IODS course project

Alvin Khng

Table of 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.*

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

## [1] "Thu Nov 05 22:55:54 2020"

## About Me

My name is Alvin Khng and I am current doing my PhD studies in radiochemistry and radiobiology. I am excited for this course as I believed in the power of data statistics. I also feel that this would accelerate the process of data analysis and management i.e. more productivity. I have learnt basic python and used R in courses but most of the time is just follow a template within minimum changes. Therefore, I hope to learn more from this course. Lastly, I chance upon this course through the usual PhD emails regarding modules etc.

## My GitHub Repository

<https://github.com/Alvinash/IODS-project>

\*\*\*  
   
  
# Regression and model validation  
  
\*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.

## [1] “Thu Nov 05 22:55:54 2020”

## ‘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 …

## Registered S3 method overwritten by ‘GGally’:

## method from

## +.gg ggplot2

![](index\_files/figure-docx/unnamed-chunk-7-1.png)<!-- -->

## [1] “a pair relationship between two variables of all possible combinations. The nature of their plots are dependent on the two variables e.g. scatteplot, box and whiskers, barplot etc.. Their correlation is shown as well.”

## 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

## [1] “A summary of data showing the type of variables, mean, minimum, maximum and 25%, 50% and 75% quartile values.”

Based on the summary of the data, three explanatory variables "attitude", "surf" and "stra" are picked to fit a regression model against exam points as they have the highest absolute correlation value.

## 

## 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

The summary shows the statistical test of the fitted model. It can be seen that only "attitude" has a statistical significant relationship with the "point"s variable, concluding that there is a statistical relationship between "attitude" and "exam points". Therefore, the model is fitted again with just the "attitude" variable.

## 

## 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

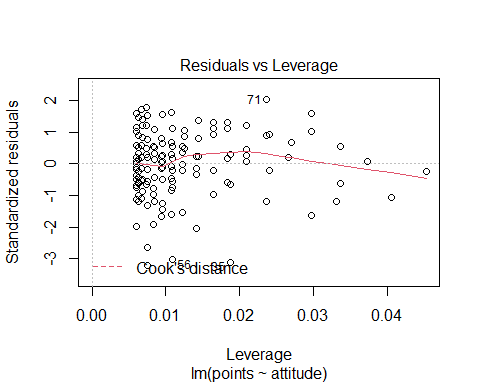
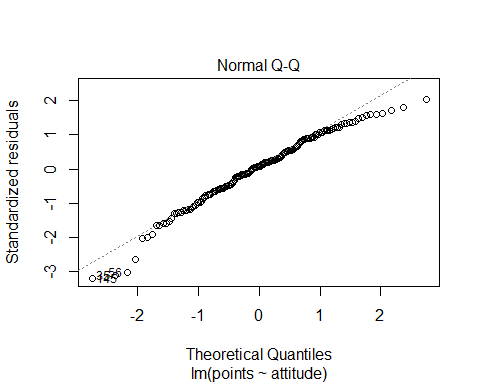
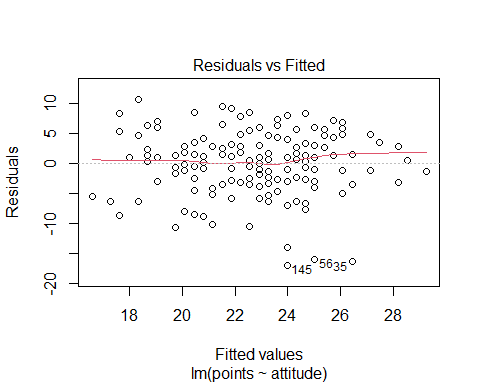
## 

## 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 new linear fit summary shows that the P value has improved by a factor of 10. the there is a positive relationship between “attitude” and “points” where every 1 unit increase in “attitude” results in a 3.5255 unit increase in “exam points”. However, the linear fit has a low multiple R-squared value of 0.1906, which indicates that only some of the variability of the data is around its mean.



The linear regression model assumes that: - The data follows a linear relationship - The errors are normally distributed - The errors are not correlated - The errors have constant variance

The normal QQ plot confirms that the data between “attitude” and “exam points” are normally distributed since the most of the data fits in on the normal line. The residual vs fitted plot shows that the data plots are similarly spread as fitted values increases, implying that the errors have constant variance.

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