Lab 5 - Linear regression

Systems modelling and data analysis
2016/2017

1 Preparing the data

1. Run RStudio
2. Set your Working Directory using the setwd() command.
3. Download, extract and then load the data from the file krakow-kurdwanow.zip. The data comes from the monthly reports from 2015 from the Kraków-Kurdwanów station (source: http://monitoring.krakow.pios.gov.pl/).

```
        download.file("http://home.agh.edu.pl/~mmd/_media/dydaktyka/
as-is/krakow-kurdwanow.zip", "krakow-kurdwanow.zip")
unzip("krakow-kurdwanow.zip")
data <- dget("./krakow-kurdwanow")
```

2 Linear regression

1. Create y and x vectors. Assign corresponding data from columns: data$PM25 and data$PM10.

```
y <- data$PM25
x <- data$PM10
```

2. Delete missing data from vectors. If there is no data in the vector y, also delete the corresponding data from the vector x. Similarly, if the vector x is missing data, also delete the corresponding data from the vector y.

```
good <- complete.cases(x,y)
y <- y[good]
x <- x[good]
```

3. Calculate correlation - find out how variables depend on each other linearly.

```
n <- length(x)
l <- (n*sum(x*y)-sum(x)*sum(y))
m <- sqrt((n*sum(x^2) - sum(x)^2) * (n*sum(y^2) - sum(y)^2))
l/m
```

4. Use the cor() function to calculate correlations.

```
cor(x,y)
```

5. Centralize the random variable.

```
ymean <- mean(y)
xmean <- mean(x)


y <- y - ymean
x <- x - xmean
```

6. Make sure the centralization is done correctly.

```
mean(y)
mean(x)
```

7. Create a function that would count the sum of the squares "vertical"distances between the points of the straight line y = ax for a given "a" argument. The function should have the following arguments:

   (a) y-coordinate of the data (vector)

   (b) x-coordinate of the data (vector)

   (c) Parameter a of straight line y = ax, for which the sum of squares is calculated.

```
sum_of_the_squered <- function(y,x,a) {
     sum <- 0
     for(i in seq_along(y)) {
        sum <- sum + (y[i] - (a*x[i]))^2
      }
      sum
   }
```

8. Create a function that will select "a" parameter from the given vector of "a" parameters (a_vector) for which the sum of squares is the smallest. The function should have the following arguments:

   (a) y-coordinate of the data (vector)

   (b) x-coordinate of the data (vector)

   (c) Vector a parameters (a_vector).

```
find_a <- function(y,x,a_vector) {
      min_sum <- Inf
      min_a <- NA
      for(a in a_vector) {
          sum <- sum_of_the_squered(y,x,a)
           if(sum < min_sum) {
                 min_sum <- sum
                 min_a <- a
```

```
            }
        }
        min_a
}
```

9. Use the created function to find the a parameter. To do this, increase the accuracy of the search parameter to 10 decimal places. View the parameter found.

```
a <- 0

for(i in 0:10) {
        a <- find_a(y,x,seq(a-10^(-i+1), a+10^(-i+1), 10^(-i)))
}
a
```

10. Use the lm() function to find the a parameter, and then view the parameter found.

```
model <- lm(y ~ x)
model$coefficients["x"]
```

11. Count and display the differences between the a parameter found with the created function and the lm() function.

```
model$coefficients["x"] - a
```

12. Draw a graph showing the data and linear regression made with the created functions and the lm() function.

```
plot(y~x)
abline(0,a, col="red", lwd=7)
abline(model, col="blue", lwd=3)
```

13. Draw graph showing data and linear regression based on the created functions and the lm() function for non-centralized data.

```
plot(data$PM25~data$PM10)
abline(ymean-a*xmean,a, col="red", lwd=5)
model <- lm(data$PM25 ~ data$PM10)
abline(model, col="white")
```

3 Exercise

1. Use the data from the monthly reports from 2015 from the Kraków-Kurdwanów station (source: http://monitoring.krakow.pios.gov.pl/). Download, extract and then load the data from the file located at: http://home.agh.edu.pl/ mmd/ _media/ dydaktyka/as-is/krakow-kurdwanow.zip

2. Limit the data to: SO2, NO2, PM10, and then delete the missing data rows.
3. Calculate the correlation coefficient for individual data: SO2 NO2 and SO2 PM10 and NO2 PM10.
4. Determine linear regression. To calculate the linear regression, select the best correlated data. Note the data order - place the first value of the pair on the y-axis and place the value given in the second position of the pair on the x-axis.
5. Draw a graph showing the data along with the regression line.
6. Calculate the sum of squares of the "vertical"distances between the points and the straight line y = ax + b for the designated arguments a and b.