MAST30025: Linear Statistical Models

Assignment 3, 2019

Due: 5pm Friday, May 31 (week 12)

- This assignment is worth 7% of your total mark.
- You may use R for this assignment, including the `lm` function unless specified. If you do, include your R commands and output.
- Your assignment must be submitted to Turnitin on the LMS as a single PDF document only. You may choose to either typeset your assignment or handwrite and scan it to produce an electronic version. Turnitin will not accept late submissions.
- Turnitin gives you an option to preview your work prior to submission. Please check this preview carefully to ensure you are submitting the correct document. After a successful submission to Turnitin, you will see a submission ID. This confirmation will also be sent to your University email address. If you do not see a submission ID, you should assume that your assignment has not been submitted successfully. Either try to submit again or contact the tutor co-ordinator (Rheanna Mainzer) immediately to arrange an alternate means of submission. Issues with Turnitin are not a valid excuse for submitting a late assignment or an incorrect version of an assignment.
- **(1 mark)** Your assignment must clearly show your name and student ID number, your tutor's name and the time and day of your tutorial class. Your assignment must be submitted in the correct format and the correct orientation. Your answers must be clearly numbered and in the same order as the assignment questions.

1. Let A be an $n \times p$ matrix with $n \geq p$.
   (a) Show that $r(A^c A) = r(A)$.
   (b) Show that $I - A(A^T A)^c A^T$ is idempotent.
   (c) Show that $r(I - A(A^T A)^c A^T) = n - r(A)$.

2. We are interested in examining the yield of tomato plants that have been grown with certain types of fertiliser. A study is conducted and the following data obtained:

Fertiliser		
1	2	3
43	33	56
45	37	54
47	38	57
46	35	
48		

   We fit the model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

   where μ is the overall mean and τ_i is the effect of using the ith fertiliser.

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

   (a) Find a conditional inverse for $X^T X$, using the algorithm given in Theorem 6.2.
   (b) Characterise all solutions to the normal equations.
   (c) Is $4\mu + 2\tau_1 + \tau_2 + \tau_3$ estimable?

(d) Find a 95% prediction interval for the yield of a tomato plant grown on fertiliser 1.

(e) Test the hypothesis that fertilisers 2 and 3 have no difference in yield.

3. Consider a linear model with only categorical predictors, written in matrix form as $\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$. Now suppose we add some continuous predictors, resulting in an expanded model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

   Now consider a quantity $\mathbf{t}^T\boldsymbol{\beta}$, where $\mathbf{t}^T = [\mathbf{t}_1^T|\mathbf{t}_2^T]$ is partitioned according to the categorical and continuous predictors. Show that if $\mathbf{t}_1^T\boldsymbol{\beta}_1$ is estimable in the first model, then $\mathbf{t}^T\boldsymbol{\beta}$ is estimable in the second model.

   If you write $X = [X_1|X_2]$, you may assume that $r(X) = r(X_1) + r(X_2)$.

   Hint: Use Theorems 6.9 and 6.3. For any vectors $\mathbf{z}_1$ and $\mathbf{z}_2$, you can write

$$\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & X_1^TX_1\mathbf{z}_1 \\ X_2^TX_1 & X_2^TX_2 & X_2^TX_2\mathbf{z}_2 \end{array}\right] = \left[\begin{array}{cc} X_1^T & \mathbf{0} \\ \mathbf{0} & X_2^T \end{array}\right] \left[\begin{array}{cc|c} X_1 & X_2 & X_1\mathbf{z}_1 \\ X_1 & X_2 & X_2\mathbf{z}_2 \end{array}\right].$$

4. Data was collected on the world record times (in seconds) for the one-mile run. For males, the records are from the period 1861–1999, and for females, from the period 1967–1996. The data is given in the file `mile.csv`.

   (a) Plot the data, using different colours and/or symbols for male and female records. Without drawing diagnostic plots, do you think that this data satisfies the assumptions of the linear model? Why or why not?

   (b) Test the hypothesis that there is no interaction between the two predictor variables. Interpret the result in the context of the study.

   (c) Write down the final fitted models for the male and female records. Add lines corresponding to these models to your plot from part (a).

   (d) Calculate a point estimate for the year when the female world record will equal the male world record. Do you expect this estimate to be accurate? Why or why not?

   (e) Is the year when the female world record will equal the male world record an estimable quantity? Is your answer consistent with part (d)?

   (f) Calculate a 95% confidence interval for the amount by which the gap between the male and female world records narrow every year.

   (g) Test the hypothesis that the male world record decreases by 0.3 seconds each year.

5. You wish to perform a study to compare 2 medical treatments (and a placebo) for a disease. Treatment 1 is an experimental new treatment, and costs \$5000 per person. Treatment 2 is a standard treatment, and costs \$2000 per person. Treatment 3 is a placebo, and costs \$1000 per person. You are given \$100,000 to complete the study. You wish to test if the treatments are effective, i.e., $H_0 : \tau_1 = \tau_2 = \tau_3$.

   (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.