Exercise 3: Creating and Combining Datasets

Launch Stata, open a new do-file and save as *Stata_Exercise3.do*. Add appropriate comments at the beginning of the do-file. Run through these exercises referring to chapter 4 in the module notes.

Exercise 3.1: Importing Data from Excel

- Firstly, open the *bl_medhis.xls* and *fup_endpoints.xls* Excel files. Make a note of the data structure (e.g. are there variable names, etc) and worksheet names.
- Continuing in the same do-file write commands to load the data from the two worksheets in *bl_medhis.xls*.
- As you import look at the data in the Data Browser.
- Save each file as a Stata dataset with the same name as the worksheet.
- Import the data from the *fup_endpoints.xls* file.
- Save as a Stata dataset with the name *fup_endpoints.dta*.

Exercise 3.2: Importing Delimited Text Files

- Firstly, open *bl_labs1.txt* in a text file editor (or use the type command). Make a note of the format of the text file.
- Load this dataset into Stata using the `import` command.
- Look at the data in the Browser before saving this file as *bl_labs1.dta*.
- Repeat for *bl_labs2.txt*, *bl_labs3.txt*, *bl_labs4.txt*.
- Repeat the above process for *bl_rand.txt* and *fup_vitals1-4.txt*.

Exercise 3.3: Importing Free Format Text Files

- Firstly, open the *bl_meds.txt* in a text file editor (or use the type command). Make a note of the format of the text file. The file contains a patient identifier and 10 fields containing names of medications taken at baseline.
- Load this file into Stata using the `infile` command. Name the 10 medication variables as med1, med2 etc.
- Check the data has loaded correctly and then save as Stata data file.
- Repeat for the *fup_meds.txt* file.

Exercise 3.4: Importing Fixed Format Text Files

- Firstly, open the *trtcodes.txt* in a text file editor (or use the type command). Make a note of the format of the text file and the variables. The file contains 3 variables – randomisation code randomisation date and treatment group (trt).
- Load this dataset into Stata using the `infix` command. Name the variables rcode, randdate and trt. Save as *trtcodes.dta*.
- Repeat for *lab_dates.txt*. The file contains 5 variables: patient id (9 characters), sex (M or F), age-group (0, 1, 2 or 3), visit number (01, 02, etc.) and visit date (YYYYMMDD).

- Save as *lab_dates.dta* (i.e. as a Stata dataset).

Exercise 3.5: Appending Data Files

- Append the four baseline laboratory datasets (*bl_labs1-4.dta*). Before doing so check the variable names are consistent in each of the datasets.
- Having appended the four files save as *bl_labs.dta*.
- Repeat for *fup_vitals1-4.dta*. Save as *fup_vitals.dta*.

Exercise 3.6: Merging Data Files

- Load the *bl_demog.dta* dataset into memory.
- Merge on the following files in order:
    - o *bl_medhis1.dta*
    - o *bl_medhis2.dta*
    - o *bl_labs.dta*
    - o *bl_rand.dta*
    - o *trtcodes.dta*
- Save the combined file as *bl_combined.dta*.

Merging part of a dataset

- Load the *fup_endpoints.dta* dataset.
- Bring in just the variables *sex*, *age* and *trt* from the *bl_combined.dta* dataset.
- Save as *fup_endpoints1.dta*.

Non-unique Merging

- Load the *fup_pot_long.dta* dataset.
- Merge on just the variables *sex*, *age* and *trt* from the *bl_combined.dta* dataset.

Merging with more than one linking variables

- Continuing with the *fu_pot_long* dataset in memory, merge on the variable *visitdate* from *lab_dates.dta* which was created in Exercise 3.5.
- Check the non-matching records. Can you see why they have not matched?
- Save as *fup_pot_long1.dta*.

Optional Exercise

Creating a dataset using Input

- In a small randomised controlled trial (RCT) comparing CALM-BP (58 patients) versus DASH (55 patient) in 113 patients with high blood pressure 18 patients in the CALM-BP group and 9 people in the DASH arm reported suffering from headaches during the treatment period:
    - o Use the input command to create a Stata dataset containing these results.
    - o Use the tabulate command to produce a table of treatment group versus headaches.
    - o What percentage of patients experienced headaches in the two groups? Is there evidence that these percentages differ?