MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

Assignment Semester 1 2023

*To achieve full marks 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 25% 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 [45 marks] - Tests your applied statistics skills**

**Question 2 [25 marks] - Tests your applied statistics skills**

**Question 3 [20 marks] - Tests your RMarkdown technical skills**

Marking Guide/Rubric for this question:

- Only 10 marks if the assignment file is compiled from RMD to HTML/Word (even if you then convert it into a PDF file).
- Full 20 marks if the assignment file is compiled directly from RMD to PDF with LaTeX.

**Question 4 [10 marks] - Tests your Git version control technical skills**

Marking Guide/Rubric for this question:

- Only 5 marks if only uploaded to the designated repository once or you didn't include your StudentID as part of the assignment filename.
- Full 10 marks if used proper Github submission workflow: if submitted at least twice into the designated repository with proper description, commit and push. Also, your StudentID is included in the assignment filename.

**You should prioritise Questions 1 and 2 and present them in a format that you are comfortable with. Using Markdown, LaTeX, and Github can require substantial investment in your time and effort.**

## A small tutorial on R Markdown

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

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 4.7Gb 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 the 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` to continue. You should be able to find the `Terminal` tab next to the `Console` tab in `RStudio`.

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 $\beta$ into your `.Rmd` file and assuming it would work is one of the most common errors) 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 trial and error 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 upload everything online and knit.

4. For those who decide 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 the `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 a PDF file, stage the changes, add proper description (Summary and Description) and push the file to the repository. Do it at least twice to demonstrate your 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
- RStudio.cloud users - creating Personal Access Tokens link

  Once RMarkdown and Git are installed (and RStudio is configured for both) on your laptop, open the following 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-STAT6180/2023-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. `Assignment-your_name_IDstudent_here.Rmd` is your starter file for your answers in Rmarkdown. We strongly recommend you to rename it with your details.
5. Open the (renamed) starter file.
6. In the default RStudio layout, you should be able to find the `Git` section in your top right-hand side window (alongside `Environemnt` and `History`) to: `stage` updated/changed files - please remember to add proper `description`.
7. When you click on a `Push` button, the `staged` files (RMD and PDF) will be uploaded to your repository. If this is your first `Push`, you should log onto `Github` to check your files have been uploaded properly.

## Question 1 [45 marks]

World Health Organisation's (WHO) specialised cancer agency, the International Agency for Research on Cancer (IARC) has designated fine particulate matter ($PM_{2.5}$) as carcinogenic to human beings. $PM_{2.5}$ particles have a diameter of 2.5 micrometers (0.0025 $mm$) or smaller and they are small enough for people to breath them deeply into lungs and sometimes $PM_{2.5}$ particles can even enter the bloodstream. Research indicates that temperature in degrees ($^{\circ}C$), relative humidity in percentage (%), wind speed in kilometers per hour ($km/h$), and precipitation in millimeters ($mm$) are potential predictors for $PM_{2.5}$ concentration in milligram per cubic meter ($\mu g/m^3$).

A random sample of the annual mean temperature, humidity, wind, precipitation and $PM_{2.5}$ concentration at 56 test locations was collected. The data is available in the file `pm25.csv` on iLearn. It is located under Assessment → Assignment → Assignment datasets.

| Variable | Description |
|---|---|
| temperature | The annual mean temperature in degrees |
| humidity | The annual mean relative humidity in percentage |
| wind | The annual mean wind speed in kilometers per hour |
| precipitation | The annual mean precipitation in millimeters |
| pm25 | The annual mean $PM_{2.5}$ concentration in milligram per cubic meter |

a. [**7 marks**] Produce a plot and a correlation matrix of the data. Comment on possible relationships between the response and predictors and relationships between the predictors themselves.

b. [**6 marks**]

- Fit a model using all the predictors to explain the `pm25` response.
- Using the full model, estimate the impact of humidity on $PM_{2.5}$ concentration. Do this by producing a 95% confidence interval that quantifies the change in $PM_{2.5}$ concentration for each extra percentage of relative humidity and comment.

c. [**14 marks**] 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 for the test statistic.
- Compute the P-Value
- State your conclusion (both statistical conclusion and contextual conclusion).

d. [**10 marks**] Validate the **full** model and comment on whether the full regression model is appropriate to explain the $PM_{2.5}$ concentration at various test locations.

e. [**2 marks**] Find the $R^2$ and comment on what it means in the context of this dataset.

f. [**3 marks**] Using model selection procedures discussed in the course, find the best multiple regression model that explains the data. State the final fitted regression model.

g. [**3 marks**] Comment on the $R^2$ and adjusted $R^2$ in the full and final model you chose in part f. In particular explain why those goodness of fitness measures change but not in the same way.

## Question 2 [25 marks]

A business wants to advertise their product in Film media by using product placement in a movie. To maximise the brand recognition from the placement, the business conducted a study recording the correct number of brands identified by individuals in an experiment that watched different types of movies. Each movie in this experiment featured six different brands.

| Variable | Description |
|---|---|
| Gender | Gender of the individual watching the movie |
| Genre | Genre of the movie being watched |
| Score | The number of correct brands recalled by the individuals after the movie |

The data is available in the file `movie.csv` on iLearn. It is located under Assessment → Assignment → Assignment datasets.

a. [**2 marks**] For this study, is the design balanced or unbalanced? Explain why.
b. [**8 marks**] Construct two different preliminary graphs that investigate different features of the data and comment.
c. [**4 marks**] Write down the **full** mathematical model for this situation, defining all appropriate parameters.
d. [**9 marks**] Analyse the data to study the effect of `Gender` and `Genre` on the brand recall Score. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests you conducted in this part and the preliminary plots in part b. You do not need to statistically examine the multiple comparisons between contrasts and interactions. Remember to
    - state the null and alternative hypothesis for each test, and
    - check assumptions.
e. [**2 marks**] Based on your results from part d), discuss the practical implications of your findings for the business that aims to maximise the brand recognition from the placement. What advice/interpretation would you provide on the effect drama genre on the brand recall Score.