Introduction to Stata

Lecture II

Tomas R. Martinez

UC3M

September, 2019

Quote of the Day

"Data is the new bacon" - Unknown

Importing Data

- "Data is the new bacon" - Unknown
- First step to start working is to read your data in Stata
- Stata data files are **.dta**
- However, most of the case the data comes in other formats
  - .xls (excel)
  - .csv
  - .txt
  - .dat
- How we deal with that?

Where can we get data?

Cross-country / aggregate data

- Penn World Table: provides data on GDP, consumption, exports, price index, etc for more than 180 countries
- FRED Economic Data: Tons of time series provided by the Federal Reserve Bank of St. Louis
- UN Comtrade: Trade data, very disaggregate by products/countries
- ILO Stat: Labor data, employment and earnings, etc
- World Economic Outlook Database: The IMF data, tons about debt, currencies, commodities...
- World Bank: Covers some development topics: health, education, etc.
- And many others... For a good summary this Harvard website has many sources: here

Where can we get data?

Micro data

- IPUMS: Tons of harmonized microdata of different countries + many other US data sets
- Eurostat: Lots of micro data from European countries, in many you have to apply access but there are some of public ones too
- LIS: Harmonized income and wealth database from different countries
- PSID, NLSY, SIPP: US individual panel data
- Usually micro data is very country specific and you have to dig in around the statistical agency webpage

Importing Data

- If your data is in **.dta** is very easy
- Go on the menu: file, open and that's it
- In your do-file you just use the command **use**

Importing Excel Data

- We are all very familiar with excel and this one of the most common sources we have
- The "easy way": Ctrl+C and Ctrl+V
- **Example:** Data on income distribution
  - World Inequality Database (Piketty Data): `https://wid.world/`
  - Spanish data from Top 1%, 5% and 10% income and thresholds
  - spain_data.xlsx

Importing Excel Data

- Clear your data and open the data editor
- Copy and paste the data there
- **Problem**: Data is imported exactly as displayed!!!
- If our system uses comma to separate decimals, we are into trouble!
- We can change this feature, of course...
- Or we just import the data in a different way!

Importing Excel Data

- Let's use the menu
- If you are using a old version of Stata:
  - Save the data as "CSV (comma delimited values)"'
  - Go the menu: File $\rightarrow$ Import $\rightarrow$ ASCII data created by spreadsheet
  - The name of the (old) command: **insheet**, check the delimiter
- If you are using a new version of Stata:
  - You do not need to save in csv
  - File $\rightarrow$ Import $\rightarrow$ Excel Spreadsheet
- The options are very intuitive, just experiment with them
- Now let's use the do-file
- **Pro-tip**: In practice I use the menu to import the data and then just copy and paste the command in the do-file!

Importing CSV Data

- CSV is a common format since it can store lots of data
- If the data is already in CSV no need to change to import
- Old versions can use the previous command
- New versions: File $\rightarrow$ Import $\rightarrow$ Text Data
- Check the options!
  - Delimiter
  - First row for variable names
  - Text Enconding

Importing CSV Data

- CSV is a common format since it can store lots of data
- If the data is already in CSV no need to change to import
- Old versions can use the previous command
- New versions: File $\rightarrow$ Import $\rightarrow$ Text Data
- Check the options!
  - Delimiter
  - First row for variable names
  - Text Enconding
- Importing delimited **.txt** data works the same way
- Try to import spain_data in both .csv and .txt

Importing Data

- It is not unusual that we have to open excel and do some pre-processing before importing
- Potential problems you may encounter
- Variable names: Stata does not accept variable names starting with numbers (among other rules)
- Solution: Change it before in excel (one trick is to include just a letter before the numbers: e.g. 1998 to y1998)
- Importing string data with latin characters $\rightarrow$ play with the enconding
- Stata imported numeric data in form of string because some missing values
- Solution: use the command **destring** maybe with option **force**
- Data is too big (because of storage type) $\rightarrow$ use **compress**

Reshaping Data

- Sometimes even after we import we want to modify the structure of our data
- Stata has a nice command for it: **reshape**
- Let's say your data has some indicators by country (in the rows) by year (in the columns) $\rightarrow$ Your data is in **wide** format
- It is easier if you have your data in the **long** format: both country and year are in the rows
- reshape long indicator, i(country) j(year) $\rightarrow$ The data will go from wide to long and we will create a new variable "year"
- The reverse operation: reshape long indicator, i(country) j(year), but the variable "j()" should exist already

Saving the Data

- After you have imported the data it is useful to save in **.dta**
- You can use the menu: File $\rightarrow$ Save (as)
- Or just use the command **save**
- **CAREFUL:** data set saved by new versions of Stata does not open in some old versions!
- Use **saveold** instead
- If you want to erase the data set you can use the command **erase**

Exercise 2

1. Go to the World Bank Open Data
2. Search for the data on poverty headcount at 1.9 USD a day and download in the excel format
3. First, try to import in Stata without modifying any of the actual data: what are the problems did you encounter?
4. Reshape the data to long format
5. After your data is ready to use save it in .dta format.