Assignment 1, Math 3346, 2008

Lecturer: John Maindonald

August 1, 2008

This exercise will work with experimental control and treatment groups in the data set `nswdemo`. The dataset `nswdemo` is included in the *DAAGxtras* package. Be sure to examine its help page. Data are from a randomized trial that was designed to assess whether a work training program helped the income prospects of individuals who had a history of employment difficulties.

1. Write brief notes on each of the columns in the data frame, noting whether columns should be treated as numeric or categorical, noting ranges of values in numeric columns, and noting numbers of NAs. For the variables that record incomes, note numbers of zeros.
   [2 marks]

2. For each column of the data and for each of the two treatment groups, do the following:

   (a) Determine the number of missing values.
   [$\frac{1}{2}$ mark]

   (b) Determine, for each of `re75` and `re78`, the number and proportion that are zero.
   [$\frac{1}{2}$ mark]

3. Now examine `re74`, comparing control and treatment data:

   (a) Compare the proportion of `NA`s between control and treatment.
   [$\frac{1}{2}$ mark]

   (b) Compare the proportion of 0's (obviously `NA`s have to be excluded) between control and treatment.
   [$\frac{1}{2}$ mark]

   (c) Limiting attention to nonzero (and non-`NA`) values, use the method of Subsection 6.3.2 in Statistical Perspectives on Data Mining to compare the distribution of `log(re74)` between control and treatment observations. Why use `log(re74)`, rather than `re74`, for the comparison?
   [3 marks]

4. Provide graphs that conveniently summarise differences between the three groups, with respect to `age` and `re75`. Issues to consider, and on which you should comment, are:

   - Is it best to examine separately i) comparisons with respect to number of zeros, and ii) distributions of non-zero values, rather than using one density plot for both? [Both approaches can be defended, depending however on the audience.]
   - For `re75`, is a logarithmic scale preferable?
   - If zeros are included in a probability density plot for the logged values, it will be necessary to add a small positive offset before taking logarithms. What magnitude of offset is sensible?

   NB: Marks will be given for layout, with a preference for a layout that lays information out in a compact and readily comprehended form.
   [4 marks]

5. Use tables to summarise differences in categorical variables between the two groups.
   [2 marks]

6. What are the major differences between the two groups, as evident from examining columns one at a time? Comment especially on any differences in the pre-training variables, i.e., all except `re78`.
   [3 marks]

7. Do any differences in the pre-training variables have implications for the way that you might analyse the data, or the reliance that you might put on the results?
   [2 marks]

8. The aim of the study was to assess the effect of training. Is it best to base the comparison between treatment and control group i) on `re78` alone, or ii) on `re78 - re75`? Justify your answer.
   [2 marks]

[**TOTAL: 20 marks**]

Due Date: August 22, 2008, 5pm

In addition to any R code that may be included in the main document, please provide the R code separately from the output. Marks will be subtracted if the R code is not provided.

Please provide assigments in a pdf file, either as hard copy or emailed to john.maindonald@anu.edu.au