#Excercises for week 2

setwd(“~/IODS-project”)

I have done some linear models. I also lost my github button, but luckily I got it back! And obviously this likes to happen on times such as sunday night!Hmm, lets see what happens, now the github diary does not work

date()

## [1] "Sun Nov 14 20:06:11 2021"

Here we go again. This next session lists all the libraries needed for completing this excercise:

library("ggplot2")

## Warning: package 'ggplot2' was built under R version 3.6.3

library("GGally")

## Warning: package 'GGally' was built under R version 3.6.3

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

Lets start with reading the dataframe from the provided link. We shall also look at the data more closely by using the summary function. I am not printing here the structure of our data, because from the environment page it can be seen.

data<- read.table(url("http://s3.amazonaws.com/assets.datacamp.com/production/course\_2218/datasets/learning2014.txt"),   
 sep = ",", header = T)  
  
summary(data)

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

This is a classic dataframe with 166 rows and 7 columns: gender, age, attitude, deep, stra, surf and points. It seems that this dataframe is the same we created with data wrangling excercise. However, I think each row is an individual and columns are results from that individual.

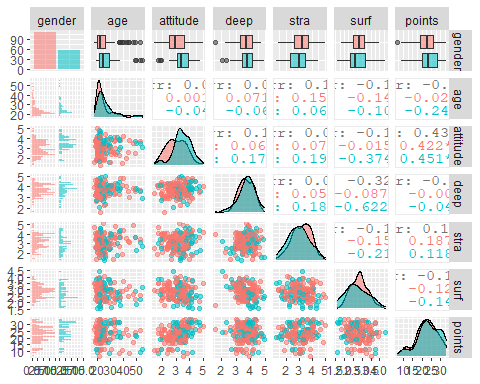
There are 110 females and 56 males in the data. Mean age is 25.51 years. Variables attitude means global attitude towards statistics, points are derived from the exam and variables deep, stra, and surf were constructed from subquestions as in previous excercise.

From now on, abbreviations will be used, but different variables mean these things:

stra = Strategic approach deep = deep approach surf = surface approach

ggpairs(data, mapping = aes(alpha = 0.3, col =gender))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



summary(data)

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

From the graphical oveview of the data, we can see that we have more women than menas a test subjects. The average age of women is smaller than men.

Men have generally a bit better attitude towards statistics, however, it is not stated as statistically significant. I don’t see other differences between gender and individual variables (boxplots are quite overlapping).

It seems that the age is not normally distributed variable, the peak is in on the right side of the graph. Thus, people included into this study seem to be generally young rather than old.

When it comes to attitude, it seems that attitudes of women is almost normally distributed while men have a bit left tilted graph. (still, this might be okay to say its normally distributed).

Deep seems to also be a bit leaning to left as a graph, it would be good to chech the distribution with some calculation method rather than looking the graphs.

Stra and surf learnings seems to be normally distributed variables, while the points is not (it has this odd tail on the right side).

It seems that there is startistically significant correlation between the attitude towards statistics and points obtained from the exam. This is true for both genders.

Also surface and deep learning seem to be strongly negatively correlated in the population including both genders and among men.

Attitude and surface learning also seem to be negatively correlated, in the whole population and among men.

Surface and strategic learning seem to be correlated in whole population, but such correlation is not seen in one-gender-only populations.

For our regression model we ought to choose (from the graphical output, three variables that have the best correlation with points).

Three having the greatest correlation values with points according to our graphical interpretation seem to be attitude (0.437, positively correlated and statistically significant), stra (0.146, not statistically significant), and surf (-0.144, so negatively correlated, not statistically significant).

linearmode<- lm(points ~ attitude + stra + surf, data = data)  
summary(linearmode)

##   
## 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

From the summary of linear model we can observe following: There is a significant correlation between attitude and points. Also, the crossing of the axis is not 0, which is also a significant observation. Other variables in this model are not significantand thus they are left out from the model.

Lets make new model with attitude, gender and deep.

linearmode<- lm(points ~ attitude + gender + deep, data = data)  
summary(linearmode)

##   
## Call:  
## lm(formula = points ~ attitude + gender + deep, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.0364 -3.2315 0.3561 3.9436 11.0859   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.6240 3.1799 4.284 3.13e-05 \*\*\*  
## attitude 3.6657 0.5984 6.125 6.61e-09 \*\*\*  
## genderM -0.4633 0.9170 -0.505 0.614   
## deep -0.6172 0.7546 -0.818 0.415   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.337 on 162 degrees of freedom  
## Multiple R-squared: 0.1953, Adjusted R-squared: 0.1804   
## F-statistic: 13.1 on 3 and 162 DF, p-value: 1.054e-07

Again, the only statsistically significant correlation is between attitude and points. For us students this is a happy finding, a better attitude we have, better results we will have:)

So my final linear model is following:

linearmode<- lm(points ~ attitude, data = data)  
summary(linearmode)

##   
## 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

The residuals in my model are reflecting how well my model predicted the actual value of y. Thus, these are the difference between actual points value and calculated points value. If residuals are symmetrically divided to either side of my plot, the value is 0. Because the median is over 0, there are more samples above the model thatn under the model. However, there are samples that my model predicts too high values especially in the lower scale (residual minimal is -16).

Coefficients have the estimated value, std error from the residuals, t-value (estimate divided by standard error) and Pr(>|t|) which is the lookup of the t-value in t-distribution table with given degrees of freedom.

In the bottom of the summary output we have the residual standard error, which is very similar for standard deviation. Thus, reflecting how much there is varition in errors.

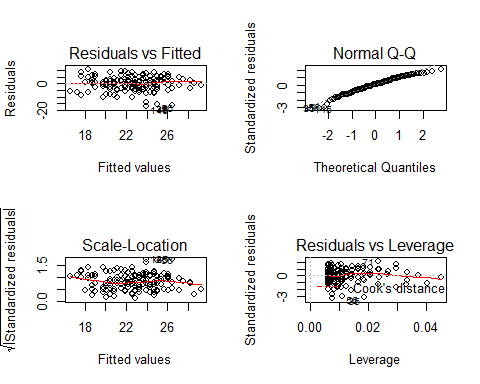
Multiple R-squared represents how well my model is reflecting my data. It is calculated by counting the explained variation of the model by the total variation of the model. If my model predicts things well, we have R-squared close to 1, while poorly performing model might even have negative values My model predicts now 19.06 percent of the whole variation. According to the datacamp, R-squared over 0.5 is good, so the model is not very accurate.

If we raise it to the second power, we will really see how well my model is reflecting my data. 3.632836 % can be seen to be caused by attitude variable.

Adjusted R-squared is multiple r-squared adjusted for the multiple hypothesis testing (if we would have several variables). Because the R-value tends to get bigger even though there would be unsignificant variables.

F-statistic: This parameter is quite like a t-test for the whole model, it gives me a p-value about how likely my model is to be just randomly fitting my data like this.

par(mfrow = c(2,2))  
plot(linearmode)



Here we are able to see, that our data contains some outliers (the residuals vs fitted-line is not straight and the outliers are indicated in the plot).

Homoscedasticity means that one variable has approximately similar variability in all values of other variable. We can see from Q\_Q plot that our model does not differ significantly from the predicted residuals presented in optimal theorethical model. I think these observations are almost perfectly in line with theorethically predicted residuals.

Heteroskedasticity means that the variability of dependent variable is altering significantly if the value of explaining variable is altered.

For observunf thse, we can look at the scale location plot where we want to really see two things: 1. line is horisontal 2. The value spread around red lines is not variating as much as fitted values (seems ok!).

And the last graph, residuals vs leverage is describing how close or far the points are from each other. We have some points with a bit higher leverage, but none of the is outside from Cook’s distance and thus deleting them might not have significant influence on our model.