MAST90044 Thinking and Reasoning with Data
Semester 1 2019
Assignment 3
Due: 8 AM, Monday 27 May

- Assignments are to be submitted (uploaded) via LMS.
- Please label your assignment with the following information:
  - your name;
  - your student number;
  - the day of your lab class;
- This assignment is worth 20% of the marks in this subject, and covers the work done in weeks 8 to 10.
- The total number of marks for this assignment is 62.
- Your assignment should show all working and reasoning, as marks will be given for method as well as for correct answers. Please spell check your document.
- Paste any R code and output into the appropriate places so that it can be seen easily along with your other work. Graphics from R can be resized within your document; make them smaller as necessary. Tutors will not help you directly with assignment questions. However, they may give some help with R.
- Solutions to the assignment questions will be made available later
- Please note that only some questions maybe marked.

Q1 Counts of fibres in skeletal tissue were made on 25 rats. There are two types of fibres: Type I and Type II. Type I fibres are further divided into three categories: (i) reticulated; (ii) punctate; and (iii) both reticulated and punctate. The aim was to set up a model which predicts the number of Type II fibres from the numbers of the three different categories of Type I fibre. Part of the data are shown below. The entire data set is in `rat_fibres.csv` on the LMS.

	Number of Type I fibres			Number of
rat	Reticulated	Punctate	Both	Type II Fibres
1	1	13	5	15
2	2	8	4	12
3	9	27	16	46
⋮	⋮	⋮	⋮	⋮
25	0	4	3	6

(a) Construct an appropriate graph of these data. Comment on the apparent strength or otherwise of the relationship between the response variable and each of the explanatory variables.

(b) Fit a multiple regression which includes all the explanatory variables, and then use tests of significance to determine the "best model". State the fitted equation of the best model, and give an interpretation of the coefficient of the most significant explanatory variable in that model.

(c) Report the value and the P-value of the F statistic for the "best" model. What this statistics is testing? what your conclusion for this model.

(d) Examine diagnostic plots for the best model. On the plot of `type II` vs `punctate` constructed in (a), circle two points that are important in the diagnostic plots, and explain their importance.

(e) Find the residual for rat number 8. Does this observation contribute more or less than expected to the variation around the fitted model?

(f) Using AIC as the criterion for comparing models, determine the most appropriate model using backward elimination. Comment on whether this model confirms or contradicts the "best model" found in (b).

(g) Using adjusted R^2 as the criterion for comparing models, determine the most appropriate model from the 3 you examined. Taking everything into consideration, which model would you adopt? Briefly explain.

[2 + 7 + 4 + 4 + 3 + 3 + 2 = 25 marks]

Q2 A study investigated the condition of two different arteries, harvested from 110 candidates for coronary artery bypass surgery: the radial artery (RA) from the arm, and the internal thoracic artery (ITA) from the chest. In this question, you will look at the outcome "medial calcification of the radial artery" (yes/no), and patient factors which may affect this outcome. It is an important outcome, because radial arteries with medial calcification may not be suitable for use in bypass surgery.

The data is in the file `graft_arteries.csv` on the LMS. Read the data into R. The name of the outcome (or response variable) has been shortened to `RAcalc`.

(a) We will consider two explanatory variables – presence or absence of diabetes, and presence or absence of hypertension. Using `table()`, create a table of the frequencies of the 8 combinations of the response variable and the two explanatory variables. Using these frequencies, create a data frame suitable for logistic regression. The first of the 4 rows of this data frame should look something like this:

```
  Diabetes Hypertension RAcalc total
1       no           no      4    41
```

(b) Firstly consider the effect of diabetes. Find an estimate for the difference between the percentage with RA medial calcification in those with diabetes, and those without diabetes. Carry out a corresponding hypothesis test and briefly describe how you would report the result.

(c) Fit a logistic regression model with RA medial calcification as the outcome, and diabetes as the explanatory variable. What is the estimated odds, comparing those with diabetes and those without?

(d) Now fit a logistic regression, with both diabetes and hypertension as explanatory variables. Test the significance of the effect of hypertension when it is added to the model with diabetes.

(e) In this data set, were the people with diabetes older, or younger, on average, than those without diabetes? It is suspected that age is associated with RA medial calcification; check if the data support this (do this simply, with summary statistics, not significance tests). If so, what are the implications for interpreting the effect of diabetes on RA medial calcification? How could this be tested?

[6 + 4 + 4 + 3 + 5 = 22 marks]

Q3 Twelve sheep are available for an experiment set up to test a new hoof hardening mixture. There are just 2 treatments in the experiment—the new mixture, and the control (no mixture). The mixture is applied by painting it onto the hooves of the sheep. For the control, water is painted onto the hooves. At the conclusion of the experiment, a digital hardness reading is taken on each hoof of each sheep.

For each of the experimental designs below, state what the experimental unit is, any blocking factors, and any flaws in the design (statistically unsound aspects).

(i) Six sheep are randomly chosen to receive the new mixture.

(ii) All twelve sheep receive the new mixture, and the results are compared to measurements taken before the experiment.

(iii) The sheep are placed in order according to body condition. The two sheep in the best condition are taken and the new mixture randomly allocated to one of them. Similarly for the two next best sheep, and so on.

(iv) The six sheep with the worst body condition receive the new mixture, to avoid overestimating its beneficial effect.

(v) The left or right hooves of each sheep are randomly chosen to receive the new mixture.

[3 + 3 + 3 + 3 + 3 = 15 marks]

Total marks = 62