THE UNIVERSITY OF MELBOURNE

Student number

Semester 1 Assessment, 2023

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 2

Submission deadline: **Friday April 28, 5pm**

This assignment consists of 4 pages (including this page) with 5 questions and 40 total marks

Instructions to Students

Writing

- This assignment is worth 7% of your total mark.
- You may choose to either typeset your assignment in LaTeX, or handwrite and scan it to produce an electronic version.
- You may use R for this assignment, including the `lm` function unless otherwise specified. If you do, include your R commands and output.
- Write your answers on A4 paper. Page 1 should only have your student number, the subject code and the subject name. Write on one side of each sheet only. 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.
- Submit your scanned assignment as a single PDF file and carefully review the submission in Gradescope. Scan again and resubmit if necessary.

Question 1 (4 marks)

Prove the formula on slide 126 of chapter 4: that is,

$$\frac{y^* - (\mathbf{x}^*)^T\mathbf{b}}{s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}}.$$

has a t distribution with $n - p$ degrees of freedom.

Question 2 (11 marks)

We wish to predict the price of apartments in Melbourne using some of their features. Let y be the apartment price per square metre, x_1 be the apartment age (in years), x_2 be the distance (in metres) to the nearest train station, and x_3 be the number of convenience stores nearby. The following data is collected:

x_1 (years)	x_2 (meters)	x_3	y (\$, $\times 10^2$)
32	84.9	10	37.9
19.5	306.6	9	42.2
13.3	562.0	5	47.3
13.3	562.0	5	43.1
5	390.6	5	54.8
7.1	2175.0	3	47.1
34.5	623.5	7	40.3

For this question, you may not use the `lm` function in R.

(a) Fit a linear model to the data and estimate the parameters and error variance.

(b) Calculate 95% confidence intervals for the parameters.

(c) Calculate a 90% prediction interval for the price per square metre of a 5 year old apartment that is 100 meters away from the nearest train station and has 6 convenience stores nearby.

(d) Test the hypothesis that the price per square metre falls by \$100 for every year that the apartment ages, at the 5% significance level.

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

Question 3 (5 marks)

Show that for a full rank linear model with p parameters, the Akaike's information criterion, defined as $-2\log(\text{Likelihood}) + 2p$, can be written as

$$n\log\left(\frac{SS_{Res}}{n}\right) + 2p + \text{const.}$$

Question 4 (12 marks)

In this question, we study a dataset of 50 US states. This dataset contains the variables:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS.Grad`: percentage of high-school graduates (1970)
- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

The dataset is distributed with R. Open it with the following commands:

```
> data(state)
> statedata <- data.frame(state.x77, row.names=state.abb, check.names=TRUE)
```

We wish to use a linear model to model the murder rate (`Murder`) in terms of the other variables.

(a) Plot the data and comment. Should we consider any variable transformations?

(b) Perform model selection using forward selection, using all variable transformations that may be relevant.

(c) Starting from the full model, perform model selection using stepwise selection with the AIC.

(d) Write down your final fitted model (including any variable transformations used).

(e) Produce diagnostic plots for your final model and comment.

Question 5 (8 marks)

For ridge regression, we choose parameter estimators $\mathbf{b}$ which minimise

$$\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2,$$

where λ is a constant penalty parameter.

(a) Show that these estimators are given by

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}.$$

(b) Calculate the ridge regression parameter estimates for the data from Q2 with penalty parameter $\lambda = 1.5$. In order to avoid penalising some parameters unfairly, we must first scale every predictor variable so that it is standardised (mean 0, variance 1), and centre the response variable (mean 0), in which case an intercept parameter is not used. (*Hint:* This can be done with the `scale` function).

(c) One way to calculate the optimal value for the penalty parameter is to minimise the AIC. Since the number of parameters p does not change, we use a slightly modified version:

$$AIC = n \ln \frac{SS_{Res}}{n} + 2\, df,$$

where df is the "effective degrees of freedom" defined by

$$df = tr(H) = tr(X(X^T X + \lambda I)^{-1} X^T).$$

For the data from Q2, construct a plot of λ against AIC. Thereby find the optimal value for λ.

End of Assignment — Total Available Marks = 40