Aarhus, 6. November 2013

Postgraduate course in
Applied Statistical Analysis with Missing Data
November 27-29, 2013

Homework

Part A

The four data sets `m0_bw.dta`, `m1_bw.dta`, `m2_bw.dta` and `m3_bw.dta` contain data from studies that all include 10 000 births. We will here focus on estimating the relationship between birth weight and smoking (using the categorization of smoking pre-coded in the data).

Do the following for each data set:

1. Make a table containing the number of missing observations for each variable in the data (use for example `-codebook, compact-`).
   Comment on the number and pattern of missingness across variables.

2. Estimate the unadjusted relationship between birth weight and smoking.
   How many observations were used in this analysis?

3. Estimate the relationship between birth weight and smoking while adjusting for sex, parity, alcohol and BMI.
   How many observations were used in this analysis?

Make a table with all four sets of unadjusted estimates and their confidence intervals, and comment on your findings. Repeat this for the adjusted estimates.

Part B

The `DrugStudy.dta` contains data from a trial comparing two drugs.
Read and 'execute' the three do-files `DrugStudy1.do`, `DrugStudy2.do` and `DrugStudy3.do`. Familiarize yourself with the output.

Part C
In this exercise you will learn how to generate data in Stata which exhibits random variation of certain kinds.

1. First create an empty dataset with 1,000 observations:

   ```
   . clear
   . set seed ddMM
   . set obs 1000
   ```

   where *ddMM* is to be replaced with your date of birth – this ensures that your random numbers are not identical to others as long as birth dates differ.
   Next you should create artificial ID-numbers:

```
. generate pid = _n
```

Now create a variable with 1000 random numbers, uniformly distributed on the interval (0; 1):

```
. generate unifvar = runiform()
```

Make a histogram to see the distribution of the new variable, and try sorting your dataset on the new variable – how would you describe the order of ID-numbers after you have done this sorting?

2. Let us now generate a random age for each individual, where we assume the distribution to be uniform between 40 and 59 years (i.e. it is just as likely for a person to be 45 as 55 – and any other age in the interval):

```
. generate age = int(runiform() * 20 + 40)
```

Examine the generated age distribution.

3. Assume that systolic blood pressure of these individuals is normally distributed with a standard deviation of 15 mmHg and a mean which is 125 mmHg for 40 year olds and then increases with 1.2 mmHg for each year a person is older. This can be generated with the following command:

```
. generate systolic = 125 + 1.2 * (age - 40) +
                                    rnormal() * 15
```

Plot systolic blood pressure against age, and fit a regression which estimates the relationship between age and systolic blood pressure.

4. When blood pressure becomes too high, patients may start treatment to lower it. We will now generate a binary variable which records treatment status, and we will let the probability of being treated increase with blood pressure:

```
. generate logodds = (systolic - 130) / 40
. generate treatment = logit(runiform()) < logodds
```

Create suitable categories of age and tabulate treatment status against age categories – why does treatment status appear to depend on age?

5. Run the following two regressions to estimate how blood pressure depends on age, and how treatment status depends on age and systolic blood pressure:

```
. regress systolic age
. logit treatment systolic age
```

The above generation of an artificial dataset is known as *stochastic simulation*, which is an integral part of multiple imputation.