THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assignment 2, 2024

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 6:00 pm on Sunday 30 August 2024

This assignment consists of 3 pages (including this page) with 3 questions and 20 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 typeset 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 commands/ output in your answer, or submit your R code.
- 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 (4 marks)

Consider the full rank linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots, n.$$

Assume that the errors are independent, normally distributed with mean 0 and variance σ^2. Derive an expression for a $100(1-\alpha)\%$ confidence interval for parameter σ.

(*Hint: This question requires you to construct a CI for σ, and not σ^2.*)

Question 2 (10 marks)

A study was conducted to predict success in the early university years. Success was measured using the cumulative grade point average (GPA). The file *gpa.csv* contains the records of 100 students. The variables that we are considering are the students' GPA after three semesters, their average high school grades in Mathematics (HSM), Science (HSS) and English (HSE).

(a) Write down the formula of a linear model to predict a student's GPA at university based on their high school grades in Mathematics, Science and English.

(b) Fit the model in part (a) to the data and estimate the coefficients and variance σ^2, using the formulas in the lecture notes.

(c) Fit the model using the function `lm()` in R.

(d) Using the formulas in the lecture notes, identify the observation with the most extreme standardised residual. Calculate its corresponding leverage, and Cook's distance.

(e) Estimate the GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

For part (f) and (g), please use both matrix calculations and `R` functions.

(f) Calculate a 95% confidence interval for the expected GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

(g) Calculate a 99% prediction interval for the GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

(h) Test the hypothesis that the parameter corresponding to HSS is 0, using both t test and F test.

Question 3 (6 marks)

Consider the data set in question 2. The data also contains the students' scores from three SAT tests: SAT Mathematics (SATM), SAT Critical Reading (SATCR) and SAT Writing (SATW).

(a) Fit a linear model to predict GPA based on high school grades and SAT scores. You may use the function `lm()` in **R**.

(b) Test for model relevance using a corrected sum of squares.

(c) Starting from the full model in part (a), use stepwise selection with AIC to select variables for your model. Use this as your final model.

(d) Test whether the parameters corresponding to HSM and SATM are equal.

End of Assignment — Total Available Marks = 20