MACQUARIE University SYDNEY·AUSTRALIA

STAT2170 and STAT6180
Applied Statistics

Assignment Semester 1. 2021

To achieve full mark you are required to complete this assignment using R Markdown to compile **a reproducible** PDF file for your submission and use the Git version control. On iLearn you only need to submit your pdf file, no need to submit your .Rmd file. For the Git repository you need to submit both files.

You need to submit your assignment via the provided submission link on iLearn by the due date. To further score marks for Question 4, you have to push the assignment file to provided Github repository.

You may discuss the assignment in the early stages with your fellow students. However, the assignment submitted should be your own work.

The R Markdown 'Cheatsheet' from the RStudio team is given here.

In your answers to the questions below, produce the appropriate `R` *output and explanation of the steps and results. Don't include any more* `R` *output than necessary and include only concise explanations.*

Rubric

The Assignment is worth 20% of the unit marks. This is an assessment task that will test both, your statistical knowledge and technical skills used in this unit.

Question 1 [30 marks] - Tests your applied statistics skills

Question 2 [17 marks] - Tests your applied statistics skills

Question 3 [10 marks] - Tests your RMarkdown technical skills

Marking Guide/Rubric for this question:

- Only 5 marks if the assignment file is compiled, eg. from RMD to HTML/Word
- Full 10 marks if the assignment file is compiled from RMD to PDF (Latex)

Question 4 [7 marks] - Tests your Git version control technical skills

Marking Guide/Rubric for this question:

- Only 3 marks if only uploaded to the designated repository once
- Full 7 marks if used proper Github submission workflow: if submitted at least twice into the designated repository with proper description, commit and push

A small tutorial on R Markdown

The following are some notes to kickstart into your `R Markdown` journey (we discusses some of these in more details in Week 6 Part B Lecture and SGTA Week 7).

1. If you see an error message of `pdflatex not found`, then you are at the right place. To knit to a pdf you need to install LaTeX on your computer. This is rather big (e.g. MacTeX is approximate 3.9Gb and MiKTex 192Mb), but a recommended option. Before installing anything, make sure you have admin right to your computer before you start. If you have encountered issues with installation of LaTeX, then you could try to install via `tinytex` which is much more light-weight. Open `R` and enter the following commands:

```
install.packages("tinytex")
tinytex::install_tinytex()
```

- See Week 6 Part B Lecture for some other alternatives.
- For Mac users, you may be asked to install `Xcode` (another rather big installation). We only need a small piece of it called the `command-line tools`. Run the following line: `xcode-select --install` in the `Terminal` app on your Mac to continue.

2. To communicate your assignment results to us (this is the knitting part), you need to know some markdown & LaTeX syntax. Learning Markdown syntax will help with your formatting while learning LaTeX syntax will allow you to typeset Mathematics (copying β into your .Rmd file and assuming it would work is one of the most common error) in your assignment. Here are some resources to get you started.

   - Markdown tutorial - 10 minutes tutorial link
   - Mathematics in R Markdown link
   - Remember Google is your best friend, and you should google whatever error messages you got. Able to debug your own code with Google (learn how to select the right keywords to improve your searches) and by trails and errors is part of the learning process. Please give this a go before reaching out for help.

   - Now create a new R Markdown document from RStudio and knit it.

3. If you are experiencing persisting or last minute (LaTeX) compiling issues, RStudio Cloud is an excellent platform. Simply uploading everything online and knit.
4. For those decided to use the RStudio Cloud platform, you will have to **download** the pdf file instead of printing to a pdf at the end. Printing to a pdf, unfortunately, will turn each page into an image and then the submission system will reject it.

5. It is also our recommendation to knit often so that you know which line(s) of code is(are) giving you the problem. (There is a keyboard short-cut for knitting.) This is not so dissimilar to when you work with the console that you only run a line at a time to identify the issue.
6. Another common mistake is that students use the code `read.csv("dat.csv")` and then assume `R` would be able to know (magically) that you are referring to `dat.csv` in a folder far far away (in `Download` folder probably) from your `.Rmd` file. At this point of the semester, you should all have your `.rproj` file and workspace setup already so, that everything will be run from there. Please go back to Week 1 lecture for more details.
7. If you are stuck, create a post on the iLearn forum! Also, check earlier posts before creating a new one. Most of the time, your issues have been discussed and resolved already.

Instructions for Git version control

To score marks on Question 4 you need to pull the assignment file from the repository, make changes to the template RMD file, compile it to PDF file, stage the changes, add proper description (Summary and Description) and pull the file to the repository. Do it at least twice to demonstrate level of skill in version control work flow. Refer to the following link to find out how Git version control works in RStudio.

- Happy Git and GitHub for the useR link
- RStudio Support blog article link

Once, RMarkdown and Git are already installed (and RStudio is configured for both) on your laptop, open the following the Github repository link provided on iLearn.

1. Accept the invitation and wait until you received a confirmation email.

2. In RStudio open New project, and choose Version Control, then choose Git.

3. Copy the repository URL, eg.https://github.com/MQ-STAT2170-STAT6180/2021-s1-stat2170-stat6180-assignment-yournamehere

    - you may add an exact folder location on your laptop,
    - when you create the project, the files will be downloaded automatically, i.e. the `pull` request will clone the repository on your laptop.

4. Open file `Assignment-your_name_IDstudent_here.Rmd` - this is your starter file for your answers in Rmarkdown.

5. In your top right-hand side window you will find Git section to: `stage` updated/changed files - please remember to add proper `description`.

6. When you click on a `Push` button, the `staged` files (RMD and PDF) will be uploaded to your repository.

Question 1 [30 marks]

A medical research team wants to investigate the survival time of patients that have a particular type of liver operation as part of their treatment. For each patient in the study, the following variables were recorded:

blood	Blood clotting Index
prognosis	Prognosis Index
enzyme	Enzyme function Index
liver	Liver function Index
age	Age of the patient, in years
gender	Gender of the patient, (Male of Female)
survival	Survival time of the patient after surgery (in days)

The data is available in the file `surg.dat` on iLearn.

a. Produce a scatterplot of the data and comment on the features of the data and possible relationships between the response and predictors and relationships between the predictors themselves.

   - You will need to remove the `gender` variable to do this.
   - Comment on why it is necessary to remove the `gender` variable to compute the correlation matrix.

b. Compute the correlation matrix of the dataset and comment.

c. Fit a model using all the predictors to explain the `survival` response. Conduct an F-test for the overall regression i.e. is there **any** relationship between the response and the predictors. In your answer:

   - Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.
   - Write down the Hypotheses for the Overall ANOVA test of multiple regression.
   - Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).
   - Compute the F statistic for this test.
   - State the Null distribution.
   - Compute the P-Value
   - State your conclusion (both statistical conclusion and contextual conclusion).

d. Using model selection procedures discussed in the course, find the best multiple regression model that explains the data.

e. Validate your final model and comment why it is **not** appropriate to use the multiple regression model to explain the `survival` time.

f. Re-fit the model using `log(survival)` as the new response variable. In your answer,

   - Use the model selection procedure discussed in the course starting with `log(survival)` as the response and start with **all** the predictors.

g. Validate your final model with the `log(survival)` response. In particular, in your answer,

   - Explain why the regression model with `log(survival)` response variable is superior to the model with the `survival` response variable

Question 2 [17 marks]

A car manufacturer wants to study the fuel efficiency of a new car engine. It wishes to account for any differences between the driver and production variation. The manufacturer randomly selects 5 cars from the production line and recruits 4 different test drivers.

kmL	The observed efficiency of the car in km/L over a standard course
car	The specific car (labelled 1, 2, 3, 4 or 5)
driver	The driver of the car (labelled A, B, C, D)

The data is available in the file `kml.dat` on iLearn.

a. For this study, is the design balanced or unbalanced? Explain why.
b. Construct two different preliminary graphs that investigate different features of the data and comment.
c. Analyse the data, stating null and alternative hypothesis for each test, and check assumptions.
d. State your conclusions about the effect of `driver` and `car` on the efficiency `kmL`. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests in c. and the preliminary plots in b.. You do not need to statistically examine the multiple comparisons between contrasts and interactions.