# Exercise 2: Analysis

*Describe the work you have done this week and summarize your learning.*

- Describe your work and results clearly.
- Assume the reader has an introductory course level understanding of writing and reading R code as well as statistical methods.
- Assume the reader has no previous knowledge of your data or the more advanced methods you are using.

*Insert all the codes, your interpretations and explanations*

# TASK 1

## Reading the data into R and checking the structure and the dimensions of the data

```
data <- read.table("data/learning2014.txt", sep=" ")
str(data)
```

```
## 'data.frame':    166 obs. of  7 variables:
##  $ Age     : int  53 55 49 53 49 38 50 37 37 42 ...
##  $ Points  : int  25 12 24 10 22 21 21 31 24 26 ...
##  $ gender  : Factor w/ 2 levels "F","M": 1 2 1 2 2 1 2 1 2 1 ...
##  $ attitude: num  3.7 3.1 2.5 3.5 3.7 3.8 3.5 2.9 3.8 2.1 ...
##  $ deep    : num  3.58 2.92 3.5 3.5 3.67 ...
##  $ stra    : num  3.38 2.75 3.62 3.12 3.62 ...
##  $ surf    : num  2.58 3.17 2.25 2.25 2.83 ...
```

```
dim(data)
```

```
## [1] 166   7
```

*The dataset I'm using in these analyses is a subset of the dataset provided by Kimmo Vehkalahti. The original data (N=183) was collected during the course Introduction to Social Sciences at the University of Helsinki in 2014-2015. The study was conducted in Finnish. Using the ASSIST (Approaches and Study Skills Inventory) questionnaire, the participating students were asked about e.g. their learning approaches (deep, strategic, and surface approach). Their global attitude towards statistics was measured using the SATS (Survey of Attitudes Toward Statistics) questionnaire. Each student's age and gender were included, as was their learning achievement measured by points they got in the course exam.*

*For the current analyses, I removed the students who got 0 points in the exam. Thus, the dataset has 166 observations and the 7 variables (Age, Points, gender, attitude, deep, stra, surf) described above.*

# TASK 2

# Graphical overview of the data and a summary of data

I installed the packages ggplot2, GGally and psych.

```
library("ggplot2")
library("GGally")
```

```
## Warning: package 'GGally' was built under R version 3.4.4
```

```
library("psych")
```

```
## Warning: package 'psych' was built under R version 3.4.4
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

## Advanced plot matrix and descriptive statistics:

```
matrix <- ggpairs(data, mapping = aes(col = gender, alpha = 0.3), lower = list(combo = wrap(
"facethist", bins = 20)))
matrix
```


```
summary(data) #all participants
```

```
##       Age            Points      gender     attitude         deep
##  Min.   :17.00   Min.   : 7.00   F:110   Min.   :1.400   Min.   :1.583
##  1st Qu.:21.00   1st Qu.:19.00   M: 56   1st Qu.:2.600   1st Qu.:3.333
##  Median :22.00   Median :23.00           Median :3.200   Median :3.667
##  Mean   :25.51   Mean   :22.72           Mean   :3.143   Mean   :3.680
##  3rd Qu.:27.00   3rd Qu.:27.75           3rd Qu.:3.700   3rd Qu.:4.083
##  Max.   :55.00   Max.   :33.00           Max.   :5.000   Max.   :4.917
##       stra            surf
##  Min.   :1.250   Min.   :1.583
##  1st Qu.:2.625   1st Qu.:2.417
##  Median :3.188   Median :2.833
##  Mean   :3.121   Mean   :2.787
##  3rd Qu.:3.625   3rd Qu.:3.167
##  Max.   :5.000   Max.   :4.333
```

```
describeBy(data, group="gender") #using the psych library to get a summary of the data divide
d by gender
```

```
##
##  Descriptive statistics by group
## group: F
##          vars   n  mean   sd median trimmed  mad   min   max range  skew
## Age         1 110 24.85 7.36  22.00   23.38 2.97 17.00 53.00 36.00  1.82
## Points      2 110 22.33 5.83  23.00   22.61 5.93  7.00 33.00 26.00 -0.36
## gender*     3 110  1.00 0.00   1.00    1.00 0.00  1.00  1.00  0.00   NaN
## attitude    4 110  2.99 0.73   2.95    2.98 0.82  1.40  5.00  3.60  0.14
## deep        5 110  3.66 0.53   3.67    3.68 0.62  1.58  4.75  3.17 -0.55
## stra        6 110  3.20 0.75   3.25    3.22 0.83  1.38  5.00  3.62 -0.15
## surf        7 110  2.83 0.46   2.83    2.82 0.49  1.83  4.00  2.17  0.19
##          kurtosis   se
## Age          2.83 0.70
## Points      -0.26 0.56
## gender*       NaN 0.00
## attitude    -0.52 0.07
## deep         0.90 0.05
## stra        -0.34 0.07
## surf        -0.43 0.04
## ------------------------------------------------------------
## group: M
##          vars  n  mean   sd median trimmed  mad   min   max range  skew
## Age         1 56 26.80 8.43  24.00   25.11 4.45 19.00 55.00 36.00  1.90
## Points      2 56 23.48 6.01  23.50   24.00 6.67  9.00 33.00 24.00 -0.51
## gender*     3 56  2.00 0.00   2.00    2.00 0.00  2.00  2.00  0.00   NaN
## attitude    4 56  3.44 0.65   3.40    3.47 0.59  1.70  4.80  3.10 -0.45
## deep        5 56  3.72 0.59   3.79    3.75 0.49  2.08  4.92  2.83 -0.47
## stra        6 56  2.96 0.80   3.00    2.96 0.93  1.25  4.50  3.25  0.02
## surf        7 56  2.70 0.64   2.62    2.69 0.62  1.58  4.33  2.75  0.29
##          kurtosis   se
## Age          3.06 1.13
## Points      -0.29 0.80
## gender*       NaN 0.00
## attitude     0.48 0.09
## deep         0.19 0.08
## stra        -0.67 0.11
## surf        -0.60 0.09
```

*The majority of the students were female (female n=110, male n=56). Based on the plots, the distributions of age, exam points and deep learning seem very similar in both genders; the majority were young adults (75% were under 27 years old, range 17-55 years), half of the participants got at least 23.0 points from the exam (range 7-33 points), and deep learning strategy was favored (50% scored 3.667 or higher, range 1.583-4.917).*

*As for the between-gender differences, attitude towards statistics seems more positive among male participants (F median = 2.95, M median = 3.40). Female students seem to lean just a bit more towards using strategic learning (F median = 3.25, M median = 3.00), but also towards using surface learning (F median = 2.83, M median = 2.62).*

*Some of the continuous variables seem to correlate with one another: exam points correlate positively with attitude towards statistics (r=0.437) and using strategic learning (r=0.146); and negatively with using surface learning (r=-0.144). Those resorting to surface learning tended to rely less on deep learning (r=-0.324). Interestingly, this kind of relationship was only seen in males (r=-0.622 vs. females r=0.087).*

# TASKS 3 & 4

# Regression model summary

Based on the correlations between the continuous variables, I chose attitude towards statistics, using strategic learning and surface learning as explanatory variables. Exam points was used as the dependent variable.

```
#The regression model:
regr_model <- lm(Points ~ attitude + stra + surf, data = data)
summary(regr_model)
```

```
##
## Call:
## lm(formula = Points ~ attitude + stra + surf, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1550  -3.4346   0.5156   3.6401  10.8952
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0171     3.6837   2.991  0.00322 **
## attitude      3.3952     0.5741   5.913 1.93e-08 ***
## stra          0.8531     0.5416   1.575  0.11716
## surf         -0.5861     0.8014  -0.731  0.46563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.296 on 162 degrees of freedom
## Multiple R-squared:  0.2074, Adjusted R-squared:  0.1927
## F-statistic: 14.13 on 3 and 162 DF,  p-value: 3.156e-08
```

*According to the model, +1 points in the exam corresponds to a) +3.4 points in the attitude towards statistics, b) +0.9 points in strategic learning approach, and c) -0.6 points in surface learning approach.*

*The model is statistically significant (p<0.001, F=14.13, df=162). However, of the three explanatory variables, the attitude variable is the only one with statistical significance (p<0.001). This model covers ~20% of the variance in the exam points, so ~80% of the variance is left "unexplained".*

```
#A new model without surface learning:
regr_model2 <- lm(Points ~ attitude + stra, data = data)
summary(regr_model2)
```

```
##
## Call:
## lm(formula = Points ~ attitude + stra, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6436  -3.3113   0.5575   3.7928  10.9295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.9729     2.3959   3.745  0.00025 ***
## attitude      3.4658     0.5652   6.132 6.31e-09 ***
## stra          0.9137     0.5345   1.709  0.08927 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.289 on 163 degrees of freedom
## Multiple R-squared:  0.2048, Adjusted R-squared:  0.1951
## F-statistic: 20.99 on 2 and 163 DF,  p-value: 7.734e-09
```

*In this model, attitude is strongly significant (p<0.001) and strategic learning non-significant at 0.05 level (p<0.1). Also this model covers ~20% of the variance in the exam points.*

```
#The final model with only one explanatory variable:
regr_model3 <- lm(Points ~ attitude, data = data)
summary(regr_model3)
```

```
##
## Call:
## lm(formula = Points ~ attitude, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9763  -3.2119   0.4339   4.1534  10.6645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.6372     1.8303   6.358 1.95e-09 ***
## attitude      3.5255     0.5674   6.214 4.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.32 on 164 degrees of freedom
## Multiple R-squared:  0.1906, Adjusted R-squared:  0.1856
## F-statistic: 38.61 on 1 and 164 DF,  p-value: 4.119e-09
```

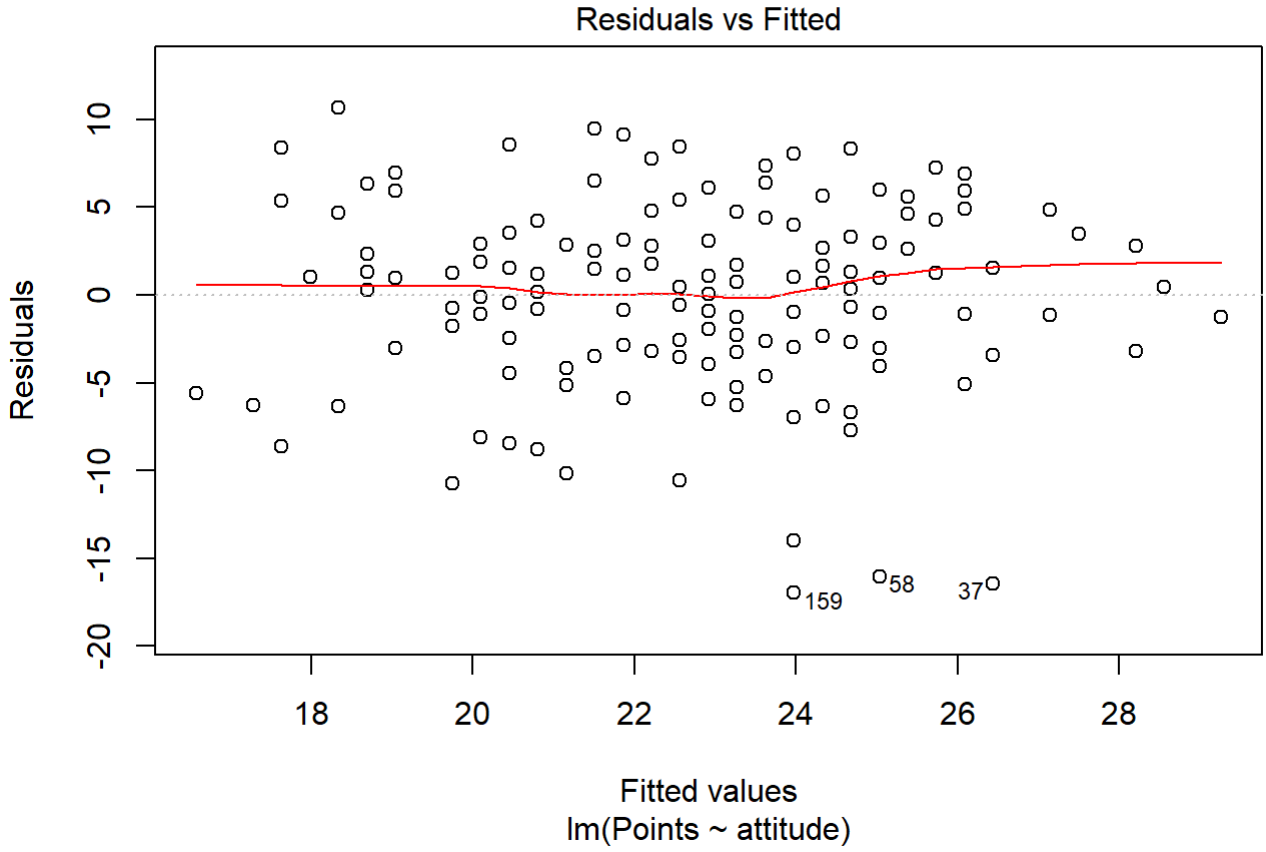
*In this model, +3.5 points in the attitude towards statistics corresponds to +1 points in the exam. With only one explanatory variable, we get a model covering almost 20% of the variance in the exam points. The model is statistically significant (p<0.001, F=38.61, df=164).*

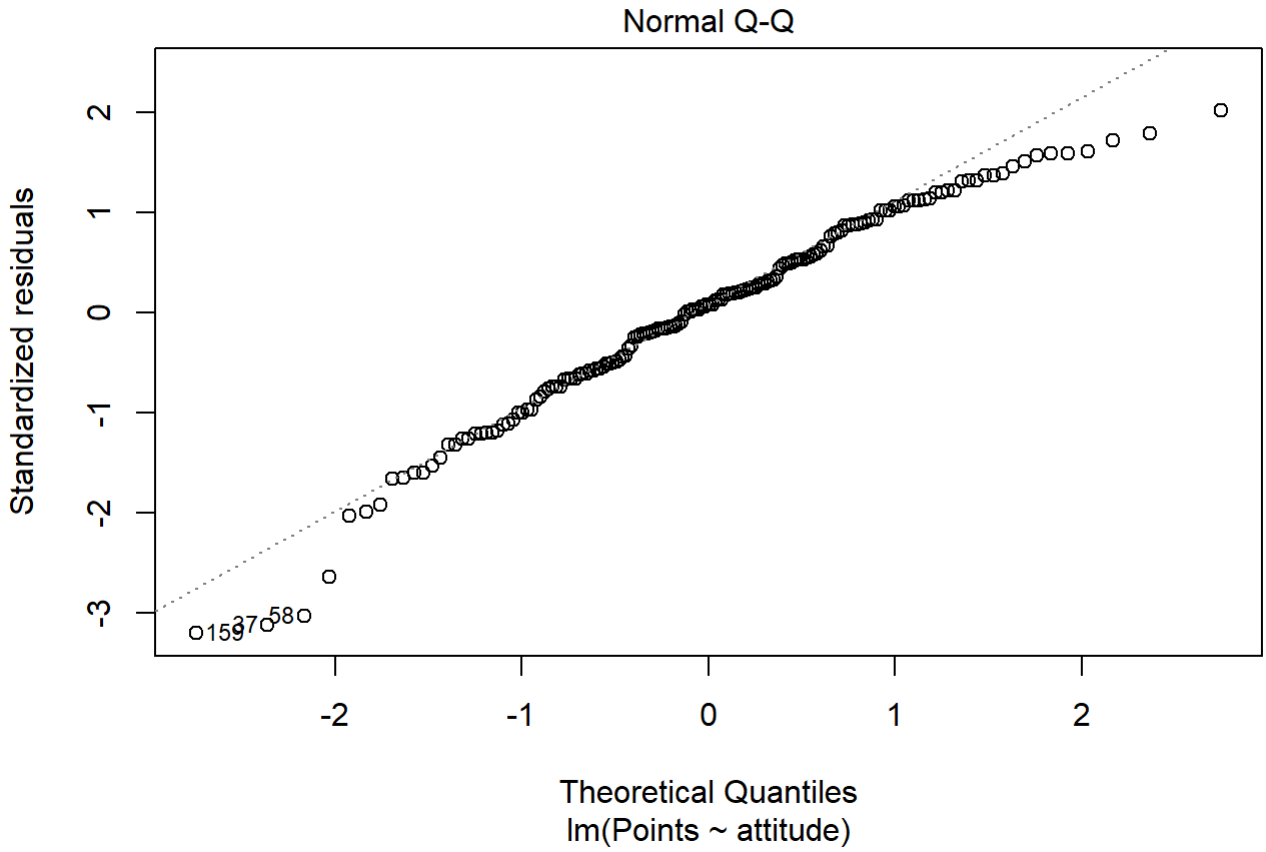
*Conclusion: attitude counts!*

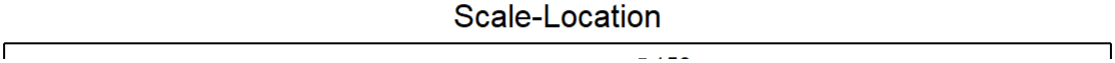
# TASK 5

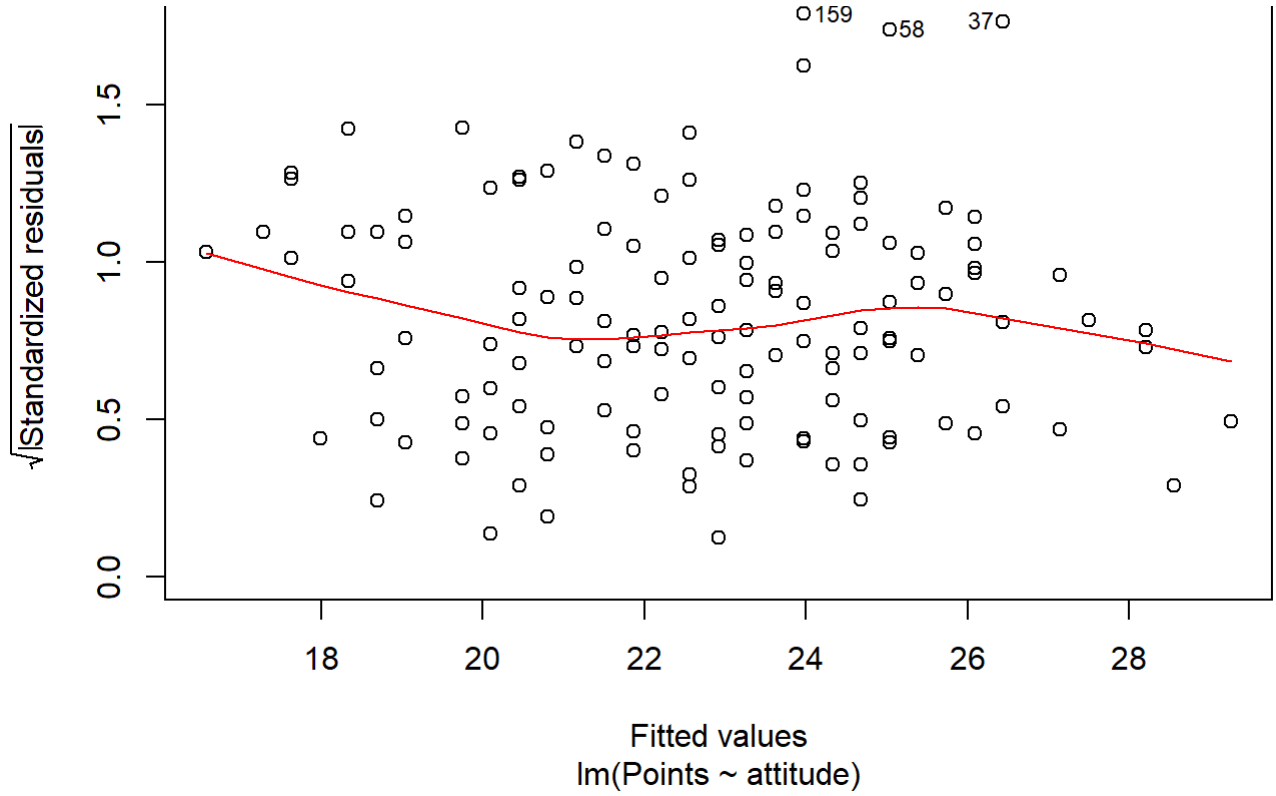
# Diagnostic plots

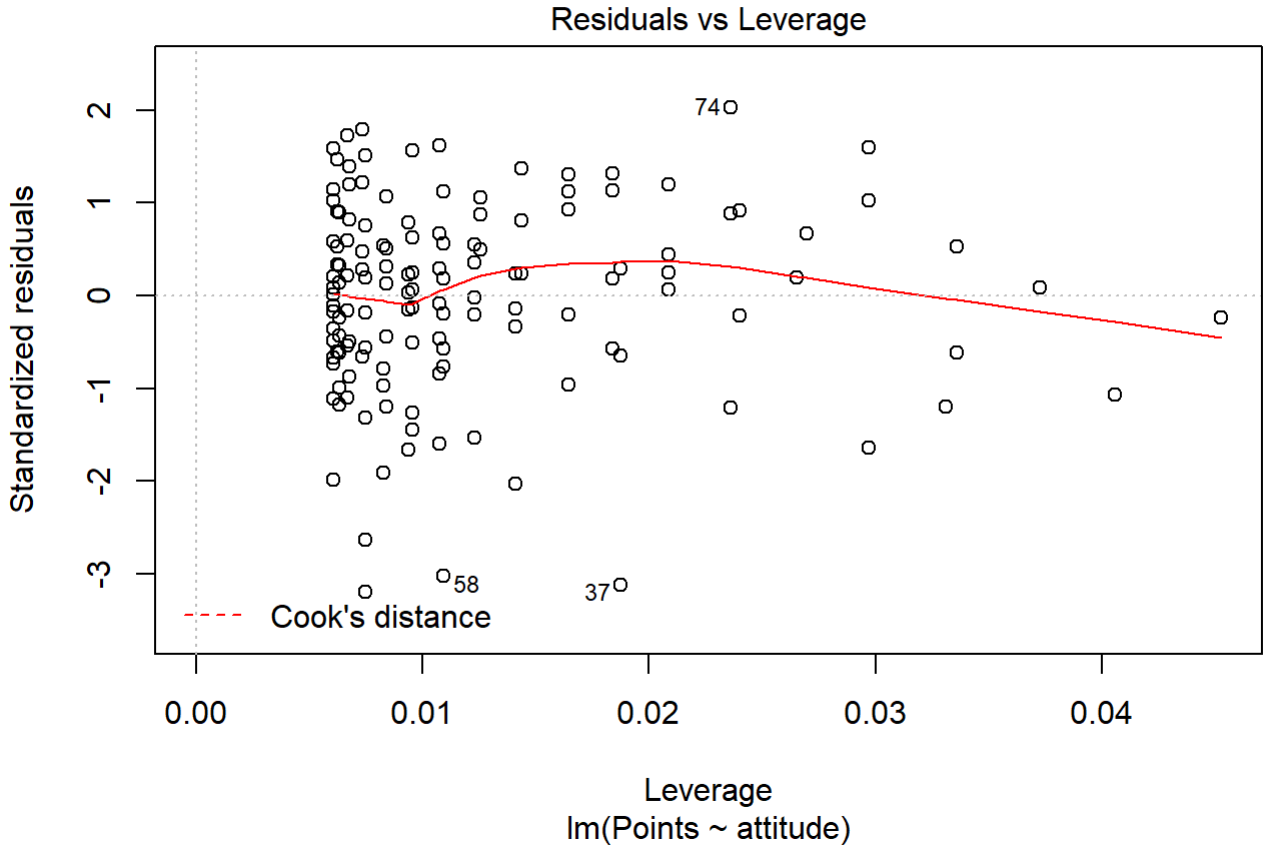
```
plot(regr_model3)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
-15
-20
18
20
22
24
26
28
159
58
37
Fitted values
lm(Points ~ attitude)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-2
-1
0
1
2
159
37
58
Theoretical Quantiles
lm(Points ~ attitude)


Scale-Location




Residuals vs. fitted values: *The residuals seem randomly distributed, so there is no dependency between residuals and the predicted values.*

Normal QQ-plot: *The standardized residuals deviate a bit from the plotline, but overall, the residuals are rather normally distributed. Our assumption that exam points are normally distributed seems to hold well.*

Residuals vs. Leverage: *All subjects are within Cook's distance, i.e. there seem to be no cases having an especially strong influence on the regression analysis.*

*These plots speak for the robustness of our regression model, i.e. the assumptions of using the model are not violated.*