MAST90104: Introduction to Statistical Learning

Week 7 Lab and Workshop

1. In a manufacturing plant, filters are used to remove pollutants. We are interested in comparing the lifespan of 5 different types of filters. Six filters of each type are tested, and the time to failure in hours is given in the dataset (on the website) `filters` (in `csv` format).

   (a) Use the `read.csv` function to read the data. Then convert the `type` component into a factor.

   (b) Using only the `matrix` command, two different X and `y` matrices for this linear model both of which are full rank, one corresponding to `cont.treatment` and one to `contr.sum`.

   (c) Fit the models and compare with the `lm` output.

   (d) Calculate s^2 using the residuals

   (e) Calculate a 95% confidence interval for the difference in lifespan between filter types 3 and 4.

   (f) Show that the hypothesis that the filters all have the same lifespan is testable.

   (g) Test this hypothesis, using matrix theory.

   (h) Test the same hypothesis using the `linearHypothesis` function from the `car` package.

2. An industrial psychologist is investigating absenteeism among production-line workers, based on different types of work hours: (1) 4-day week with a 10-hour day, (2) 5-day week with a flexible 8-hour day, and (3) 5-day week with a structured 8-hour day. A study is conducted and the following data obtained of the average number of days missed:

	Work plan		
	1	2	3
Mean	9	6.2	10.1
Number	100	85	90

   They also find $s^2 = 110.15$.

   (a) Test the hypothesis that the work plan has no effect on the absenteeism.

   (b) Test the hypothesis that work plans 1 and 3 have the same rate of absenteeism.

3. We study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

   (a) Express this as a two-factor model with no interaction in matrix form.

   (b) Express this as a two-factor model with interaction in matrix form.

   (c) Express the hypothesis that there is no interaction in terms of your parameters. Eliminate any redundancies.

   (d) Input this data into R. Plot an interaction plot between breed and diet.

   (e) Test for the presence of interaction.

   (f) What is the degrees of freedom used for the interaction test?

   (g) From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3?

   (h) Fit an additive model. What is the estimated amount of milk produced from breed 2 and diet 3 now?

(i) Test the hypothesis (under the additive model) that the 2nd and 3rd diets are equivalent in terms of milk produced.

(j) Find a 95% confidence interval, under the additive model, for the amount of milk produced from breed 2 and diet 3. Use both matrix calculations and the `estimable` function from the `gmodels` package.

(k) Find the same confidence interval under the interaction model.

(l) Why is the second interval wider than the first?

4. Suppose each row of a dataset has a response variable and two factors, which have 2 and 3 possible levels respectively. The dataset has 2 rows for each possible combination of factor levels. We model this with a less than full rank model with one parameter for the overall mean, and one parameter for each level of each factor, assuming that the overall mean is adjusted additively by each factor. Write down the linear model in both equation and matrix form.

5. Let

$$A = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ 0 & 1 & -3 & -2 \end{bmatrix}.$$

(a) Show that $r(A) = 2$.

(b) Construct two different full rank matrices which generate the same column space as A.

6. It is known that toxic material was dumped into a river that flows into a large salt-water commercial fishing area. We are interested in the amount of toxic material (in parts per million) found in oysters harvested at three different locations in this area. A study is conducted and the following data obtained:

Site 1	Site 2	Site 3
15	19	22
26	15	26

(a) Write down the linear model in matrix form.

(b) Write down the normal equations.

(c) Reparameterize the model to a full rank model.

(d) Find a solution for the normal equations.

7. In the one-way classification model, show that any linear combination of $\bar{y}_1 - \bar{y}_\cdot, \ldots, \bar{y}_k - \bar{y}_\cdot$ can be written as a linear combination of $\bar{y}_1, \ldots, \bar{y}_k$. Does the converse hold?