LESSON 12. MULTIPLE LINEAR REGRESSION

Class content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- Types of variables
- Multiple regression in R
- Prediction

Exercises:

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?**

- Remember these notations: Independent Variable is something which is contributing to make something, whereas if we look closer towards the Dependent variable then it is dependent on the input from the Independent Variable.
- **Independent:** Smoking and Biking are Independent Variables.
- **Dependent:** Heart-Disease is Dependent variable because it requires input from Smoking and Biking.
- **Index:** It is the towns, which are acting as segmentation.

b) Determine the regression line for the model and the corresponding R^2.

Solution:

1. Start by using the data.frame function, to construct a dataframe.

```
#Opgave 1.b
Sygdom <- 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(69.4,65.7,54.4,65
```

2. Now, use the Regression function.

```
Regression_Sygdom <- lm(Disease~Biking+Smoking, data=Sygdom)
```

3. Now, create the summary.

```
summary(Regression_Sygdom)
```

c) Are all independent variables significant to the model? Consider a 95% confidence level.

Solution:

1. Look at the p-value and compare them with 0,05=95%.

- Intercept: 4,99e-05 is smaller than 0,05 and that is significant.
- Biking: 0,22 is bigger than 0,05 and that is not significant.
- Smoking: 0,00027 is smaller than 0,05 and that is significant.

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?

Solution:

1. Start by using the regression function.

```
#Opgave 2.a
Regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
```

2. Now use the summary to find the p-values.

```
summary(Regression_Healthcare)
```

```
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
```

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)?

Solution:

1. Add +REGION in the Regression_Healthcare.

```
#Opgave 2.a
Regression_Healthcare <- lm(TREATCOST~CAREDAYS+REGION,data=Healthcare)
```

2. Use the summary function to find the values.

```
summary(Regression_Healthcare)
```

```
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
```

c) Expand the analysis done in item a) and b) and include all the other independent variables in the analysis (remember to check if the variables are correctly recognized in R). Which variables are significant to the model (95% confidence level)?

Solution:

1. Add the other variables and EXCLUDE the variable Index.

```
#Opgave 2.a
Regression_Healthcare <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+STATUS+CAREDAYS+INTENSIVEDAYS+AGE+SEX+INSURANCE+REGION,data=Healthcare)
```

2. Use the summary function.

```
summary(Regression_Healthcare)
```

```
(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
INSURANCE1    -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
```

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.

Solution:

1. Use the function of Regression and place the Y before the ~(tilde sign).

```
#Opgave 3.a
Regression_real_estate <-lm(Y~X1+X2+X3+X4, data=real_estate)
```

2. Use the summary function.

```
summary(Regression_real_estate)


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
```

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.

Solution:

1. Use the following equation and insert the values.

$$HouseSold = 17{,}89 - 0{,}12 \cdot X1 - 0{,}002 \cdot X2 + 0{,}56 \cdot X3 + 12{,}04 \cdot X4$$

2. Convert 1 km to 1 meter, and write the values which match X1, X2, X3 and X4.

$$X1 = Unit\ Area\ for\ House = 10$$

$$X2 = Public\ Transport = 1\ km = 1000\ meter$$

$$X3 = Convenience\ Store = 1$$

$$X4 = High = 3$$

3. Now, insert the values and thereafter

$$HouseSold = 17{,}89 - 0{,}12 \cdot 10 - 0{,}002 \cdot 1000 + 0{,}56 \cdot 1 + 12{,}04 \cdot 3 = 51{,}37$$

4. Conclude the answer

- We can in the end conclude, that the predicted value would become 51,37.

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

Solution:

1. Use the predicted function.

```
predict(Regression_real_estate,data.frame(X1=10,X2=1,X3=1,X4=1000))
```

2. Conclude from the Predicted Function.

```
> predict(Regression_real_estate,data.frame(X1=10,X2=1,X3=1,X4=1000))
       1
12066.88
>
```

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.

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).

- No

   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).

- No

   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).

- No

   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).

- Yes

Solution:

1. Use the Regression function.

- From the tasks we have learned that the thing we want to research is written before the "tilde-sign".
- Thereafter, the independent variables are written afterwards.

```
#Opgave 4.a
Regression_car <- lm(Selling_Price~Original_Price+Kms_Driven, data=car)
```

2. Use the summary-function to find the significance.

```
summary(Regression_car)


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
```

3. Use the predict function, to insert predict values from Original Price and Kilometers Driven.

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