An Introduction to STATA

Some things you may need to know:

1) The command **clear** clears all of the data in the current memory. You lose any changes to the current dataset that you didn't save.
2) You can **save** and **open** files in STATA very easily. Those options are in the drop down "file" menu just like they are in many other applications.
3) The command **set mem** tells STATA how much memory it will need to work with the dataset you want to open. When you open STATA, it will tell you how much memory is being allocated to the data (usually it is 1 MB). If you need to increase the memory size (say you want 9MB), just type: **set mem 9m**.
4) The page up button on your keyboard tells STATA to go to the previous command typed.

Entering a dataset into STATA:

- STATA format data files have the file extension .dta
- The data for this problem set is already in STATA format so you just need to open the file from the website and it will open in STATA
- You can also type **use** and then the filename (the whole name, e.g. "H:\...\...dta") in quotations.
- To enter data that is in text format, use the command **insheet using *filename***. **Insheet** reads text (ASCII) files where there is one observation per line and the values are separated by tabs or commas. In addition, the first line of the file can contain the variable names or not.
- To enter data from an excel spreadsheet, the easiest thing to do is to open it in excel and then save it as a text file and use the **insheet** command.
- You can view the data in the data editor.
- There are two commands for combining datasets, **append** and **merge**. **Merge** joins corresponding observations from the dataset currently in memory with those from the STATA-format dataset stored as filename into single observations (so the two datasets must have at least one variable in common, such as an ID #, social security number, etc.)

  Merge using filename

  Append appends a STATA-format dataset stored on disk to the end of the dataset in memory.

  Append using filename

Basic Commands:

- **Label var "…"** is the command to label a variable (with the label in quotation marks)
- The **sum** command (followed by the variable name) gives you summary statistics on a variable (mean, std dev, min, max and number of observations).
- The **tab** command (followed by the variable name) gives you a frequency distribution for the variable.
- The **if** command (preceded by another command) restricts the command to the data specified. E.g. to summarize a variable called wages to only women (if you have a dummy variable that equals 1 for a woman), you would write:

  ```
  Sum wages if woman==1
  ```

  (Note that a double equal sign is used.)

  You could restrict the data used further using **&**, **>**, **<, !=** (this last symbol means "is not equal to)

  ```
  Sum wages if woman==1 & age < 40
  ```

  To use an "**or**" command, use the symbol |

  ```
  Sum wages if woman==1 | age < 40
  ```

- The **gen** command creates a new variable. For example, suppose you have an age variable and you want to create a dummy variable that equals 1 for a child and 0 otherwise:

  ```
  Gen child=1 if age <= 12
  Replace child=0 if age >12
  ```

  (note the single equal sign for the creation of a variable). All values for which age is missing will be missing for child as well.
- The **keep** command, followed by a list of variables, tells STATA to drop all of the variables. If you want to drop observations, not variables, you can type:

  ```
  Keep if child==1
  ```

  STATA will drop all observations for which child equals zero or is missing.
- The **drop** command works just like **keep** except you are telling STATA which variables (or observations) to drop

Regression Analysis:

- The command for a basic, OLS regression is **reg**. The format is:

  ```
  Reg Y X
  ```

  Where Y is the dependent variable and X is the independent variable (you can list as many control variables as you want, e.g **reg wages educ age income**)
- A constant term is assumed with the reg command. If you don't want a constant term you have to type

  ```
  Reg Y X, noconstant
  ```

- The command for instrumental variables estimation is **ivreg**. The format is:

  ```
  Ivreg Y X1 (X2 = Z)
  ```

Y is the dependent variable, X1 is the independent variable, X2 is the variable you are instrumenting (the endogenous variable) and Z is the instrument.

- The **robust** command after the regression syntax specifies that robust (Huber-White) standard errors be estimated.

  Reg Y X , robust

Graphs:

- Any two-way (two-variable) graph has the following syntax:

  Twoway plottype Y X

  Where *plottype* can be any type of two-way plot (e.g. scatter, line, etc.)
- If your variables are labeled, those labels will be used on the axes of the graph
- You can change the style of the markers (e.g. make them hollow) using the command **mstyle**
- STATA has many options for graphs that you can look up under help