IODS course project

Contents

Introduction to Open Data Science - Course Project

About the project

Write a short description about the course and add a link to your GitHub repository here. This is an R Markdown (.Rmd) file so you should use R Markdown syntax.

The link to my GitHub repository is https://github.com/statnaia/IODS-project The link to GitHub Pages is https://statnaia.github.io/IODS-project/

```
# This is a so-called "R chunk" where you can write R code.

date()
```

```
## [1] "Sat Nov 13 17:11:24 2021"
```

The text continues here.

Chapter 2: Regression and model validation

Dataset: JYTOPKYS3
- The dataset is an international survey of Approaches to Learning done by Kimmo Vehkalahti in 2014-2015
- The dataset learning2014 consist 166 rows and 7 variables.

```
#reading the dataset
learning2014 <- read.csv("D:/Desktop/Courses/Data Science/IODS-project/Data/learning2014.csv", sep=" ",

#checking the structure and dimensions of the dataset
str(learning2014)
```

```
## 'data.frame':    166 obs. of  7 variables:
##  $ gender  : chr  "F" "M" "F" "M" ...
##  $ Age     : int  53 55 49 53 49 38 50 37 37 42 ...
##  $ 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 ...
##  $ Points  : int  25 12 24 10 22 21 21 31 24 26 ...
```

```
dim(learning2014)
```

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

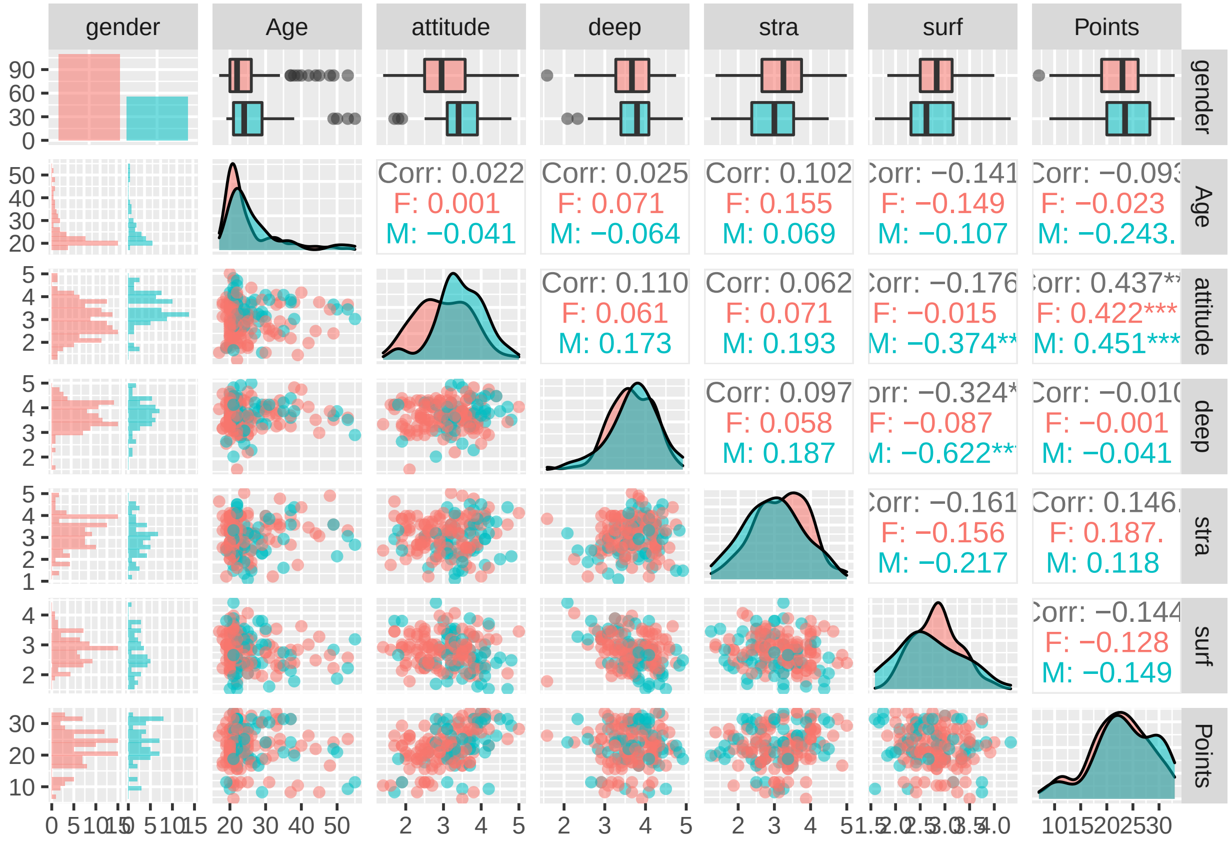
First we explore the data by constructing scatter plots, PDFs and correlations between the variables by gender. Pink color denotes the information on female participants of the survey, and cyan color denotes the information on males. Number of females is approximately twice larger than the number of males. Most of the respondents are under age 35-40. The boxplots and PDFs that denote Global attitude toward statistics, Deep approach, Surface approach, Strategic approach and Total points look quite similar for both genders. Overall males have somewhat higher scores for attitude that females and vice versa for Surface approach. Surface approach scores are negatively correlated with all other variables.

The correlations between variables are in general quite low, and non-significant in many cases. This also can be seen from the scatter plots, the relationships between variables seem mostly quite random. Surface approach scores are negatively correlated with all other variables, but are significant only for males when correlated with Deep approach and Global attitude toward statistics. On the other hand, variables Global attitude toward statistics and Total points are significantly positively correlated with each other for both males and females.

```
# access the GGally and ggplot2 libraries
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
# create a more advanced plot matrix with ggpairs()
p <- ggpairs(learning2014, mapping = aes(col = gender, alpha = 0.3), lower = list(combo = wrap("facethis

# draw the plot
p
```



Having studied the relationships between variables, it seems that the Global attitude toward statistics might explain the variation in Total points the best. Nevertheless, the use of the two other variables: Strategic approach and Surface approach might improve the model. A summary of a multiple linear regression model is shown below.

```r
# creating a multiple regression model with attitude, strategic learning, and surface learning as expla
# target variable is Points
my_model <- lm(Points ~ attitude + stra + surf, data = learning2014)

# print out a summary of the model
summary(my_model)
```

```
##
## Call:
## lm(formula = Points ~ attitude + stra + surf, data = learning2014)
##
## 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
```

As the statistical significance is marked by stars (t value and Pr(>|t|) columns), Global attitude toward statistics is in fact significantly positively correlated with exam points, but the other two variables are not. The p-values for these two variables are greater than the .05 value, which is generally accepted to test the significance. Therefore, the null hypothesis is not rejected.

Based on these results, I remove these two parameters and make a new model:

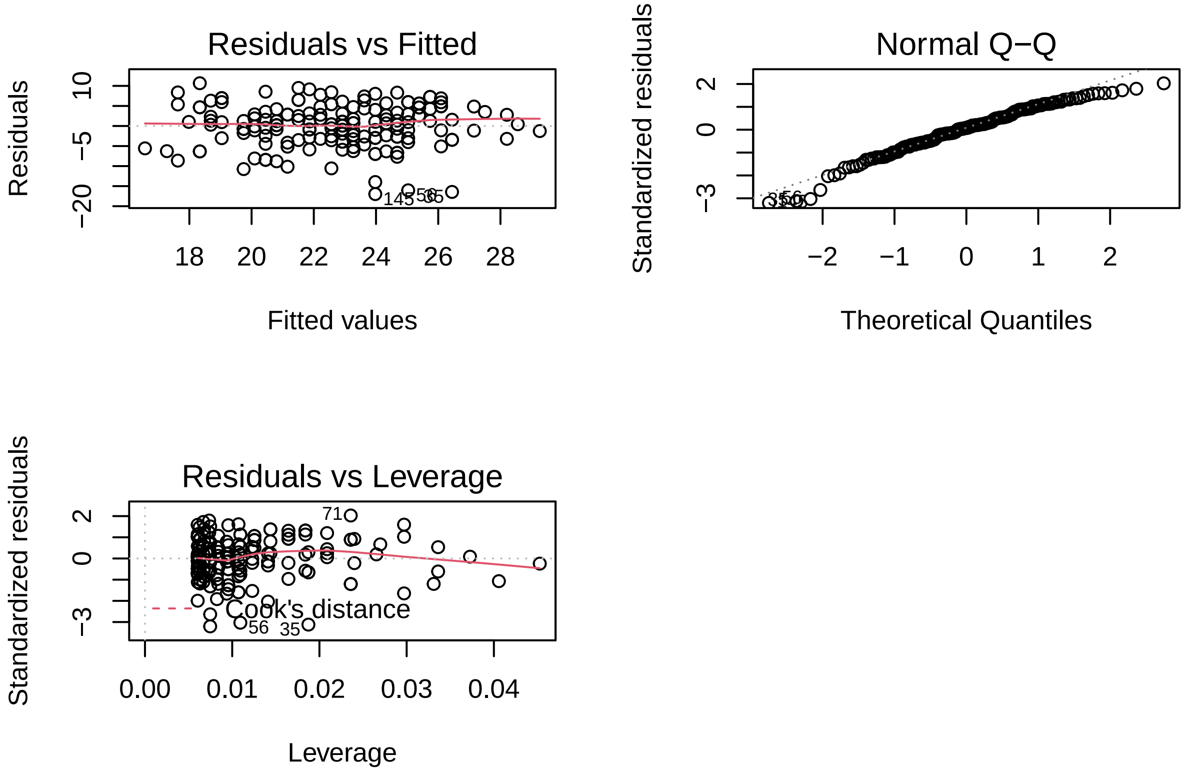
```
#remove stra and surf and run model again
my_model2 <- lm(Points ~ attitude, data = learning2014)

# print out a summary of the model
summary(my_model2)
```

```
##
## Call:
## lm(formula = Points ~ attitude, data = learning2014)
##
## 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
```

The simple linear model performs better than the multimpe regression model in our case. The residuals for this model are slightly smaller than for the previous model, indicating a better model fit. The model fit is described the value of the Multiple R squared: 0.19, indicating that the model can explain 19 percent of the variance in our dependent variable. In the case of this simple linear regression, this means that differences in attitude explain about a fifth of the variance in exam points.

From the “Residuals vs Fitted” plot we can see, that the relationship between the residuals and the fitted values is quite random, which indicates that the size of the errors is not dependent on the explanatory variable. In the “Normal Q-Q” plot we see that the errors are reasonably normally distributed, and thus fit the normality assumption, and the results in the “Residuals vs Leverage” plot imply, that no single observation has unusually high impact on the model. The model diagnostics show a reasonably good fit to the data.

```
# draw diagnostic plots
par(mfrow = c(2,2))
plot(my_model2, which = c(1,2,5))
```



Insert chapter 3 title here

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.

```
date()
```

```
## [1] "Sat Nov 13 17:11:30 2021"
```

Here we go again…

(more chapters to be added similarly as we proceed with the course!)