THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assignment 3, 2023

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 11:59 pm on Monday 2 October 2023

This assignment consists of 4 pages (including this page) with 3 questions and 40 total marks

Instructions to Students

Writing

- Please submit a scanned or other electronic copy of your work via the Learning Management System. Your submission should be a single PDF file. You may submit an additional .R file of your code.
- You can type to 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 commands/ output in your answer.
- 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 (16 marks) The file *fat.csv* contains records of the average butterfat content (percentages) of milk for random samples of twenty cows (ten two-year old and ten mature (greater than four years old)) from each of five breeds. The variables are:

- **Butterfat**: butter fat content by percentage
- **Breed**: a factor with levels "Ayrshire", "Canadian", "Guernsey", "Holstein-Fresian" and "Jersey"
- **Age**: a factor with levels "2year" and "Mature"

Hint: When importing the data to R, you will need to specify `Breed` *and* `Age` *as factor*

(a) Fit an additive two-factor model to the data.

(b) From this model, estimate the difference between the butter fat content of two-year old cows and mature cows.

(c) Fit a linear model with interaction to the data. Calculate a confidence interval for the difference between butterfat content of mature Jersey cows and mature Guernsey cows.

(d) Test the hypothesis that the butter fat content of Canadian cows has no dependence on age.

(e) Test for the presence of interaction between age and breed of cows.

Question 2 (12 marks) Depletion of the ozone layer allows the most damaging ultraviolet radiation to reach the Earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable). To measure the relationship, researchers sampled the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. The data from this study is given in the file `ozone.csv`. There are 3 variables:

- **Inhibit**: percent inhibition of primary phytoplankton production in water
- **UVB**: UVB exposure
- **Surface**: a factor with levels "Deep" and "Surface"

Hint: When importing the data to R, you will need to specify `Surface` *as factor*

(a) Plot the percentage inhibition against UVB exposure, use different colours for observations at the surface and in the deep. Is there any evidence of an interaction?

For part b and c, you should not use `lm()` function.

(b) Using matrix calculation, fit a model for percentage inhibition with interaction between UVB exposure and depth level. You should use treatment contrasts in this question.

(c) Estimate the percent inhibition of primary phytoplankton production in water below the surface when UVB exposure is 0.02.

Question 3 (12 marks) You should not use R's `glm()` function for this question. The data frame anaethestic contains data on the concentrations, `conc`, of anaesthetic given to groups of patients of size, `m`, and the subsequent number of patients that responded satisfactorily to the anaesthetic, `satis`. Here is a table of the data frame:

conc	m	satis
0.8	7	1
1.0	5	1
1.2	6	4
1.4	6	4
1.6	4	4
2.5	2	2

We are interested in building a model for patients' response

(a) Fit a binomial regression model to the data using a logit link.

(b) Construct the 95% CIs for the parameter estimates.

(c) Perform a likelihood ratio test for the significance of the dose coefficient.

(d) Estimate the probability of satisfactory response for a patient who received anaesthetic of concentration 2.0, together with a 95% CI.

End of Assignment — Total Available Marks = 40