Statistisk Dataanalyse 2023

Instruktør: Vivek Misra

Pointgivende Aktiviteter

- **I har 2 Pointgivende Aktiviteter i faget Statistisk Dataanalyse.**
  - Hvert Pointgivende Aktivitet tæller 5% og bidrager samlet til 10% i den endelige karakter, da der er 2 tests.
  - **FORDELE:**
    - Hjælper den enkelte studerende med at forhøje karakteren.
    - Hjælper i ”værste tilfælde” med at bestå, hvis studerende ligger på overkanten til at dumpe.
    - Hjælper med at give mavefornemmelse over Eksamensspørgsmål som kommer til faget.
  - **ULEMPE:**
    - Studerende har mindre chance for at forhøje karakteren og skal gøre ekstra indsats til endelig eksamen.
    - Studerende har mindre chance for at blive reddet til den endelig eksamen, hvis dumpet.
  - **DER ER INGEN RE-PRØVER I POINTGIVENDE AKTIVITET**
    - Deltager man ikke testen, afholdes testen IKKE igen da det afholdes engang i specifikke tider.

Snyd og Plagiat!

- **DET ER STRENGT FORBUDT, AT SNYDE TIL POINTGIVENDE AKTIVITETER, TESTS, EKSAMEN OG RAPPORT ETC.**
  - **BRUGEN AF CHATGPT OG ANDRE AI ER STRENGT FORBUDT FRA SDU'S SIDE TIL EKSAMENS AKTIVITETER! (WARRANT).**
  - I MÅ HELLER **IKKE** BRUGE VIMIS22.GITHUB.IO TIL POINTGIVENDE AKTIVITETER OG EKSAMEN.
    - JEG DEAKTIVERER SIDEN, SÅLEDES AT I IKKE KAN KOMME IND. DET ER JER, DER ER ANSVARLIG FOR BRUG AF JERES EGEN HJÆLPEMIDLER!
  - **NOTE: Fejler man Pointgivende Aktivitet eller klarer sig dårligt. Vær ikke bange!**
    - **Se det som en mulighed for forbedring, og kom ENDELIG gerne til Øvelsestimerne og få spurgt om tingene.**
  - **RESULTAT PÅ POINTGIVENDE AKTIVITET, KOMMER STRAKS EFTER AFSLUTNING AF PRØVEN.**
    - **AFHOLDELSE: PÅ ITSLEARNING.COM VED FORELÆSNINGSLOKALET**
  - **KOMMUNIKATION IKKE TILLADT PÅ TVÆRS AF ELEVER, HVERKEN VERBALT ELLER GENNEM INTERNETTET. DETTE SES OGSÅ SOM SNYD UNDER POINTGIVENDE AKTIVITET & EKSAMEN.**
  - **ANBEFALING TIL PA: Er du tvivl om et spørgsmål, gå med første mavefornemmelse.**
    - **VEND TILBAGE TIL DEN TVIVLENDE SPØRGSMÅL, NÅR DU HAR LØST DE ANDRE OPGAVER!**
  - **PS: Hvis du er Ordblind, så kontakt venligst lektoren eller sekretæren for ekstra tid!**

Exercise Class NR10

Solutions to the Tasks

Task 1 – Description

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

Task 1 – Solution

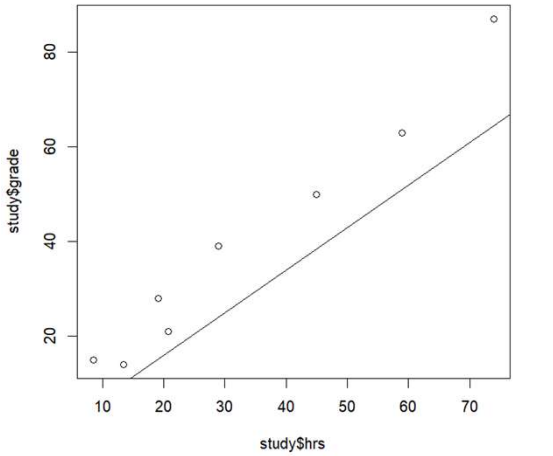
- First we will create the data.frame, where the values are inserted inside.
- Then we will use the correlations-coefficent, plot and regressionsmethod to analyse the dataframe.

```
study <- 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(study$hrs,study$grade)
plot(study$hrs,study$grade)
regression_study <- lm(hrs~grade,data=study)
abline(regression_study)
regression_study
```

Task 1 – Solution

```
> study <- 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(study$hrs,study$grade)

        Pearson's product-moment correlation

data:  study$hrs and study$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
```

```
> plot(study$hrs,study$grade)
> regression_study <- lm(hrs~grade,data=study)
> abline(regression_study)
> regression_study

Call:
lm(formula = hrs ~ grade, data = study)

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



Task 2 – Description

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?

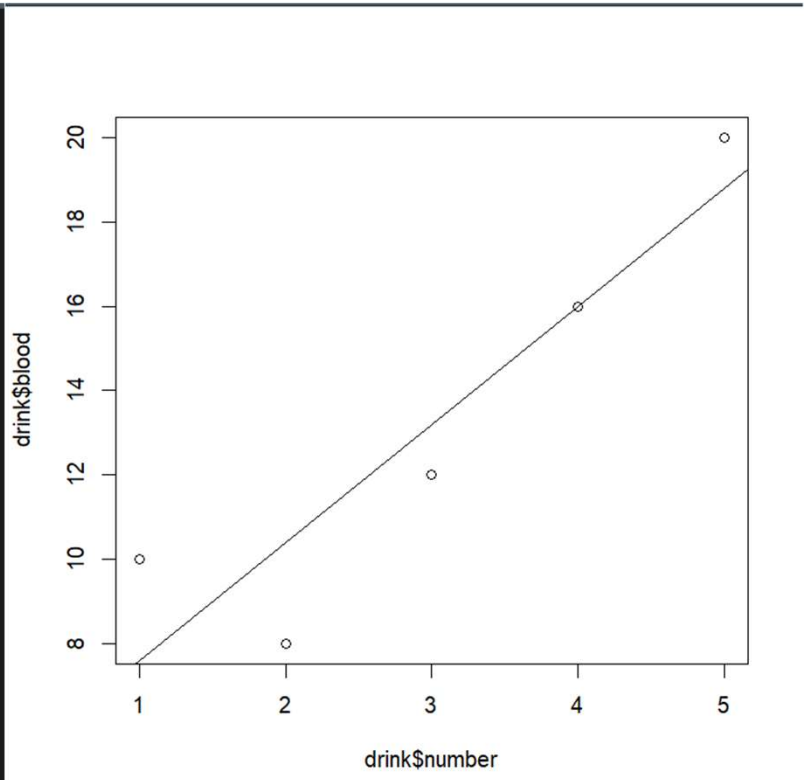
Task 2 – Solution

- First we will create the data.frame, where the values are inserted inside.
- Then we will use the correlations-coefficent, plot and regressionsmethod to analyse the dataframe.
- Then we will use the predict-function to calculate the y-value.

```
drink <- data.frame(number=c(1,2,3,4,5),blood=c(10,8,12,16,20))
cor.test(drink$number,drink$blood)
plot(drink$number,drink$blood)
regression_drink <- lm(blood~number,data=drink)
abline(regression_drink)
regression_drink
predict(regression_drink,data.frame(number=4.2),interval="predict")
```

Task 2 – Solution

```
> cor.test(drink$number,drink$blood)

        Pearson's product-moment correlation

data:  drink$number and drink$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

> plot(drink$number,drink$blood)
> regression_drink <- lm(blood~number,data=drink)
> abline(regression_drink)
> regression_drink

Call:
lm(formula = blood ~ number, data = drink)

Coefficients:
(Intercept)       number
        4.8          2.8

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



Task 2 – Solution

- 2.A: Yes there is a significant relationsship, because the p-value is under 0,05, which 0,02.
- 2.B: The regression equation is: 4,8x+2,8
- 2.C: The variance is 0,91*0,91=0,844
- 2.D: The predicted value of blood alcohol concentration is 16,86.

Task 3 – Description

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?

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

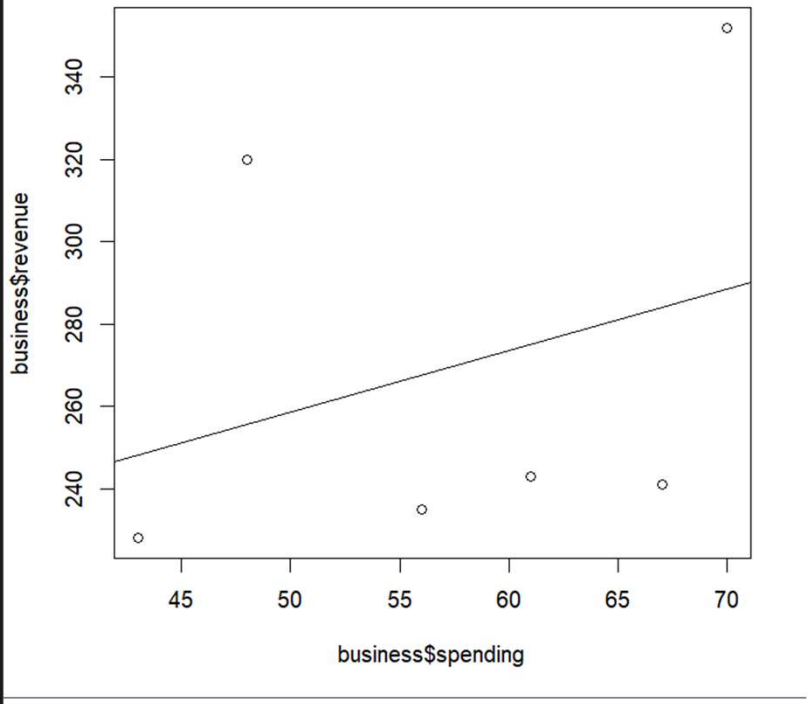
Task 3 – Solution

- First we will create the data.frame, where the values are inserted inside.
- Then we will use the correlations-coefficent, plot and regressionsmethod to analyse the dataframe.
- Then we will use the predict-function to calculate the y-value.

```
business <- data.frame(spending=c(43,48,56,61,67,70),revenue=c(228,320,235,243,241,352))
cor.test(business$spending,business$revenue)
plot(business$spending,business$revenue)
regression_business <- lm(revenue~spending,data=business)
abline(regression_business)
regression_business
predict(regression_business,data.frame(spending=50),interval="predict")
```

Task 3 – Solution

```
> cor.test(business$spending,business$revenue)

        Pearson's product-moment correlation

data:  business$spending and business$revenue
t = 0.63471, df = 4, p-value = 0.5601
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6747042  0.8944687
sample estimates:
      cor
0.3024893

> plot(business$spending,business$revenue)
> regression_business <- lm(revenue~spending,data=business)
> abline(regression_business)
> regression_business

Call:
lm(formula = revenue ~ spending, data = business)

Coefficients:
(Intercept)     spending
    183.660        1.499

> predict(regression_business,data.frame(spending=50),interval="predict")
       fit      lwr      upr
1 258.5934 83.74861 433.4381
```



Task 3 – Solution

- We can see, that the regression is not significant.
- And along with this, it has a weak correlationscoefficent, which is 0,30.
- We can see, that the predict value for spending 50 million in revenue is equal to 258,5934 danish krones.

Task 4 – Description / Solution

4. 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.

- This is clearly a correlation, because the text is telling us about the increase in the number of people who are exercising in Denmark and the increase in the number of people who are commiting crimes in US.
  - This tells us about the relationsship between an independent - and a dependent variable.

Task 5 – Description

5. 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

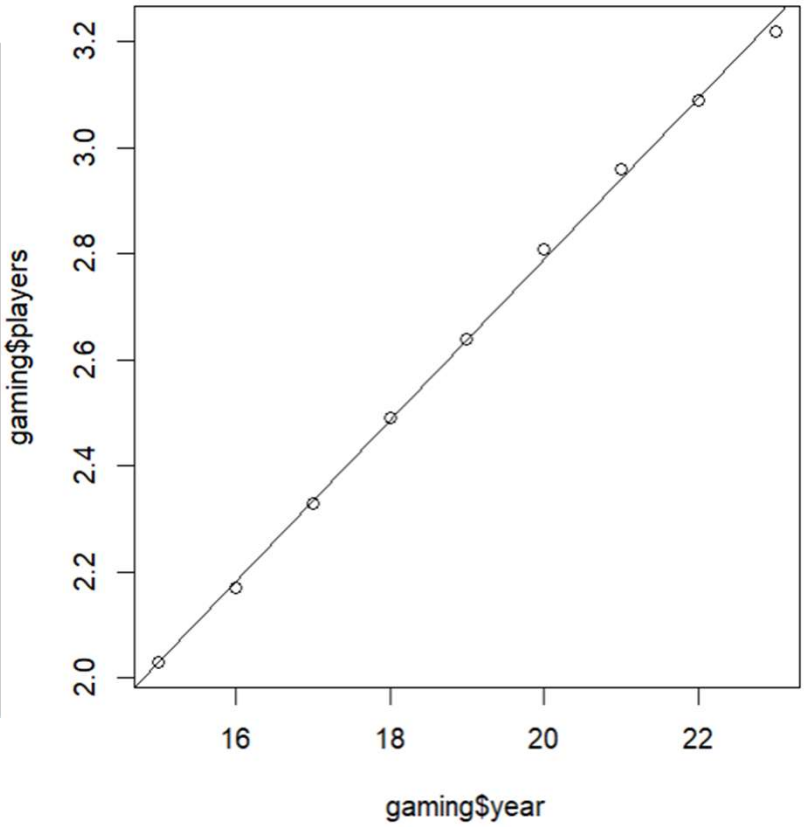
Task 5 – Solution

- First we will create the data.frame, where the values are inserted inside.
- Then we will use the correlations-coefficent, plot and regressionsmethod to analyse the dataframe.
- Then we will use the predict-function to calculate the y-value.

```
gaming <- 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(gaming$year,gaming$players)
plot(gaming$year,gaming$players)
regression_gaming <- lm(players~year,data=gaming)
abline(regression_gaming)
regression_gaming
predict(regression_gaming,data.frame(year=40),interval="predict")
```

Task 5 – Solution

```
> gaming <- 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(gaming$year,gaming$players)

        Pearson's product-moment correlation

data:  gaming$year and gaming$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

> plot(gaming$year,gaming$players)
> regression_gaming <- lm(players~year,data=gaming)
> abline(regression_gaming)
> regression_gaming

Call:
lm(formula = players ~ year, data = gaming)

Coefficients:
(Intercept)         year
    -0.2439       0.1517

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



Task 5 – Solution

- We can see, that the regression is statistically significant.
- We can see, that the graph has a strong correlation coefficent.
- We can see, that in 2040 there are going to 5,82 billion global-players.

Tak for i dag!

Instruktør: Vivek Misra