Assignment 2: Regression and Model Validation

- This week, we are learning about regression and model validation. In this exercise, we are using a data from the International Survey of Approaches to Learning. In this survey, the students were asked to assess themselves on the scale of 1-5 about various statements related to their learning (ex: "I organize my study time carefully to make the best use of it"). The questions could be classified into three different learning approaches: strategic learning, surface learning, and deep learning. Beside the questions about the student's learning approach, the data also includes questions about the student's attitude on statistics, age, exam points, and gender.
- We have wrangled the original data to obtain a new dataframe which contains the students' average assessment score for each learning approaches, average attitude towards statistics, and their personal data.
- In this report, we will use the new dataframe to analyze the relationship between the variables.

```
date()
```

```
## [1] "Fri Nov 25 18:12:07 2022"
```

Reading the data

```
data <- read.csv(file = "learning2014.csv", sep=",", header = TRUE)
```

Examining the structure and dimension of the data

```
dim(data)
```

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

```
str(data)
```

```
## '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 ...
```

The data has 166 observations and 7 variables, which are:

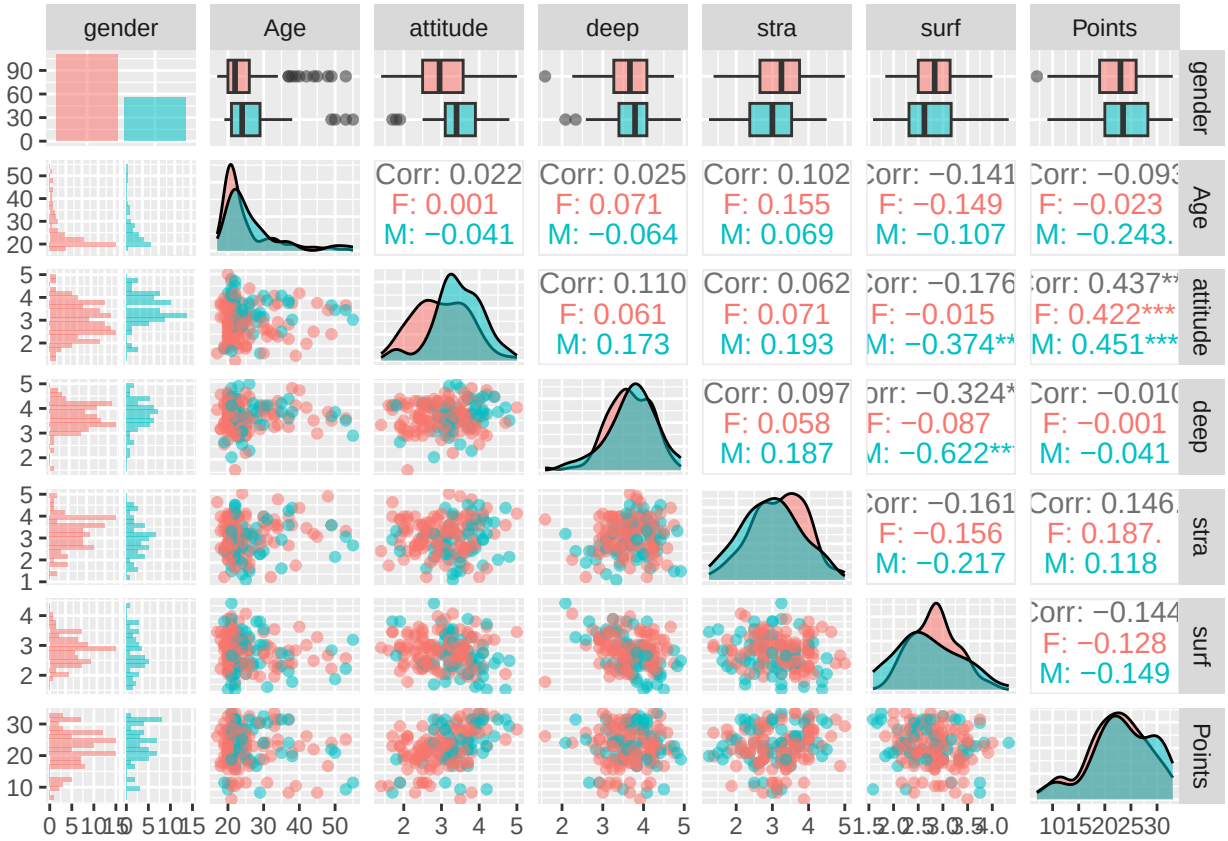
Variable name	Description
gender	Gender of the student (character, F/M)
Age	Age of the students in years (integer)
attitude	The average score of the student's attitude towards statistics (numeric)
deep	The student's average score on deep learning approach (numeric)

Variable name	Description
stra	The student's average score on strategic learning approach (numeric)
surf	The student's average score on surface learning approach
Points	The student's exam points (integer)

Overview of the data

```
#accessing the libraries
library(ggplot2)
library(GGally)
```

```
#drawing scatter plot matrix
ggpairs(data, mapping = aes(col=gender, alpha=0.3), lower = list(combo = wrap("facethist", bins = 20)))
```



The data for female students are represented in pink, while the data for male students are represented in green. The frequency graph shows us that we have a lot more data from female students. The students' age has a highly skewed distribution graph, which tells us that most of the students are on the younger side, but there are a few much older students as well. We can see that the student's attitude towards statistics are moderately positively correlated with their exam points, which is not surprising. An interesting correlation can be seen between surface and deep learning scores. The two variables have a negative correlation, which means that students who score higher on surface learning tend to score lower in deep learning, and vice versa. However, a moderately high negative correlation is only observed in male students, not in female students. A

similar phenomenon can also be observed between surface learning score and the students' attitude towards statistics.

```
summary(data)
```

```
##     gender               Age           attitude          deep
##  Length:166         Min.   :17.00   Min.   :1.400   Min.   :1.583
##  Class :character   1st Qu.:21.00   1st Qu.:2.600   1st Qu.:3.333
##  Mode  :character   Median :22.00   Median :3.200   Median :3.667
##                     Mean   :25.51   Mean   :3.143   Mean   :3.680
##                     3rd Qu.:27.00   3rd Qu.:3.700   3rd Qu.:4.083
##                     Max.   :55.00   Max.   :5.000   Max.   :4.917
##       stra            surf           Points
##  Min.   :1.250   Min.   :1.583   Min.   : 7.00
##  1st Qu.:2.625   1st Qu.:2.417   1st Qu.:19.00
##  Median :3.188   Median :2.833   Median :23.00
##  Mean   :3.121   Mean   :2.787   Mean   :22.72
##  3rd Qu.:3.625   3rd Qu.:3.167   3rd Qu.:27.75
##  Max.   :5.000   Max.   :4.333   Max.   :33.00
```

Here, we can see the descriptive statistics of each variable in the data. The students' age ranges from 17 to 55 years, while the average is 25.5 years old. The students' attitude towards statistics averages at 3.14, which means that the students have a slightly more positive attitude towards statistics in general. Among the three learning approaches, deep learning has the highest average score while surface learning has the lowest average score. Meanwhile, the students' exam point ranges from 7 to 33 and averages at 22.7.

Building the model

```
library(tidyverse)
```

```
#3 variables as explanatory variables
fit <- data %>%
  lm(Points ~ attitude + stra + deep , data = .)
summary(fit)
```

```
##
## Call:
## lm(formula = Points ~ attitude + stra + deep, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5239  -3.4276   0.5474   3.8220  11.5112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3915     3.4077   3.343  0.00103 **
## attitude      3.5254     0.5683   6.203 4.44e-09 ***
## stra          0.9621     0.5367   1.793  0.07489 .
## deep         -0.7492     0.7507  -0.998  0.31974
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.289 on 162 degrees of freedom
## Multiple R-squared:  0.2097, Adjusted R-squared:  0.195
## F-statistic: 14.33 on 3 and 162 DF,  p-value: 2.521e-08
```

The low p-value and positive estimate suggest that the student's attitude towards statistics is significantly and positively related to their exam points. We also found an evidence of a positive relationship between the students' strategic learning score and their exam points (p-value = 0.075, significant at the 0.1 level). However, we did not find any evidence of a relationship between the students' deep learning score and their exam points. Next, we will remove this variable and fit the model again without it.

```
#remove insignificant variable
fit2 <- data %>%
  lm(Points ~ attitude + stra, data = .)
summary(fit2)
```

```
##
## Call:
## lm(formula = Points ~ attitude + stra, 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
```

We obtained a similar result with the previous model. With a very low p-value, the students' attitude towards statistics was still found to be a highly significant predictor of their exam point. The students' exam score increases by 3.46 points on average with each increase in the students' attitude score, assuming that their strategic learning score remains constant. Meanwhile, the students' strategic learning score was found to be significant at 0.1 level (p-value = 0.08). The student's exam score increases by 0.91 points on average with each increase in the students' strategic learning score, assuming that their attitude score remains constant. In another word, students who have a more positive attitude towards statistics and/or practices strategic learning approach tend to have a higher exam point.

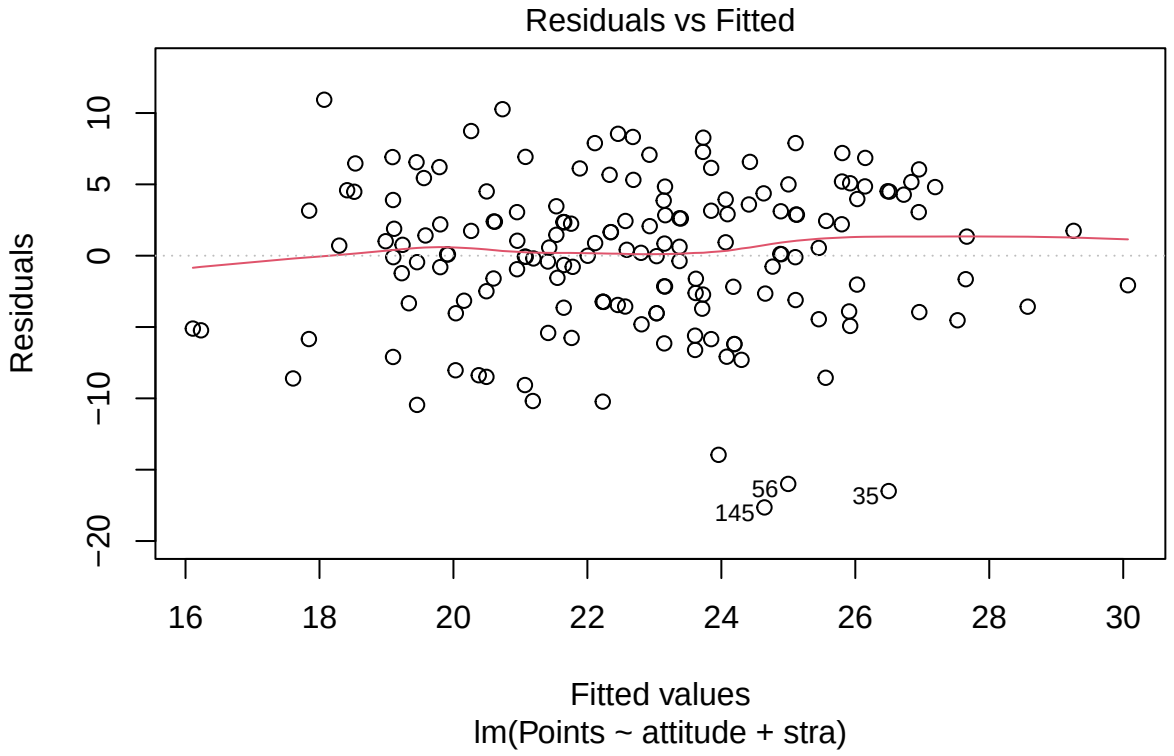
The adjusted R-squared of this model was 0.1951, which means that this model explains about 19.51 percent of the variations in the student's exam point. Considering that this is a very simple model, this is already a quite high R-squared value.

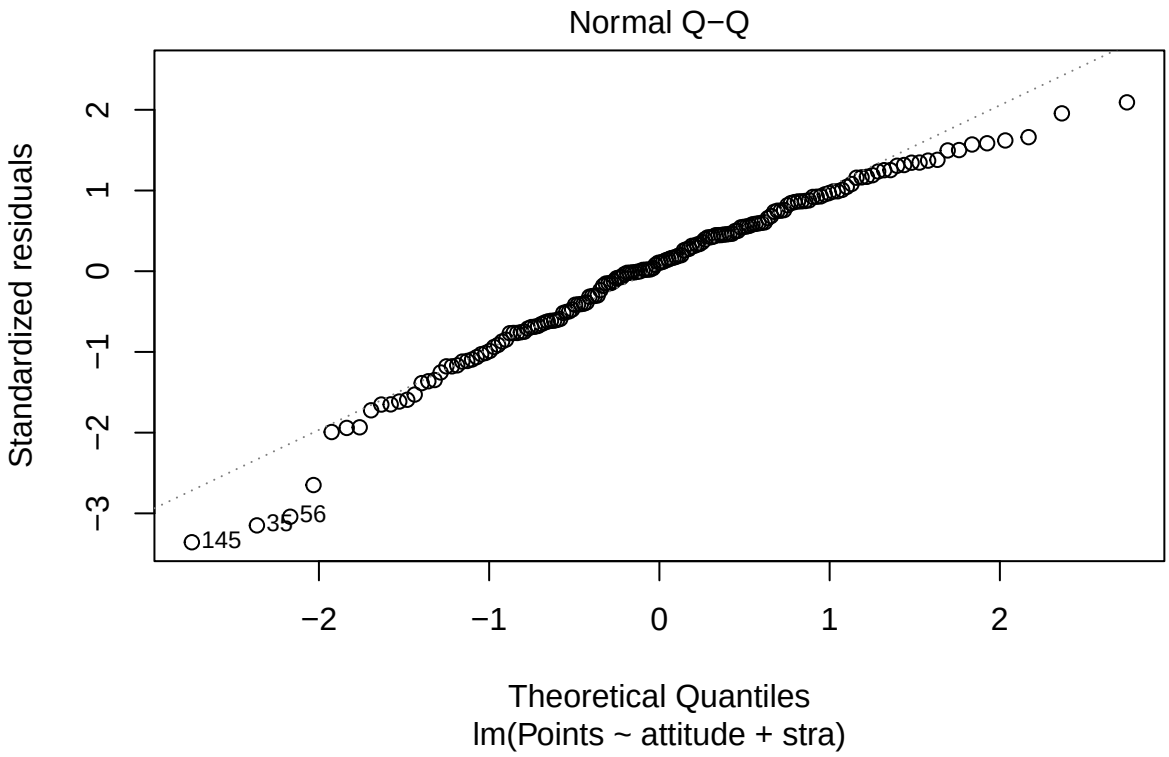
Diagnostics

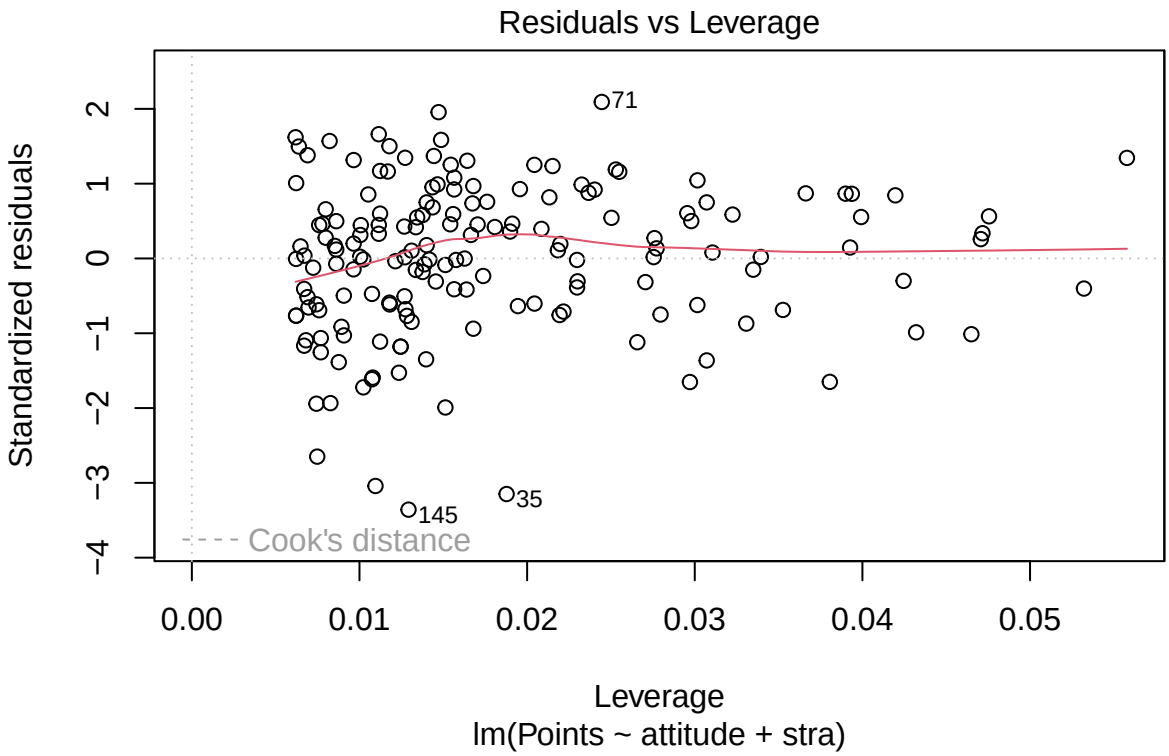
In linear regression, we assume a linear correlation between the variables, and the error term/residual is normally distributed. We also assume that the variance of the residuals are equal across all predicted

values (homoscedasticity). The residuals vs. fitted values plot and the Q-Q plot can be used to check these assumptions.

```
plot(fit2, which = c(1, 2, 5))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Points ~ attitude + stra)
145
35
56



In the residuals vs. fitted values plot, we can see that the residuals seem to be randomly scattered. It does not seem to display any concerning patterns, such as a curve (suggesting non-linearity) or a trombone pattern (suggesting heteroscedasticity). Based on this plot, it seems that the data is linear and homoscedastic (the variance of the residuals tend to be equal across all predicted values).

The Q-Q plot compares the standardized residuals to their theoretical quantiles (the values they should have if the normality assumption is fulfilled). If the assumption is fulfilled, the points should fall across the straight line. In this plot, we can see that the points seem to form a slight upward curve. This means that the distribution of the residuals are actually a bit left skewed.

The residuals vs leverage plot is used to check if there is any outliers that might affect the model heavily. There seem to be several extreme values in the data, but none of them fall outside of the Cook's distance line, so they are not necessarily considered to be influential to the model.