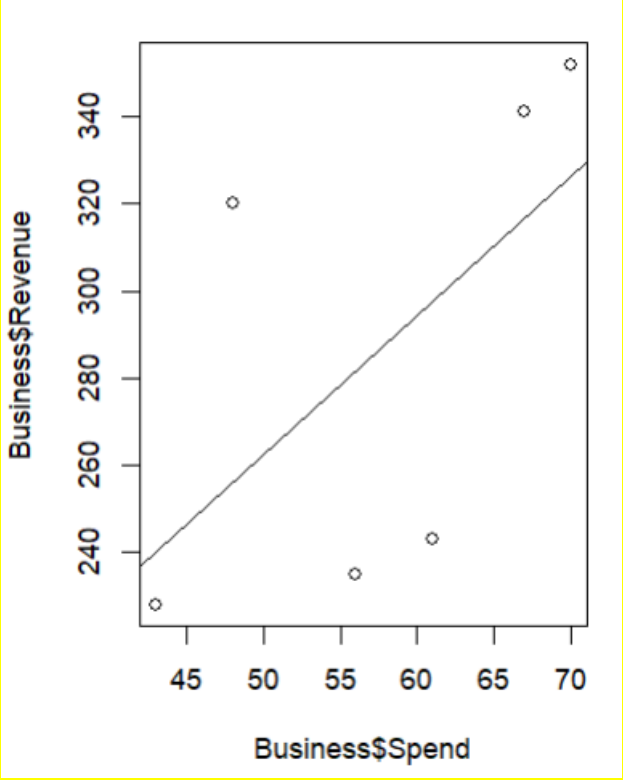
```
"Opgave 2"
Drikke <- data.frame(Number=c(1,2,3,4,5),Blood=c(10,8,12,16,20))
cor.test(Drikke$Number,Drikke$Blood)
plot(Drikke$Number,Drikke$Blood)
regression_Drikke <- lm(Blood~Number,data=Drikke)
abline(regression_Drikke)
regression_Drikke
predict(regression_Drikke,data.frame(Number=4.2),interval="predict")
#A.) Vores forhold er signifikant mellem antal af glas og blodkoncentrationen.
#B.) I tilfældet, kan det ses at vi har funktionen 4,80*x+2,80
#C.) Vi kigger på Korrelationskoefficenten og opløfter i anden!
# Svaret giver os 0,844 = 94,4%
#D.) Vi kan se, at vores svar bliver 16,56 ved Blodsukkeret!
```

B. Brug denne kommando til at kigge på p-værdien og derved sammenligne det med 0,05=95%.

```
> cor.test(Business$Spend,Business$Revenue)

        Pearson's product-moment correlation

data:  Business$Spend and Business$Revenue
t = 1.4663, df = 4, p-value = 0.2164
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4235098  0.9479547
sample estimates:
      cor
0.5912752
```

C. Lav regressionsanalyse, hvor du efterfølgende danner tendenslinje gennem abline.

```
> regression_Business <- lm(Revenue~Spend,data=Business)
> abline(regression_Business)
> regression_Business

Call:
lm(formula = Revenue ~ Spend, data = Business)

Coefficients:
(Intercept)        Spend
    103.043        3.191
```



- Vi kan se, at vi får følgende funktionsforskrift: $y = 103{,}043 + 3{,}191 \cdot x$

D. Derefter udskriv regressionsanalysen og dermed anvend predictions-funktion hvis nødvendigt.

```
> predict(regression_Business,data.frame(Spend=50),interval="predict")
       fit      lwr      upr
1 262.5708 101.4464 423.6951
>
```

b. 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.

- I tilfældet kan det ses at fordi en værdi rejser sig op ad og den anden følger med s kan vi derfor sige at der er snak om en korrelation og ikke om kausation!

c. 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

A. Start med, at opskrive de to kolonner i to sektioner i data.frame.

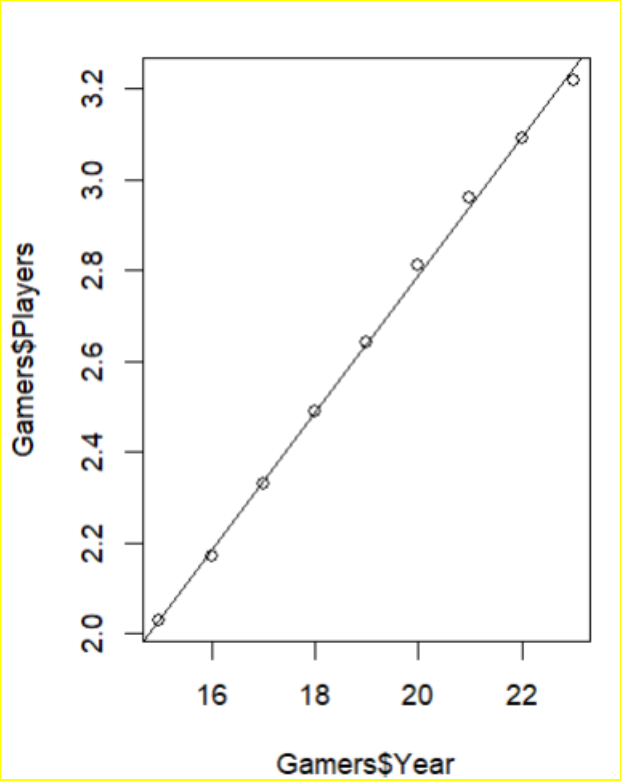
```
"Opgave 5"
Gamers <- data.frame(Year=c(15,16,17,18,19,20,21,22,23),Players=c(2.03,2.17,2.33,2.49,2.64,2.81,2.96,3.09,3.22))
cor.test(Gamers$Year,Gamers$Players)
plot(Gamers$Year,Gamers$Players)
regression_Gamers <- lm(Players~Year,data=Gamers)
abline(regression_Gamers)
regression_Gamers
predict(regression_Gamers,data.frame(Year=40),interval="predict")
#Vi kan se,at vi har en stærk positiv korrelationsværdi signifkant :)
#Vi kan se, at vores funktionsforskrift bliver -0,24+0.15*x
#Vi kan se, at vores værdi i 2040 bliver til 5,82 som befinder sig mellem lwr og upr :)
```

B. Derefter brug korrelationstesten, og plotgrafens kommando.

```
> cor.test(Gamers$Year,Gamers$Players)

        Pearson's product-moment correlation

data:  Gamers$Year and Gamers$Players
t = 77.977, df = 7, p-value = 1.501e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9971538 0.9998839
sample estimates:
      cor
0.9994249
```

C. Lav regressionsanalyse, hvor du efterfølgende danner tendenslinje gennem abline.

```
> regression_Gamers <- lm(Players~Year,data=Gamers)
> abline(regression_Gamers)
> regression_Gamers

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

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



D. Derefter udskriv regressionsanalysen og dermed anvend predictions-funktion hvis nødvendigt.

```
> predict(regression_Gamers,data.frame(Year=40),interval="predict")
       fit      lwr      upr
1 5.822778 5.719151 5.926405
>
```