MAST90104: Introduction to Statistical Learning

Assignment 2

Late assignments will be penalised two marks for each day overdue.
Extensions should be supported by a medical certificate. Requests for extensions must come on or before the due date for the assignment.

Please submit a scanned or other electronic copy of your work via the Learning Management System in one file - see this link for instructions

The .pdf must have in **one file**:

- handwritten or typed answers to the questions
- handwritten or typed R code used to produce your answers
- graphics required to answer the questions

If you have more than one file submitted, *only the last LMS .pdf file with your name on it will be marked.*

This assignment is worth 5% of your total mark.

You may use R for this assignment, including the `lm` function unless otherwise specified.
Include your R commands and output.

1. Consider a general full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $p > 2$ parameters. Derive an expression for a joint $100(1-\alpha)\%$ confidence region for parameters β_i and β_j, where i and j are arbitrary.

2. An experiment is conducted to estimate the annual demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained:

Cars sold ($\times 10^3$)	Cost ($$k$)	Unemployment rate (%)	Interest rate (%)
5.5	7.2	8.7	5.5
5.9	10.0	9.4	4.4
6.5	9.0	10.0	4.0
5.9	5.5	9.0	7.0
8.0	9.0	12.0	5.0
9.0	9.8	11.0	6.2
10.0	14.5	12.0	5.8
10.8	8.0	13.7	3.9

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 variance.

(b) Which two of the parameters have the highest (in magnitude) covariance in their estimators?

(c) Find a 99% confidence interval for the average number of $8,000 cars sold in a year which has unemployment rate 9% and interest rate 5%.

(d) A prediction interval for the number of cars sold in such a year is calculated to be $(4012, 7087)$. Find the confidence level used.

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

3. For this question we use the data set `UCD.csv` (available on the LMS). This data set, collected on 158 UC Davis students (self-reported), includes the following variables:

ID = the ID for that student

alcohol = average number of alcoholic drinks consumed per week

exercise = average hours per week the student exercises

height = the student's height (in inches)

male = indicator variable, 1 if male and 0 if female

dadht = the student's father's height

momht = the student's mother's height

We seek to predict a person's height, based on the given data.

(a) Fit a linear model using all of the variables (except ID).

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

(c) Use forward selection with F tests to select variables for your model.

(d) Starting from a full model, use stepwise selection with AIC to select variables for your model. Use this as your final model; comment briefly on the variables included.

(e) Test whether the parameters corresponding to father's and mother's heights are equal.

(f) Comment on the suitability of your final model, using diagnostic plots.

4. A study was conducted to determine the effect of the size of the root system on the growth of Douglas-fir seedlings when they are planted out. Seedlings were obtained from three seed lots, and when they were planted out their root volume was classified as small (RV1), medium (RV2), or large (RV3). The heights of the seedlings were then measured at the end of the first growing season. The data from the experiment is given in the file `douglas.csv`.

(a) Fit a linear model with interaction to the data. Calculate a confidence interval for the difference between the heights of large (RV3) and medium (RV2) seedlings in the B349 seed lot.

(b) Is it possible to estimate, from this model, an overall difference between the heights of large and medium seedlings?

(c) Test the hypothesis that the height of seedlings from the J052 plot has no dependence on root volume.

(d) Generate an interaction plot for the data. Is there any evidence of an interaction?

(e) Test for the presence of interaction between root volume and seed lot.

(f) Perform forwards selection to determine a final model.

(g) Is it possible to estimate, from this model, an overall difference between the heights of large and medium seedlings?

5. You wish to perform a study to determine if 3 treatments each produce no effect using a completely randomised design. To do this, you will test the hypothesis $H_0 : \mu + \tau_1 = \tau_1 - \tau_2 = \tau_2 - \tau_3 = 0$. You are given resources to study 50 sample units.

(a) Determine the optimal allocation of the number of units to assign to each treatment.

(b) Perform the random allocation. You must use R for randomisation and include your R commands and output.