IODS course project

Harri Lindroos

Table of Contents

# About the project

Quite excited about the course. I’ll just add something here, so we can get started. Maybe just this [**link**](https://github.com/harrlind/IODS-project).

# Reading the data

Let’s start by reading a prepared dataset “learning2014”:

learning2014 <- read.csv("learning2014.csv")

Now when the dataset is succesfully read, it is wise to observe it. Firstly, the dataset should have 166 observations and 7 variables concerning students (our observations), their attitudes, learning processes, and their respected success in Introductionary Statistics Course. Secondly, the exam points variable should be over zero (>0) in all cases. That means, that students with zero points are left out of our data and our analysis. This can be seen by printing a summary of the dataset:

summary(learning2014)

## gender Age Attitude deep stra   
## F:110 Min. :17.00 Min. :14.00 Min. :1.583 Min. :1.250   
## M: 56 1st Qu.:21.00 1st Qu.:26.00 1st Qu.:3.333 1st Qu.:2.625   
## Median :22.00 Median :32.00 Median :3.667 Median :3.188   
## Mean :25.51 Mean :31.43 Mean :3.680 Mean :3.121   
## 3rd Qu.:27.00 3rd Qu.:37.00 3rd Qu.:4.083 3rd Qu.:3.625   
## Max. :55.00 Max. :50.00 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

Observations and variables can be checked by printing the dimensions of our data:

dim(learning2014)

## [1] 166 7

Everything seems to be in order and we can start analysing the data. Let’s begin with a graphical overview of the whole dataset.

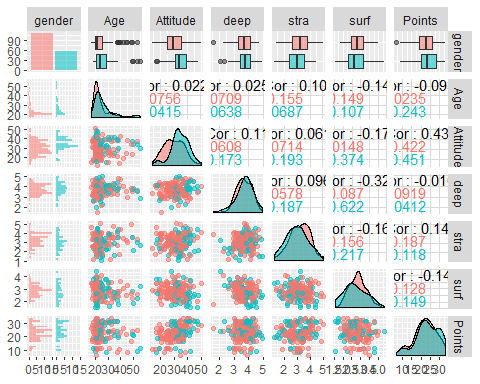
# Graphical overview and analysis

A good way to have a quick overview of our data is to print a scatterplot matrix. In base R this can be done easily with “pairs()” function. In our analysis we want to have a more precise look at the variables and their relations. In order to do this, we must first install (“install.packages()”) two packages “ggplot2” and “GGally”. After the installation we can introduce them in a following way:

library(ggplot2)  
library(GGally)

Now we can create and draw our scatterplot matrix:

learning2014matrix <- ggpairs(learning2014, mapping = aes(col = gender, alpha = 0.3), lower = list(combo = wrap("facethist", bins = 20)))  
  
learning2014matrix



Starting from the top row this matrix helps us to see all the variables in relation to gender (y-axis). Now it is easy to see, for example, that a majority of our dataset’s observations are females (red colour), and that the median age of male observations is higher than median age among females. The one thing that interests us the most is each variables relation to “points” variable and now this will be the centre of our analysis. In this analysis the last column is very helpful. From there we can see the correlations between each variable and points. The value of correlation (r) is between -1 (negative) and 1 (positive), and as the value approaches zero (0), the correlation weakens. The correlation values seen in our scatterplot tell us, that the explanatory variables with highest absolute correlations are “points” (highest positive correlation with r = 0.437), “strategic questions” (r = 0.146), and “surface questions” (highest negative correlation with r = -0.144).

In overall this tells us, that among our datas students there is a strong correlation with attitudes towards statistics and the respected exam success in one particular course of statistics. There is also a positive correlation with so called strategic learning approach and points. On the other hand there is a negative correlation with so called surface learning approach that depicts the level of superficiality contained in learning process. Among other variables the deep learning approach is closest to zero correlation (r = -0.010). We can see these relations among variables in more detail by fitting a regression model where exam points is the dependent target variable and all the other variables fuction as explanatory variables. Let’s take a closer look at the summary of this regression model:

pointsmodel <- lm(Points ~ Attitude + stra + Age + deep + surf, data = learning2014)  
  
summary(pointsmodel)

##   
## Call:  
## lm(formula = Points ~ Attitude + stra + Age + deep + surf, data = learning2014)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.4759 -3.2812 0.3804 3.5381 11.4897   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.30059 5.24076 3.492 0.00062 \*\*\*  
## Attitude 0.34320 0.05699 6.022 1.14e-08 \*\*\*  
## stra 0.96551 0.53917 1.791 0.07523 .   
## Age -0.09650 0.05335 -1.809 0.07234 .   
## deep -1.04374 0.78179 -1.335 0.18375   
## surf -1.10529 0.84009 -1.316 0.19016   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.249 on 160 degrees of freedom  
## Multiple R-squared: 0.2311, Adjusted R-squared: 0.2071   
## F-statistic: 9.62 on 5 and 160 DF, p-value: 4.803e-08

Like with the scatterplot, we can easily observe the effect of “attitude” to “points” but now in different way. Instead of the correlation value we should examine the values of t-test presented in this summary. From the values t-test we can estimate if they give reason to reject the null hypothesis that the actual value of given parameter is zero. In the case of attitude variable we have a high t-value (6.022) and, relating to that, a low p-value (1.14e-08). This gives us a reason to reject the null hypothesis and a reason conclude that there is a relation between these too variables. A quick way to see this is to pay attention to the stars (\*\*\*) at the end of attitude variables p-value: this tells us that the p-value is very low and the t-value is high.

We should also pay attention to the value of multiple R-squared or, in the case of multiple regression, to the value of adjusted R-squared presented in this summary. This decimal tells us how well our regression model explains the variation in points. Our numbers for R-squared and adjusted R-squared are 0.2311 and 0.2071. (When we have many explanatory variables the difference between these two becomes more important.) This tells us that our model accounts for 23%/21% of the variation.

Next we are going to continue our analysis by leaving only the variables with the stongest correlations, and by fitting a regression model with these as explanatory variables. Here is the summary of this model:

pointsmodel2 <- lm(Points ~ Attitude + stra + surf, data = learning2014)  
  
summary(pointsmodel2)

##   
## 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) 11.01711 3.68375 2.991 0.00322 \*\*   
## Attitude 0.33952 0.05741 5.913 1.93e-08 \*\*\*  
## stra 0.85313 0.54159 1.575 0.11716   
## surf -0.58607 0.80138 -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

Comparing these too models we can see how the estimated values for all of the parameters and also for the t- and p-tests have changed. But one thing stays the same as we still have high t-value, low p-value and the three stars for the attitude variable. All in all our second model has a multiple R-squared value of 0.2074 or success rate of ~21% in accounting for the variation (or 0.1927 and ~19% with adjusted numbers).

The discernible values and success of attitude variable in explaining our total variation gives us a reason to fit a regression model with only one explanatory variable:

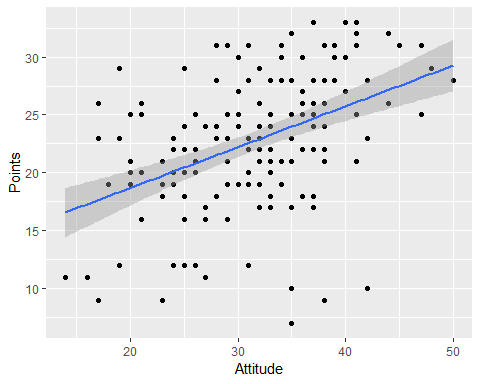
pointsmodel3 <- lm(Points ~ Attitude, data = learning2014)  
  
summary(pointsmodel3)

##   
## 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) 11.63715 1.83035 6.358 1.95e-09 \*\*\*  
## Attitude 0.35255 0.05674 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

Now we can see how the value of the response variable (that is “points”“) changes when the corresponding explanatory variable (that is”attitude“) changes by one unit. in our model the parameter for the attitude variable is 0.35255 and that means, that a rise of one point in attitude corresponds with the rise of 0.35255 exam points with the standard error of 0.05674. Our multiple R-squared is 0.1906 (or 0.1856), and this last model fares fares about as well as our previous model.

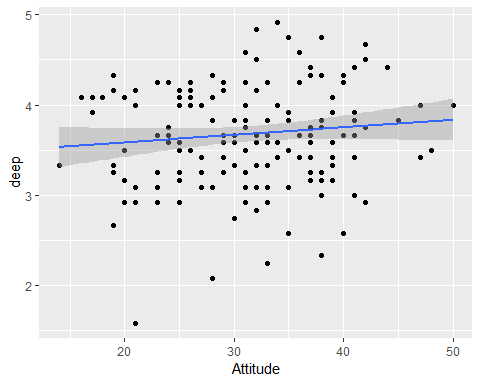
Here is the plot for attitude vs. points and also for deep questions vs. points. Now we can clearly compare our most succesful variable with the variable with absolute correlation value closest to zero:

qplot(Attitude, Points, data = learning2014) + geom\_smooth(method = "lm")



Compare:

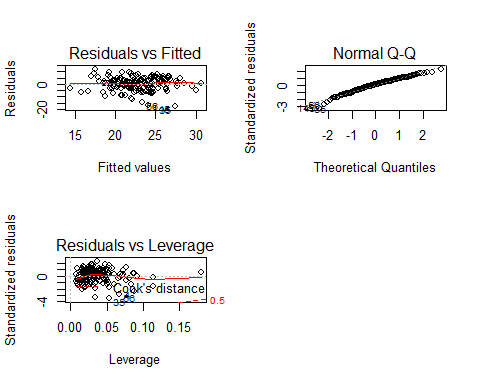
qplot(Attitude, deep, data = learning2014) + geom\_smooth(method = "lm")



# Model validation with diagnostic plots

Lastly we can explore the validity of ou prefered model of only one explanatory variable:

par(mfrow = c(2,2))  
plot(pointsmodel,which = c(1, 2, 5))



These plots show us that the residuals are normally distributed (the spread on the first graph, the fitness with the line in the second and reasonably low leverage values).