

Introduction to Stata

Lecture III

Tomas R. Martinez

UC3M

September, 2019

“In God we trust. All others must bring data.” - W. E. Deming

- We have seen how to read “structured data” in Excel, csv or txt
- Sometimes the data has a very loose structure → use a **dictionary**
- **What is a dictionary?**
- Set of instructions on how to read the data

Importing Data with dictionary

- We have seen how to read “structured data” in Excel, csv or txt
- Sometimes the data has a very loose structure → use a **dictionary**
- **What is a dictionary?**
- Set of instructions on how to read the data

- Many data sources have their own dictionary
- **Example:** IPUMS (Mexico census data)
 - IPUMS dictionary are usually very complete
 - Not only it open the data, but also label it
 - Data includes: Income, age, sex, education, marital status
- What if the data does not have their own dictionary?

- We have to build it (tedious process)
- What you need to know about your data
 - The variables (of course)
 - The positions of your variable
 - If possible the data type (string, numeric...)

- Let's use the command **infix**
- infix *specifications* using *datafile*
- Specifications are the variables name and the position in the data
- E.g. infix country 1-3 using country.txt
- Look at the IPUMS example
- If the *specifications* we can also write a .dct file and call it using infix

Names and Labels

- To rename a variable use command **rename**
- We can also label the variables (command: **label**)
- **label variable** *varnames* "**variable label**"
- You can also give labels to the values of variables
- This is done in two steps:
 - First define the labels using **label define**
 - Then set the labels using **label values**

Exercise 3

- 1 We are going to build our own dictionary for the PNAD data (Brazilian household survey)
- 2 First take a look at the file *dictionary_PNAD.xls*, check if it has the desired information
- 3 Now, open a do-file and use the command **infix** to import the file *PES2011.txt*
- 4 If the command is too long you can write line breakers “*///*” in the end of each line, or change the delimiter (**help delimit**)
- 5 Now rename and include the label for variables gender and race
- 6 Finally, relabel the values (e.g. instead of 2, the variable gender should have “male”)