

Stata Workshop

At MINAGRI

Roshni Khincha and Sakina Shibuya

DIME, World Bank

August, 2018



Section 1

Why learn stata?

Excel vs Stata

Can I use Excel?

The main reasons to use Stata

- In Excel you make changes directly to the data and save new versions of the data set
- In Stata you make changes to the instructions on how to get from the raw data to the final analysis and save new versions of the instructions
- Since Stata is a more statistics oriented software, processing the data to create analytical products can be a lot easier.

The main reasons to use Stata

- Powerful tool with many capabilities:
 - Descriptive statistics
 - Inference statistics
 - Complex data analysis
- But it's also good for beginner programmers:
 - User friendly interface
 - Relatively easy programming language that can be learned while you're using the software

What's the fuss about do-files?

- Its through the do-file you communicate your work to other members in your team, both current and future
- Think of the do-files as instructions on how to get from raw data to final report
- For a simple task you can enter commands manually. But for more complex tasks you need to write a recipe, or a list of instructions

Stata interface

The Stata interface

The screenshot shows the Stata interface with several callouts highlighting key components:

- Drop down menus**: Points to the menu bar (File, Edit, Data, Graphics, Statistics, User, Window, Help).
- Short cuts**: Points to the toolbar icons.
- Review window**: Points to the Command window showing the command history.
- Result Window**: Points to the main output area displaying the results of the commands.
- Variable window**: Points to the Variables list on the right, showing the list of variables in the dataset.
- Variable and data properties window**: Points to the Properties window on the right, showing the properties of the selected variable (headroom).
- Command window**: Points to the Command window at the bottom, showing the current command being executed.

The Command window shows the following commands:

```
. use "C:\Users\WB462869\Dropbox\FC Training - Kris\Intro Stata - Track 1\practice_data.dta"
> tice_data.dta", clear
(1978 Automobile Data)

. tabulate rep78
```

The Result Window displays the following table:

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
Total	69	100.00	

The Command window shows the following command:

```
. regress price headroom weight
```

The Result Window displays the following table:

Source	SS	df	MS	Number of obs =	F(2, 71) =	Prob > F =	R-squared =	Adj R-squared =	Root MSE =
Model	201873307	2	100936654	16.1		0.00	0.31	0.29	2470
Residual	433192089	71	6101297.03						
Total	635065396	73	8699525.97						

The Command window shows the following command:

```
. regress price headroom weight
```

The Result Window displays the following table:

	price	Coef.	Std. Err.	t	P> t	[95% Conf. Interva
headroom	-663.7758	390.383	-1.70	0.093	-1442.177	114.62
weight	2.393377	.4249418	5.63	0.000	1.546067	3.2406
_cons	925.3939	1282.38	0.72	0.473	-1631.6	3482.3

The Command window shows the following command:

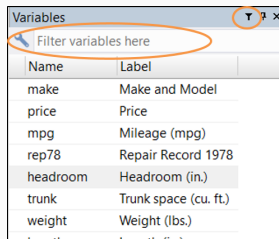
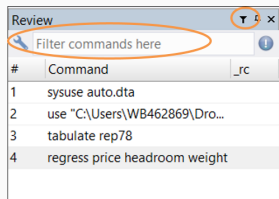
```
. regress price headroom weight
```


The review window

- Provides a history of your actions
- A convenient way to bring back your previous commands and modify it to do something new
- Double click on a command you want to use again and it will appear in your command window
 - You can also click in command window and select the commands in the result window by using *PageUp/PageDown* buttons (or *fn+ArrowUp/ fn+ArrowDown* on Mac)
- If a command is **red** in the review window, it means it did not finish because an error

Filtering in variable and review windows

- Both the variable and the review window will soon be very crowded. You can then search both of them for commands/variables
- If you do not see the search bar, click the little funnel symbol



How to open a data set in Stata

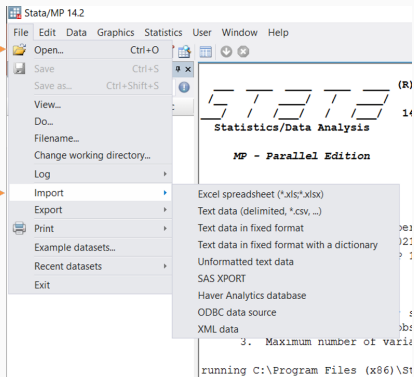
Three ways to tell Stata what to do

- Drop-down menus
 - An easy place to start but quickly becomes inefficient
- Command window
 - Faster than menus but require that you are familiar with the command
- Do-file
 - The only feasible way to run long instructions
 - Use menus and command window to figure out what you need to write, then copy to a do file

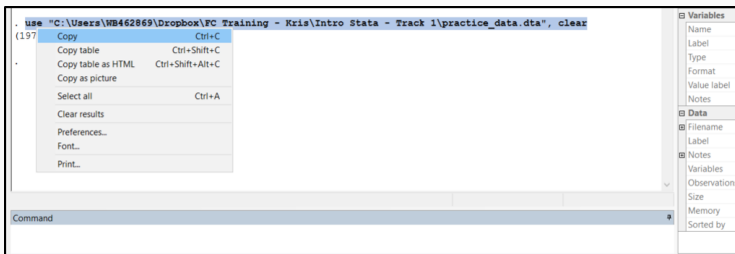
Open a dataset - menus

Open data sets that are already in Stata format here (.dta)

Open all other types of data sets (.csv, .xlsx etc.) here



Open a dataset - command window



- When you use the menus, Stata produces the code for that action (except for Data Browse)
 - Highlight, right-click and copy the code
 - Paste the code in the command window
 - Hit enter

Task 1

1. Open Stata and then open the EICV household data set **cs_s0_s5_household.dta** using the menu: File → Open. Navigate to where you saved the material for this lab. Select the data set and click *Open*
2. Browse to check that you have data: Data → Data Editor → Data Editor Browse
3. Describe to get additional information on the data: Data → Describe data → Describe data in memory or in a file.
 - A new window will open
 - Select In memory and press OK

Task 1

- You can see that one the second command printed information on your screen.
 - The first part is the command used
 - The second part are the results

```
Contains data from C:\Users\WB519128\Dropbox\Work\WB\Mission - Rwanda Feeder Roads\Sta
> ta Training\Data\cs_s0_s5_household.dta
  obs:      14,419
  vars:       76
  size:    8,709,076
                                28 Jun 2016 09:56
```

variable name	storage type	display format	value label	variable label
hhid	double	%10.0g		Household Identifier_in cross section
province	double	%10.0g	province	Province
district	double	%10.0g	district	District (also stratum CS; panel)
ur2012	double	%10.0g	ur2012	Urban/Rural 2012 (4 categories)
ur2_2012	double	%10.0g	ur2_2012	Urban/Rural 2012 (2 categories)
region	double	%10.0g	region	Region
weight	double	%10.0g		Sampling weight_CS
clust	double	%10.0g		Cluster
rwanda	double	%10.0g	rwanda	All Rwanda

Task 1

- You can perform both tasks by typing the in your command prompt.
This will yield the same results
- Type *browse* in the command window and press enter
- Type *describe* and press enter

Exploring a data set opened for the first time

The EICV data

- For our exercises we will explore part of EICV 4 data
- The data is a household survey collected between 2013 and 2014 by NISR
- It is a cross-section of more than 14 thousand Rwandese households both in rural and urban areas
- Close to 2 thousand of these households form a panel have been also interviewed in EICV 3

Types of variables

- In Stata, each variable (column) has to be either:
 - string (text) values are red when browsing
 - numeric (number) values are black or blue when browsing
- Numbers **can** be stored as text, but text **cannot** be stored as number
 - Not possible to do computations on numbers stored as text
- Categorical variables should be stored as numeric variables and have labels

How the data looks

hhid	province	district	ur2012	ur2_2012	region	Weight
100004	Kigali Cit	Nyarugenge	Urban	Urban	Kigali Cit	71.45979
100005	Kigali Cit	Nyarugenge	Urban	Urban	Kigali Cit	71.45979
100006	Kigali Cit	Nyarugenge	Urban	Urban	Kigali Cit	71.45979
103589	Southern P	Gisagara	Peri urban	Rural	Rural Sout	154.7477
103718	Southern P	Gisagara	Rural	Rural	Rural Sout	165.6057
103719	Southern P	Gisagara	Rural	Rural	Rural Sout	165.6057
103720	Southern P	Gisagara	Rural	Rural	Rural Sout	165.6057
105133	Southern P	Nyamagabe	Semi urban	Urban	Other Urba	152.6599
105134	Southern P	Nyamagabe	Semi urban	Urban	Other Urba	152.6599
105135	Southern P	Nyamagabe	Semi urban	Urban	Other Urba	152.6599

How the data actually is

hhid	province	district	ur2012	ur2_2012	region	weight
100004	1	11	1	1	1	71.45979
100005	1	11	1	1	1	71.45979
100006	1	11	1	1	1	71.45979
103589	2	22	3	2	3	154.7477
103718	2	22	4	2	3	165.6057
103719	2	22	4	2	3	165.6057
103720	2	22	4	2	3	165.6057
105133	2	25	2	1	2	152.6599
105134	2	25	2	1	2	152.6599
105135	2	25	2	1	2	152.6599

Useful commands

- browse: see all data in spreadsheet format
- describe: list of all variables in memory
 - Total number of variables & observations (size of matrix)
 - Variable name, type, format, value label name, variable label
- summarize: Basic statistics for numeric variables
 - Obs (Number of observations), Mean, Std. Dev. (Standard deviation), Min (Minimum), Max (Maximum)
- tabulate: frequencies

More commands

- *codebook*: displays the following for each variable
 - Type (more detail than describe)
 - Number of unique values and number of missing values
 - Range and units
 - Examples of values (strings); tabulations (categorical); or mean, sd and percentiles (continuous)
 - Warnings if embedded blanks (may or may not be ok)
- *labelbook*: displays the following for each stored value label
 - Label definitions
 - Which variables labels are applied to
- *list*: lists all variables and observations
 - Can qualify: *list if price < 5000, list in 1/10*
- *summarize*, *detail* : percentiles, variance, skewness, kurtosis

Task 2

1. Open the **cs_s0_s5_household.dta** again. Use the command prompt this time.
2. Explore the dataset
 - browse - see the different colors in the columns
 - describe - check the storage type column
 - summarize - are there any statistics that might not make sense to interpret?
3. Learn more about the variable *s5bq3a*, the household estimated rent amount. What values does it take on? What is minimum, maximum, mean of this variable? How many unique values does it have?

```
. use "$data\cs_s0_s5_household.dta", clear  
. tabulate s5bq3a  
. summarize s5bq3a  
. codebook s5bq3a
```

Task 2

- Learn more about the variable *ur2012*, to learn about the proportion of urban and rural households in Rwanda
 - `. tabulate ur2012`
 - Create now a pie chart: Graphics → Pie chart, select *ur2012* as Category variable and press OK
- Now, create a pie-chart graph for the variable *s5cq7*, the type drinking water source used. This time, use the command prompt!
 - Use the code printed by the previous graph and replace the name of the variable

- Using help - Type *help summarize* to get documentation on the summarize function
- Using search- Type *search regression* to get general documentation on running regressions in Stata
- Google -Search what you want to do. There are many resources online (e.g. Statalist)

Section 2

How can we delete irrelevant variables?

Delete variables

- blah blah

```
. use "$data\cs_s0_s5_household.dta", clear  
. keep hhid province district ur2012 s5cq2 s5cq4 s5cq8 s5cq15 s5c  
> q23 s5bq2 s5cq22 s5cq13 s5cq17
```

Section 3
