



Semester 1 Assessment, 2021

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 2

Submission deadline: **Friday April 30, 5pm**

This assignment consists of 4 pages (including this page)

Instructions to Students

Writing

- There are 5 questions with marks as shown. The total number of marks available is 40.
- This assignment is worth 7% of your total mark.
- You may choose to either typeset your assignment in L^AT_EX or handwrite and scan it to produce an electronic version.
- You may use R for this assignment, including the `lm` function unless 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 the page.

Scanning

- 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. Check PDF is readable.

Submitting

- Go to the Gradescope window. Choose the Canvas assignment for this assignment. Submit your file as a single PDF document only. Get Gradescope confirmation on email.
- It is your responsibility to ensure that your assignments are submitted correctly and on time, and problems with online submissions are not a valid excuse for submitting a late or incorrect version of an assignment.

Question 1 (4 marks)

Prove Theorem 4.8: show that the maximum likelihood estimator of the error variance σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n}.$$

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	x_2	x_3	$y (\times 10^3)$
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.

- Fit a linear model to the data and estimate the parameters and variance.
- Find a 90% confidence interval for the expected price per square metre of a 10 year old apartment that is 100 meters away from the train station and has 6 convenience stores nearby.
- Find the standard error of $\beta_1 - \beta_3$.
- Test the hypothesis that the price per square metre falls by \$1000 for every year that the apartment ages, at the 5% significance level.
- Test for model relevance using a corrected sum of squares.

Question 3 (5 marks)

Consider two full rank linear models $\mathbf{y} = X_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1$ and $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2$, where all predictors in the first model ($\boldsymbol{\gamma}_1$) are also contained in the second model ($\boldsymbol{\beta}$). Show that the SS_{Res} for the first model is at least the SS_{Res} for the second model.

Question 4 (10 marks)

In this question, we study the `mtcars` dataset. This dataset contains data published by the US magazine *Motor Trends* in 1974, on fuel consumption of cars for 32 different models. It includes the variables:

- `mpg`: miles/(US) gallon
- `disp`: displacement (cu. in.)
- `hp`: gross horsepower
- `drat`: rear axle ratio
- `wt`: weight (1000 lbs)
- `qsec`: 1/4 mile time

The dataset is distributed with R. Open it, select the appropriate variables, and take a logarithmic transformation of the data with the following commands:

```
> data(mtcars)
> mtcars = log(mtcars[, c(1,3:7)])
```

We wish to use a linear model to model `mpg` in terms of the other variables.

- (a) Plot the data and comment.
- (b) Perform model selection using forward selection.
- (c) Starting from the full model, perform model selection using stepwise selection with AIC.
- (d) Write down the final fitted model from stepwise selection.
- (e) Produce diagnostic plots for your final model from stepwise selection and comment.

Question 5 (10 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) Show that \mathbf{b} is biased if $\lambda \neq 0$.
- (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

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

We will use the data from Q2. In order to avoid penalising some parameters unfairly, we must first standardise the variables; this also means an intercept parameter is not used. You can do this with `scale`:

```
> X <- scale(X[, -1], center=T, scale=T)
> y <- scale(y, center=T, scale=T)
> p <- 3
```

Construct a plot of λ against AIC. Thereby find the optimal value for λ .

End of Assignment — Total Available Marks = 40