Coursework for MATH96067: Introduction to Statistical Learning: 2023

This coursework counts for 10% of the overall mark for the module.
This coursework should take you between six and eight hours to complete.
Date Coursework Set: 16th February 2023.
HAND-IN DEADLINE: 2nd March 2023, by 1pm.
HAND-IN METHOD: Electronically, via the Turnitin drop box on the course Blackboard page.

Your coursework must be anonymous. Please ensure that no identifying items are present anywhere on your coursework. E.g., please DO NOT write your name or CID or anything else that identifies you.

Task.
You should find a suitable data set for analysis by regression methods (e.g. multiple linear regression, ridge, lasso) that were **taught in Block 1.** You should set out the reasons for the choice of your data set and what is of interest in your data set that is intended to be revealed by the regression method. You should analyse your data set using regression methods **in R** and then write a report about what you did and what your conclusions were.

You should imagine you are in a working environment and writing your report for your technically knowledgable manager, who has to communicate/sell your conclusions to more senior management and stakeholders.

> **YEAR 4 students only:** Those students taking the unit as a year 4 unit should undertake the following extra component. Using your set of data, formulate and carry out a separate statistical analysis on your data. You have an extra 0.5 pages to write up this analysis and your conclusions, but the overall additional figure/table count remains at five (see below for limits).

Advice on how to undertake the task and associated conditions/rules follow:

1. Your data set can originate from any source. For example, this could be the Internet, from an academic paper or book, or data that you have collected yourself. Your report should make it clear where the data originates from and provide appropriate referencing.

2. Reasons for the choice of data set include: the data is of current topical interest, the data is related to an important subject, the outcomes of the analysis are particularly interesting or revealing, the data relate to your hobbies or interests but there has to be reason for studying the data using regression methods. This list of reasons is not exhaustive and you might have others.

3. You are NOT permitted to use any dataset that is built into R or one of its packages (but, obviously, you are permitted to use R functions and packages to analyse the data).

4. Your report should be submitted as a PDF file. The written content **can NOT exceed two A4 pages using no less than a 10-point font**. Your written report can be supplemented by

up to five additional figures or numerical tables with their associated captions (you do not have to include five, but this is the maximum). You can use any text processing system to produce your report (e.g. Microsoft WORD or LaTeX), but the submitted document has to be a *single* PDF file. You are permitted to produce dynamic or Shiny-type graphics online, but this is optional. If you do, each online figure/table counts to one of your allotted figures/tables. Provide the link to the dynamic graphic in your written report that can be clicked on to access the graphic.

5. No figures and tables are permitted to be presented or inserted into the two pages of text (2.5 for Year 4 students). The (max 5) figures and tables should be presented on pages following the text and should be referenced from the text. [This is because if figures or tables are inserted into the text, it can then be difficult to assess the length of the textual part of the report. We need to accurately gauge the length the textual part of your report so as to be fair to all students.]

6. There should be no separate title page.

7. Your report **should not** reproduce detailed mathematical development of regression methods as presented in lecture notes. Remember: you only have two pages to write about the data, your reasons, the analysis and reasons for using that analysis, the results, and conclusions.

8. Amongst other things, marks will be available for: (i) an interesting data set and good reasons for analysis using regression methods; (ii) a competently executed analysis of the data; (iii) a clear and cogent well-written report that includes a description of the results and conclusions; (iv) clear and informative figures/tables where necessary.