Statistisk Dataanalyse 2023

Instruktør: Vivek Misra

Task 1 - Description

1) You are a researcher interested in social factors that influence heart disease. You survey 15 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Town	Heart.disease	Smoking	Biking
A	2.9	69.4	2.8
B	3.1	65.7	13.8
C	4.1	54.4	9.1
D	6.4	65.1	2.2
E	6.7	55.9	25.1
F	6.8	51.8	11.0
G	8.6	53.1	26.3
H	8.6	62.8	16.0
I	9.6	48.8	17.6
J	12.1	35.3	14.4
K	15.9	4.8	29.3
L	14.2	2.0	10.0

a) What is the dependent variable and the independent variables in this study? What would you expect about the relationship between the dependent variable and each of the independent variables?
b) Determine the regression line for the model and the corresponding R^2.
c) Are all independent variables significant to the model? Consider a 95% confidence level.

Task 1.A & 1.B - Solution

Town	Heart.disease	Smoking	Biking
A	2.9	69.4	2.8
B	3.1	65.7	13.8
C	4.1	54.4	9.1
D	6.4	65.1	2.2
E	6.7	55.9	25.1
F	6.8	51.8	11.0
G	8.6	53.1	26.3
H	8.6	62.8	16.0
I	9.6	48.8	17.6
J	12.1	35.3	14.4
K	15.9	4.8	29.3
L	14.2	2.0	10.0

- Task 1.A - As seen on the table to the right, we can say:
  - Heart-disease is a dependent variable, because it is the one which is resulted from different factors.
  - Smoking and Biking are independent variables, which have a contribute to heart-disease.
- Task 1.B – Here is the following code which I have written.
  - First we will create a dataframe, due to the reason we dont have an import file.
  - Then we will use the regression-command so that we can create relation between the independent and dependent variables.
  - Thereafter we will print the values.

Task 1.B - Solution

- Here we have the following command:

```
sickness <-
data.frame(Town=c('A','B','C','D','E','F','G','H','I','J','K','L'),Disease=
c(2.9,3.1,4.1,6.4,6.7,6.8,8.6,8.6,9.6,12.1,15.9,14.2),Smoking=c(69.4,65.7,5
4.4,65.1,55.9,51.8,53.1,62.8,48.8,35.3,4.8,2.0))

sickness$Town <- as.factor(sickness$Town)

sickness$Biking <- as.numeric(sickness$Disease)

sickness$Smoking <- as.numeric(sickness$Smoking)

regression_sickness <- lm(Disease~Biking+Smoking,data=sickness)

summary(regression_sickness)
```

Task 1.B & Task 1.C - Solution

- Here we can see, that we have the following results:

```
Call:
lm(formula = Disease ~ Biking + Smoking, data = sickness)

Residuals:
       Min         1Q     Median         3Q        Max
-2.182e-15 -5.772e-17  2.185e-17  2.526e-16  1.263e-15

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) -3.077e-15  2.409e-15 -1.277e+00    0.234
Biking       1.000e+00  1.443e-16  6.927e+15   <2e-16 ***
Smoking      2.543e-17  2.675e-17  9.510e-01    0.367
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.765e-16 on 9 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.244e+32 on 2 and 9 DF,  p-value: < 2.2e-16
```

- Task 1.C :
  - We can see, that Intercept is not significant.
  - Biking and Smoking are both significant, which means that they contribute to the heart-disease.

Task 2 - Description

2) A health insurance company was hired to provide a better overview of the healthcare expenses associated with hospitalization of patients in Denmark. The company has therefore collected data of 138 patients, who were admitted to different hospitals located in three different Danish regions. The data collected is found in healthcare.xlsx. A description of each variable is found in another tab of the same spreadsheet.
   a) The employees from the health insurance company have hypothesized, from the beginning, that the treatment cost (TREATCOST) can be predicted by the number of days the patient has been admitted to the hospital (CAREDAYS). Build a simple linear regression model to investigate this relationship. How much of the variation in TREATCOST can be explained by the variation in CAREDAYS?
   b) An employee raised the hypothesis that the treatment cost will also be affected by the region in which the patient was hospitalized. Develop a regression model where you include both CAREDAYS and REGION as independent variables. On average, how much will the treatment cost increase/decrease if the patient is hospitalized one day more? How much will the treatment cost increase/decrease if the patient was hospitalized in the region of Syddanmark in comparison to being hospitalized in the Capital region (Hovedstaden)?

Task 2.A - Solution

- Here we have the following command:

```
"Opgave 2.a"
regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
summary(regression_Healthcare)
#Vores intercept a-værdi er ikke signifikant.
#Vores caredays er signifikant!
```

- Here are the results:

```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
> summary(regression_Healthcare)

Call:
lm(formula = TREATCOST ~ CAREDAYS, data = Healthcare)

Residuals:
   Min     1Q Median     3Q    Max
-88247 -16660  -3700  12419 221222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     6463       5927   1.091    0.277
CAREDAYS       16572       1169  14.174   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35460 on 136 degrees of freedom
Multiple R-squared:  0.5963,    Adjusted R-squared:  0.5933
F-statistic: 200.9 on 1 and 136 DF,  p-value: < 2.2e-16
```

Task 2.B & 2.C - Solution

- Here we have the following command:

```
"Opgave 2.b"
regression_Healthcare <- lm(TREATCOST~CAREDAYS+REGION,data=Healthcare)
summary(regression_Healthcare)
#Intercept og Caredays er de samme!
#Region er ikke signifikant!
```

- Here are the results:

```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS+REGION,data=Healthcare)
> summary(regression_Healthcare)

Call:
lm(formula = TREATCOST ~ CAREDAYS + REGION, data = Healthcare)

Residuals:
   Min     1Q Median     3Q    Max
-83258 -15901  -4006  13782 227935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27323      11521   2.372   0.0191 *
CAREDAYS       14848       1413  10.508   <2e-16 ***
REGION2       -15238       8773  -1.737   0.0847 .
REGION3       -20064       9543  -2.102   0.0374 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35140 on 134 degrees of freedom
Multiple R-squared:  0.6095,    Adjusted R-squared:  0.6008
F-statistic: 69.72 on 3 and 134 DF,  p-value: < 2.2e-16
```

- Task 2.B:
  - What is Significant and what is not?
    - Intercept is Significant.
    - Caredays is Significant.
    - Region2 is not Significant.
    - Region3 is not Significant.

Task 2.B & 2.C - Solution

- Here we have the following command:

```
"Opgave 2.c"
regression_Healthcare <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+STATUS+CAREDAYS+INTENSIVEDAYS+AGE+SEX+INSURANCE+REGION,data=Healthcare)
summary(regression_Healthcare)
#I tilfældet, kan det ses at de fleste er ikke signifikante, undtagen Intercept og Laboratory
```

- Here are the results:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9259.7321 17059.8659   0.543 0.588249
MEDICINE        2.3546     0.7592   3.101 0.002382 **
LAB             1.5984     0.4292   3.724 0.000295 ***
XRAY            1.5215     0.3246   4.688 7.11e-06 ***
INHALATOR       1.7714     0.2984   5.936 2.69e-08 ***
STATUS1     -3027.9156  7058.7724  -0.429 0.668692
CAREDAYS     5019.6769  1106.0858   4.538 1.31e-05 ***
INTENSIVEDAYS 2933.1817 2635.3474   1.113 0.267838
AGE           -49.3816   180.3819  -0.274 0.784720
SEX1         2218.6264  3292.7592   0.674 0.501689
INSURANCE   -7079.7291  5987.7084  -1.182 0.239300
REGION2      5778.8742  4953.5438   1.167 0.245587
REGION3      3486.2600  5444.1373   0.640 0.523104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18820 on 125 degrees of freedom
Multiple R-squared:  0.8954,    Adjusted R-squared:  0.8854
F-statistic:  89.2 on 12 and 125 DF,  p-value: < 2.2e-16
```

- Task 2.B:
  - What is Significant and what is not?
    - Intercept is Significant.
    - Caredays is Significant.
    - Region2 is not Significant.
    - Region3 is not Significant.

Task 3 - Description

3) A real estate agent wants to better understand what are the factors that influence the price of houses sold in the region of greater Copenhagen. For that, she hires a group of statisticians and provides data on house age in years (X1), distance to public transportation in meters (X2), number of convenience stores (X3), house condition, where 1 = poor, 2 = medium, 3 =high (X4), and house price of unit area in 1,000 DKK (Y) for 413 houses. The data is available in the file "real_estate.txt".

a) Determine the appropriate regression model.

b) With the regression line equation you found in item a), predict the house price of unit area for a house that is 10 years old, is 1 km from public transport, has 1 convenience store close by and is in a high level condition.

c) Repeat item b) but using the predict() function in R. What is the estimated predicted house value?

Task 3.A - Solution

- Here we have the following command:

```
"Opgave 3.a"
regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
summary(regression_real_estate)
#I tilfældet,kan det ses at alle p-værdier fra Intercept til X4 er signifikante
"Opgave 3.b"
#Bemærk, at vi har omregnet fra 1 km til 1000 meter ved X2.
#Fordi vores R ikke virker, så kan vi bruge formlen til Eksamen!
17.89-0.12*X1-0.002*X2+0.56*X3+12.04*X4
17.89-0.12*10-0.002*1000+0.56*1+12.04*3
#Vi kan se at den forudsagte værdi bliver 51,37
"Opgave 3.c"
predict(regression_real_estate,data.frame(X=10,X2=1000,X3=1,X4=3))
```

- Here are the results:
  - Please look to the right:

```
> regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
> summary(regression_real_estate)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = real_estate)

Residuals:
    Min      1Q  Median      3Q     Max
-27.540  -3.304  -0.494   2.774  64.808

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.8976989  1.6576646  10.797  < 2e-16 ***
X1          -0.1204197  0.0300323  -4.010 7.23e-05 ***
X2          -0.0024376  0.0003643  -6.691 7.33e-11 ***
X3           0.5619746  0.1467081   3.831 0.000148 ***
X4          12.0496250  0.6320173  19.065  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.74 on 409 degrees of freedom
Multiple R-squared:  0.757,     Adjusted R-squared:  0.7546
F-statistic: 318.6 on 4 and 409 DF,  p-value: < 2.2e-16
```

Task 3.B & 3.C - Solution

- Here we can see the result for the predicted value:
  - Remind you, that there have been difficulties in using the predict function.
  - So we have used the lineare formula to create the prediction equation.

```
17.89-0.12*X1-0.002*X2+0.56*X3+12.04*X4
17.89-0.12*10-0.002*1000+0.56*1+12.04*3
#Vi kan se at den forudsagte værdi bliver 51,37
"Opgave 3.c"
predict(regression_real_estate,data.frame(X=10,X2=1000,X3=1,X4=3))
```

Task 4 - Description

4) A car dealer has collected data on all used cars he sold within last year. He makes a dataset called car.txt with the information collected for all 301 cars.
The dataset contains the following variables:

Selling_Price: Price in which the car is being sold (in Euros).
Original_Price: Price when the car was first bought (in Euros).
Kms_Driven: Number of kilometers the car is driven.
Fuel_Type: Fuel type of car (Petrol/Diesel).
Transmission: Gear transmission of the car (Automatic/Manual).

Based on the given data, use the appropriate regression model to predict the selling price of a car that was originally bought by 7500 Euros, it was driven for 8000 kilometers, uses petrol and has an automatic gear.

Task 4- Solution

- Here we have the following command:

```
"Opgave 4 og Opgave 5"
regression_car <- lm(Selling_Price~Original+Kms_Driven,data=car)
summary(regression_car)
predict(regression_car,data.frame(Original_Price=7500,Kms_Driven=8000))
```

Task 5 - Description

5) Using the same context from exercise 4, which statement is correct in relation to the model you developed to predict cars' selling price?
   a) Increasing the original price of the car in 1 euro will result in a decrease of 469 euros in the car's selling price (considering all other variables are fixed).
   b) Increasing the original price of the car in 1 euro will result in an increase of 1.59 euros in the car's selling price (considering all other variables are fixed).
   c) The selling price of a car that is run by diesel is estimated to be 1619 Euros lower than a car run by petrol (considering all other variables are fixed).
   d) The selling price of a car that has manual gear is estimated to be 1589 Euros lower than a car that has automatic gear (considering all other variables are fixed).

Task 4- Solution

- Here we have the following result of the prediction:
  - Remind you, that our Predict-Function is not working due to technical issues.

```
Residuals:
     Min       1Q   Median       3Q      Max
-14330.4   -898.3   -365.0    788.0  12388.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.331e+03  2.058e+02   6.466 4.11e-10 ***
Original_Price  5.356e-01  1.571e-02  34.084  < 2e-16 ***
Kms_Driven     -2.043e-02  3.493e-03  -5.849 1.30e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2303 on 298 degrees of freedom
Multiple R-squared:  0.796,     Adjusted R-squared:  0.7947
F-statistic: 581.5 on 2 and 298 DF,  p-value: < 2.2e-16
```

```
predict(Regression_car,data.frame(Original_Price=7500,Kms_Driven=8000))
```

Tak for denne Semester!

Held og Lykke med Eksamen, og håber vi mødes i fremtiden ☺

Instruktør: Vivek Misra