THE UNIVERSITY OF MELBOURNE

Student number	

Semester 2 Assignment, 2023

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 11:59 pm on Sunday 22 October 2023

This assignment consists of 4 pages (including this page) with 2 questions and 30 total marks

Instructions to Students

Writing

- Please submit a scanned or other electronic copy of your work via the Learning Management System. Your answer to all questions should be a single PDF file.
- You can type or handwrite your answer. If you handwrite your answer, write on A4 paper. Write on one side of each sheet only. Follow the instructions below for scanning and submitting of your assignment.
- If you use R, please include the R output in your answer and submit your R code as an additional file.
- Avoid only showing the R code and/or output as your answer without any explanation. You may lose marks if your answer is unclear.
- Page 1 should only have your student number, the subject code and the subject name. Each question should be on a new page. The question number must be written at the top of each page.

Scanning and Submitting

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Your scanned assignment must be a single PDF file. Check that all pages are present and readable before submitting.

Blank page

Question 1 (15 marks)

The data *winequality-red.csv* includes data from the paper *Modeling wine preferences by data mining from physicochemical properties* by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis (2009). The data consists of 1599 observations of red variants of the Portuguese "Vinho Verd" wine. The variables are:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol : percent alcohol content of the wine
- quality : Output variable (score between 0 and 10)

The quality score in this data set ranges from 3 to 8, but we will recode the levels as *bad*, *average* and *good*:

```
wine$quality = factor(wine$quality)
levels(wine$quality ) <- c("bad","bad","average","average" , "good","good")
```

(a) Fit a multinomial logit model to predict the wine quality by category, considering all available predictors. Refine the model using stepwise selection.

(b) Repeat the analysis with an ordinal model, with "good" being the highest level of the output. Comment on any differences.

(c) We have a new observation with these attributes:

```
newobs
# fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
#           7.9              0.4         0.2            1.7       0.1
# free.sulfur.dioxide total.sulfur.dioxide  density   pH  sulphates alcohol
#                  10                   36    0.997  3.3        0.9      10
```

Compute the probabilities that this wine variant is a "bad", "average" and "good" wine, according to the refined multinomial and ordinal models. You should NOT use the function `predict()` for this question.

(d) Under the ordinal model, what is the odds ratio of being classified as "bad" or "average" of a wine variant with chlorides level 0.08 compared to a variant with chlorides level 0.2, given that the other attributes of the two variants are the same?

Question 2 (15 marks) Consider the following simple regression model

$$y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1),$$

where $i = 1, \ldots, n$. The file *simplereg.csv* contains data of 30 observations from this model.

(a) Write the likelihood for this model in terms of β.

(b) The prior distribution for β is $N(0, 100)$. Derive the posterior distribution of β given the data.

(c) Write a Metropolis-Hastings algorithm to sample from the posterior of β. The proposal $Q(\beta'|\beta)$ is $N(\beta, 0.5^2)$, where β is the current value.

(d) Implement the algorithm in R and run for 20000 iterations. Discard the first 5000 iterations as burn-in. Compare the distribution of the resulting samples with your answer in part (b).

You should set a `seed` *value before running the algorithm so that your result can be reproduced and verified.*

End of Assignment — Total Available Marks = 30