Assignment 1, Math 3346, 2009

Lecturer: John Maindonald

August 4, 2009

This exercise will work with experimental control and treatment groups in the data set nswdemo, and with non-experimental comparison groups in the data sets cps1 and psdi1. All these datasets are included in the *DAAG* package.

Data in nswdemo 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. The remaining two datasets are comparison datsets from neighboring areas. They have been investigated, in the econometric literature, for possible use as non-experimental controls.

1 Preliminaries

A first step is to attach the package DAAG that has all the data:

```
> library(DAAG)
```

Creation of a Suitable Data Object

Here, we will create one large dataframe that holds all the data. (An alternative, described below, is to form a list in which each separate group is a different list element.)

```
> ## Note the coding for the experimental control & treatment groups
> table(nswdemo$trt)
> ## Code the non-experimental controls as 2 (cps1) and 3 (psid1)
> cps1$trt <- rep(2, dim(cps1)[1])
> psid1$trt <- rep(3, dim(psid1)[1])
> allsets <- rbind(nswdemo, cps1, psid1)
> allsets$trt <- factor(c("exp-ctl", "exp-trt", "cf-cps1", "cf-psid1")
+                      [allsets$trt+1])
> ## Check that the numbers in the different groups make sense
> table(allsets$trt)
```

Counts of number of missing values

The following creates functions that can be used for various counting tasks:

```
> ## Function that retrieves number of NAs for a single column
> countNAcol <- function(x)sum(is.na(x))
> ## Now use the function aggregate() to apply countNAcol() to
> ## allsets, indexed by trt
> aggregate(allsets, list(group = allsets$trt), FUN = countNAcol)
```

Notice that we can replace countNAcol by any other function that returns a single value. The following counts the number of zeros in each of the columns re74, re75 and re78.

```
> countzeros <- function(x)sum(!is.na(x) & x==0)
> aggregate(allsets[, c("re74", "re75", "re78")], list(group=allsets$trt),
+           FUN=countzeros)
```

Alternative: The following alternatives use a further call to sapply(), in place of use of aggregate():

- Use sapply() with the allsets data frame that was created above:

  ```
  > ## Function that retrieves number of NAs for each column of a data frame
  > countlist <- function(z, statsfun=length)sapply(z, statsfun)
  >   # Notice that statsfun has been given a default argument
  > ## Apply this to the allsets data frame
  > sapply(split(allsets, allsets$trt), FUN=countlist, statsfun=countNAcol)
  ```

  Notice that, as used here, sapply() has three arguments. The first is the list of data frames to which the function given as the second argument will, in turn, be applied. The third argument, statsfun=countNAcol, is passed to countlist(), thus supplying the second of its two arguments.

- Create a list in which the data frame corresponding to each group is a separate list element:

  ```
  > listsets <- list("cf-cps1"=cps1[,-1], "cf-psid1"=psid1[,-1],
  +                  "exp-ctl"=subset(nswdemo, trt==0)[,-1],
  +                  "exp-trt"=subset(nswdemo, trt==1)[,-1])
  > ## Check that the numbers in the different groups make sense
  > sapply(listsets, nrow)
  > ## Now do the calculations with listsets
  > sapply(listsets, FUN=countlist, statsfun=countNAcol)
  ```

Use of sapply() in this way is actually more general. For example, a possible argument is statsfun=range, so that two values are returned. The labeling is then less informative than one might like!

2 Exercises

Here then are the formal exercises:

1. Look up help(nswdemo), help(cps1) and help(psid1). Try also

   ```
   > str(nswdemo); str(cps1); str(psid1)
   ```

   Use these sources of information to write brief notes on each of the columns of data, noting whether columns should be treated as numeric or categorical.
   [1 mark]

   Write brief notes documenting each of the functions countNA(), countzeros() and countlist()
   [2 marks]

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

   (a) Determine the number of missing values.
   [1 mark]

   (b) Determine, for each of re74, re75 and re78, the number and proportion that are zero.
   [1 mark]

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

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

(b) Compare the proportion of 0's in each of `re74`, `re75` and `re75` (obviously `NA`s have to be excluded) between the four groups.
[$\frac{1}{2}$ mark]

(c) For columns that are numeric with more than two unique values, determine for each of the four groups the range of values.
[1 mark]

4. Provide graphs that conveniently summarise differences between the four 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 four 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 25, 2009, 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