

Managing and analyzing PEN data using Stata

Ronnie Babigumira

PEN Workshop, 11-12/January/2008

Outline

An

Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- 1** Introduction
 - Background
 - Advice on Data
- 2** Getting to Work
 - Getting Data into Stata
- 3** Data Management in Stata
- 4** Summarizing and reporting data
 - Descriptive Statistics
 - Regression with Stata
- 5** Getting results out of Stata
- 6** Some References

Disclosure

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

This is not

- A statistics course
- An econometrics course

So what is it about

Alleviating data pains using Stata. Specifically

- 1 Why Stata
- 2 How Stata works
- 3 How to manage your data in Stata.
- 4 How to summarize and present your data using Stata
- 5 How to analyze your data using Stata

We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

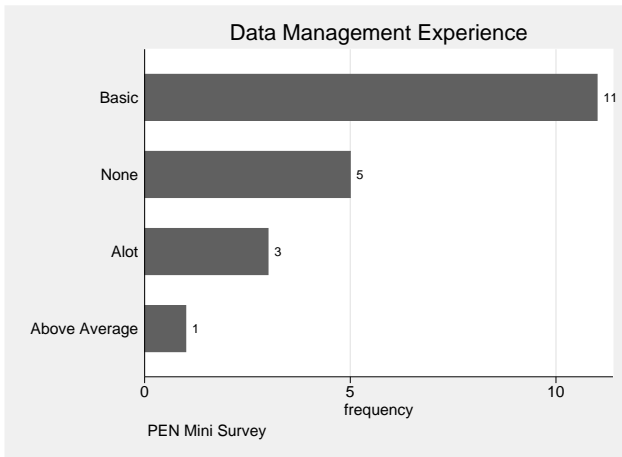
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

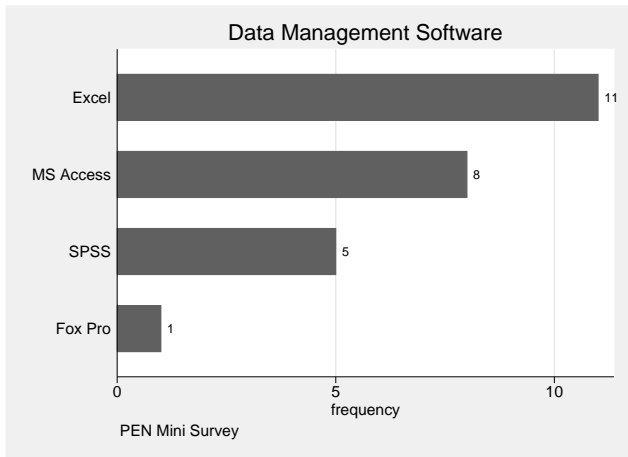
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

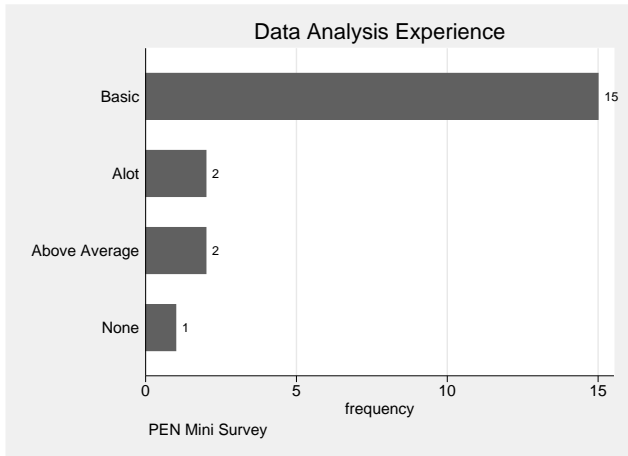
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

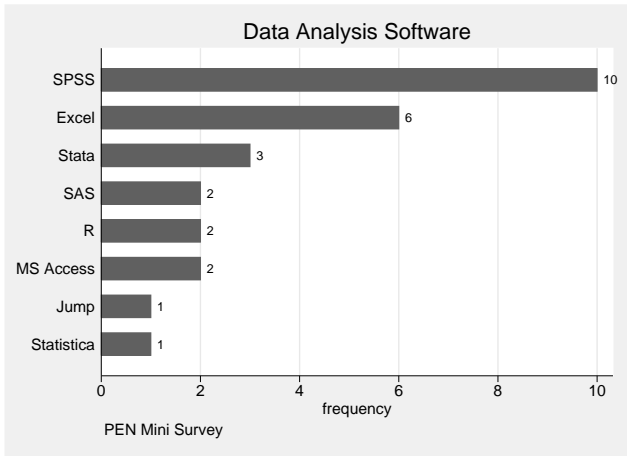
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

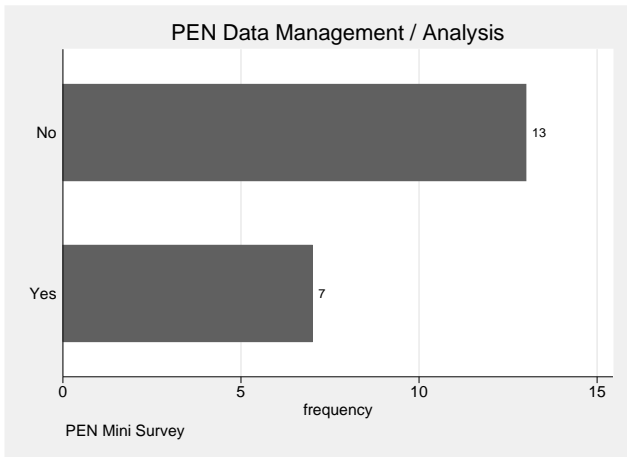
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

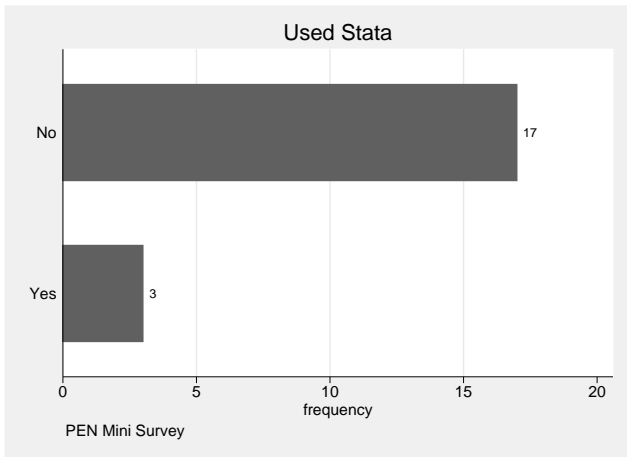
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

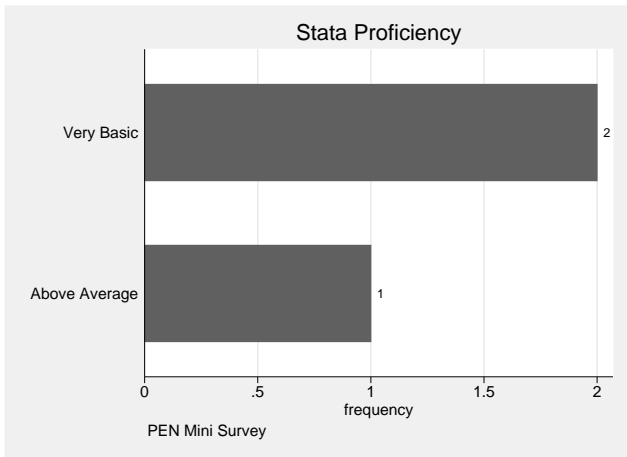
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



We the people: *The PEN mini-survey results (n=20)*

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

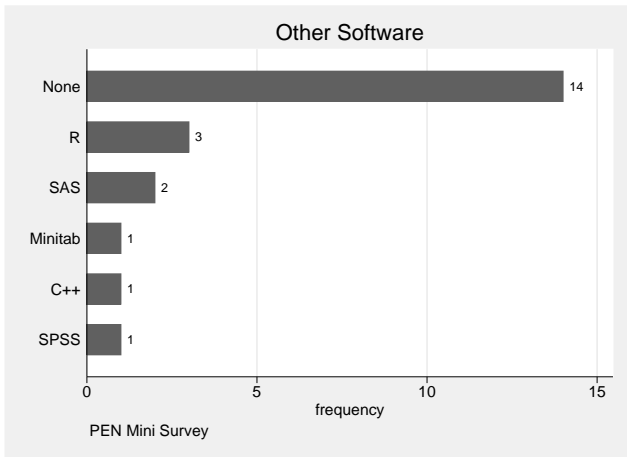
Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary



Who is Stata

An
Introduction
to Stata

Ronnie
Babigumira

Introduction
Background
Advice on Data

Getting to
Work
Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

One of the big three general purpose statistical programs (*see discussion in Mitchell 2006*¹). Increasingly popular in academia and research centers.

- Version 1 Born in 1985. Current running as version 10.
- For those who dared, it is love on first try. However, in the past, a so-so graphing engine and GUI was its Achilles' heel.
- Milestone: version 8 Graphics overhauled, extensive G.U.I. New users would easily click and point themselves through an assignment and there was no longer a need to migrate to another program for graphs.

¹ Mitchell, M,N (2006): Strategically using General Purpose Statistics Packages: A Look at Stata, SAS and SPSS, Statistical Consulting Group, UCLA Academic Technical Services

Why Stata

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

...(with Stata), you can go beyond the menu and go inside the kitchen, not to peel the potatoes or fillet the fish, but to order something not on the menu and discuss it with the chef... ²

Pervasive Principles

- Command language.
- Reproducible research.
- Platform independent.
- Backward compatible.

What sayeth the Mitchell report

- Intuitive data management and manipulation, wide range of Statistics
- Specialized statistical capabilities at no additional cost (e.g. xt: Time series, svy: Survey, st: Survival, Robust S.E's e.t.c.)
- Fully programmable publication quality graphics
- Inexpensive (c.f. SPSS SAS) and widely used

²Cox,N,J (2005) A brief history of Stata on its 20th Anniversary. Stata Journal 5 (2)

Software

- 1 Word processors or text editors
 - 2 Spreadsheets: excel lotus open office calc
 - 3 Statistical programs: SPSS, Stata, SAS
 - 4 Database programs: MS Access
- Most common are spreadsheets (familiarity, ubiquitous) but also have most issues and require discipline
 - Best would be the database packages (less error prone, validation rules, smaller manageable table e.t.c.)
 - Other specialized solutions (most are built on database principles) programs such as Epi Info (<http://www.cdc.gov/EpiInfo/>) EpiData (<http://www.epidata.dk/>), CAQDD

Data Entry in Excel

Rectangular: Basic structure for most statistical software where, columns are variables (questions), rows are observations (households), cells contain the data/values (responses)

Fields/Columns/Qns

Observations/Rows

Values/Data

	A	B	C
1	hhid	age	
2	1001	56	
3	1002	83	
4	1003	26	
5			
6			
7			

Problem excel sheets

An Introduction to Stata

Ronnie Babigumira

Introduction

Background
Advice on Data

Getting to Work

Getting Data into Stata

Data Management in Stata

Summarizing and reporting data

Descriptive Statistics
Regression with Stata

Getting results out of Stata

Summary

Microsoft Excel - Book1.xls

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

F7 = 324

	A	B	C	D	E	F	G	H	I	J
1	This is a header record									
2	This is another unnecessary line									
3	date	company	var1	var2	sex	bad name				
4	1/1/1990	IBM	12	123	m	990				
5	1/1/1991	ATT	43	43	f	22				
6	1/1/1992	Microsoft	556	67	m	n/a				
7	2/3/1995	Citicorp	65	76	f	324				
8	5/23/1998	Chase	788	89	m	546				
9	12/31/1997	Wal-Mart	n/a	99	f	776				
10	9/15/1998	McDonald	456	200	m	879				
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										

Sheet1 Sheet2 Sheet3

Ready NUM

Annotations:

- The first line must be only variable names.
- The first line must be only variable names.
- Variable names must follow Stata rules, e.g., no spaces.
- No good: Missing values must be represented by blank cells.

http://dss.princeton.edu/online_help/stats_packages/stata/excel2stata.htm

Better excel sheet

An Introduction to Stata

Ronnie Babigumira

Introduction

Background

Advice on Data

Getting to Work

Getting Data into Stata

Data Management in Stata

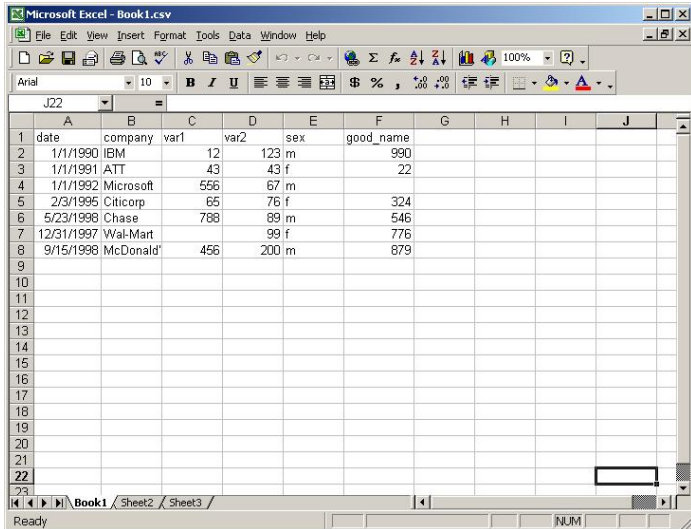
Summarizing and reporting data

Descriptive Statistics

Regression with Stata

Getting results out of Stata

Summary



	A	B	C	D	E	F	G	H	I	J
1	date	company	var1	var2	sex	good_name				
2	1/1/1990	IBM	12	123	m	990				
3	1/1/1991	ATT	43	43	f	22				
4	1/1/1992	Microsoft	556	67	m					
5	2/3/1995	Citicorp	65	76	f	324				
6	5/23/1998	Chase	788	89	m	546				
7	12/31/1997	Wal-Mart		99	f	776				
8	9/15/1998	McDonald	456	200	m	879				
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										

http://dss.princeton.edu/online_help/stats_packages/stata/excel2stata.htm

Preparing excel data for entry in Stata

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- Keep the rectangular format
- First line should have Stata variable names (32 characters or less, no spaces, or "special characters" except the underscore [`_`], not starting with an underscore or number). Second line begins the data.
- No blank rows or columns between data
- Missing numeric data should be coded as an empty cell, not a space, dot, or any other non-numeric data. Often, 0, 9, or 99 is used to code missing numeric data; this is fine as long as these are not also valid values for that variable.
- No commas in numbers. Stata thinks they are a delimiter and will not read the data properly. Remove commas from numeric values before saving the file
- Note:
 - Computers work with numbers, the more, the merrier
 - Do not mix tasks: data entry is data entry.

Caution

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

No serious data analysis and management in spreadsheets.
Lots of online material on pitfalls. e.g. Spreadsheet
Addiction, by Patrick Burns (<http://www.burns-stat.com/>).

The GUI

Start up Stata: Start > Programs > Stata 10 > Intercooled Stata

The screenshot shows the Stata 10.0 GUI with the following components:

- Command Window:** Shows a list of commands being executed. A red box labeled "Previous Commands" highlights lines 11 through 30. Line 26, `sysuse auto`, is highlighted in blue, with a red box labeled "Results" to its right.
- Variables Window:** Displays the list of variables in the current dataset. A red box labeled "Variables in Memory" highlights this window.
- Command Window (Bottom):** A red box labeled "Command Window" highlights the bottom-most command window.
- Output:** The main window displays the results of the `sysuse auto` command, including variable information and dataset details.

variable name	storage type	display format	value label	variable label
make	str18	%-18s		Make and Model
price	int	%8.0gc		Price
mpg	int	%8.0g		Mileage (mpg)
rep78	int	%8.0g		Repair Record 1978
headroom	float	%6.1f		Headroom (in.)
trunk	int	%8.0g		Trunk space (cu. ft.)
weight	int	%8.0gc		Weight (lbs.)
length	int	%8.0g		Length (in.)
turn	int	%8.0g		Turn Circle (ft.)
displacement	int	%8.0g		Displacement (cu. in.)
gear_ratio	float	%6.2f		Gear Ratio
foreign	byte	%8.0g	origin	Car type

Sorted by: foreign

. describe

Contains data from C:\Program Files\Stata10\ado\base\aa\auto.dta
obs: 74 1978 Automobile Data
vars: 12 13 Apr 2007 17:45
size: 3,478 (99.9% of memory free) (_dta has notes)

variable name	storage type	display format	value label	variable label
make	str18	%-18s		Make and Model
price	int	%8.0gc		Price
mpg	int	%8.0g		Mileage (mpg)
rep78	int	%8.0g		Repair Record 1978
headroom	float	%6.1f		Headroom (in.)
trunk	int	%8.0g		Trunk space (cu. ft.)
weight	int	%8.0gc		Weight (lbs.)
length	int	%8.0g		Length (in.)
turn	int	%8.0g		Turn Circle (ft.)
displacement	int	%8.0g		Displacement (cu. in.)
gear_ratio	float	%6.2f		Gear Ratio
foreign	byte	%8.0g	origin	Car type

Sorted by: foreign

Command

Exploring the GUI

Checking out the "1978 Automobile Data", a data set that ships with Stata. **Make a mental note of the contents of the 4 windows here**

The screenshot shows the Stata 10.0 interface with the following windows and content:

- Command Window:** Shows a list of commands from line 11 to 30. Line 18 is highlighted with a red box labeled "Previous Commands".
- Results Window:** Displays the output of the `describe` command, including variable names, storage types, display formats, values, and labels. A red box labeled "Results" is overlaid on the right side.
- Variables in Memory Window:** Shows a table of variables currently loaded in memory, with a red box labeled "Variables in Memory" overlaid on the left side.
- Command Window (bottom):** Shows the command `. describe` entered.

At the bottom of the Stata window, the path `d:\vfor\barcelona\stata_course` is visible.

Exploring the GUI

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

- Load the auto data into Stata's memory (*Hint: Look for **example datasets** on the menu*)
- Launch the data browser (Data Menu).
- Describe the data (*Hint: Look under the **Data** menu*), what can you say about the data
- Produce a one-way table of frequency for the variable foreign
- Produce some summary statistics for the variable price
- Make one table of Summary Statistics (mean, min, range, max, skewness) for mpg, price, weight (*Hint: Look for **tabstat** under the table menu*)
- Make a scatter plot of price against weight. Try and customize the graph by adding a title
- Regress price on weight, length, mpg, and foreign. What is the effect of the source of a car on its price?
- Did you notice any changes in the command review window

Exploration repeated with commands

An
Introduction
to Stata

Ronnie
Babigumira

Introduction
Background
Advice on Data

Getting to
Work
Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

- Close and restart Stata.
- Try to repeat the previous exercise using commands

Solution

```
clear
sysuse auto
browse
describe
tabulate foreign
summarize price
tabstat price mpg weight, statistics(mean min range max skewness)
tway scatter price weight, title(Plot of Price vs Weight)
regress price length weight mpg foreign
```

To Click or not To Click . . . Dont

An
Introduction
to Stata

Ronnie
Babigumira

Introduction
Background
Advice on Data

Getting to
Work
Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

The real power in using Stata lies in typing in commands

- **Reproducibility:**

Give us the data and the code and we can reproduce your results

- **Traceability:**

Errors can easily be traced

- **Comprehension:**

If you can write it, you probably understand it

- **Documentation:**

Add notelets and snippets documenting your moments of genius

- **Laziness:**

Kick it

The Grammar of Stata

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

```
[prefix: ] command [varlist] [= exp] [if] [in] [weight] [using filename ],  
[options]
```

command: Bare minimum. For example, `.describe`

[Prefix]: Some commands precede a Stata command and modify its behaviour

[Varlist]: List of one or many variables, you can use wildcards

[= exp]: Used in commands where algebraic expressions produce a new or update variable

[if] & [in]: Conditions and ranges

[using]: Some commands access data so you or write to files

[, options]: **All Stata commands have help files. help command name should show you how to use the command as well as what options a command has**

- If you don't know the name of the command, use Stata's search command.
`search keyword`

Order in the house

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

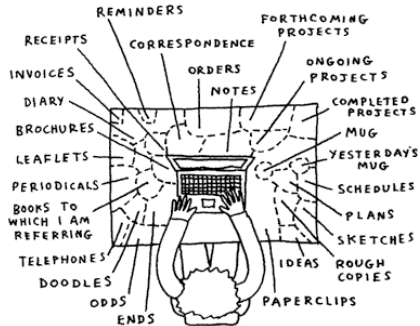
Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

HOW I ORGANISE MY DESK

THE AREAS ARE ALLOTTED THUS:



CartoonChurch.com

Goodbye to “My Documents”

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

"My Documents" is actually

"C:\Documents and Settings\All Users\Me\My Documents"

- Create directories (folders) for your work.
- It will save you using full paths for loading and saving files.
- You will always know where your stuff is (use informative names).
- Make it even better by adding subfolders for data, logs, and reports.

Going Places

- Where are you `.pwd`
- Go to another directory `.cd`
- Create a new folder `.mkdir`
- What stuff do you have in your directory `.dir`

A few variables and observations

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- For a few variables and observations such as

hhid	pid	sex	age
1	1	1	50
1	2	2	30
1	3	2	10
2	1	1	56
2	2	2	70
3	1	1	35

- You can use `.edit` to enter data directly into Stata's "spread sheet"
- Or `.input` to enter the data from the command interface
- Try each of the methods

Reading Spreadsheet Data

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive

Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Method 1: Cut and Paste (Quick and Dirty)

- Open the excel sheet `excel_entry_demo.xls`
- If Stata is running, clear its memory
- Open Stata's data editor and paste (`.edit`)

Method 2: Smart and Safe

- Open spreadsheet in excel, save it as as a Text Tab delimited file (will have extension `.txt`)
- Fire up Stata and use `insheet` to load data. See `.help insheet` on how to do it.

Other methods of reading data

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

■ `.infile`

■ `.infix`

■ `.odbc`

How Stata organizes data

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

		Variables			
		Make	Price	MPG	Weight
O B S		AMC Pacer	4,749	17	3,350
		AMC Spirit	3,799	22	2,640
		Buick Century	4,816	20	3,250
		Buick Electra	7,827	15	4,080
		Buick LeSabre	5,788	18	3,670
		Buick Opel	4,453	26	2,230
		Buick Regal	5,189	20	3,280

Columns are variables and
Rows are observations

Numeric

- May contain sign, an integer part, a decimal point, a fraction part or even exponential
- May not contain commas (an excel artifact you need to address before hand)
- Missing values generally represented as `.` even though Stata has 27 missing numeric values. See `.help missing`

Strings

- A sequence of printable characters
- Often contain identifying information
- Cannot be used directly in statistical analysis though data can be sorted and merged on a string variable

■ Variable labels

- Gives a more detailed description of a variable
- `.label variable varname "My variable label"`

■ Value labels

- Categorical variables have values. E.g, variable sex is entered as 1, 0 (values) which are "Female" "Male" (value labels). To assign value labels
- *Step 1* Define the label
`.label define sexlbl 1 "Female" 0 "Male"`
- *Step 2* Associate the values with the label `.label values sex sexlbl`

■ Data Labels

- Tells you what the data is about
- `.label data "An Informative Data Label"`

■ Notes:

- You can add notes such as version of data, publication you were preparing it for etc
- `.note: My note`
- Add as many notes as you want. To view, simply type `.notes`

Documenting Data Example

An
Introduction
to Stata

Ronnie
Babigumira

Introduction
Background
Advice on Data

Getting to
Work
Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data
Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

From your code book you know that

```
houscode      Household code
hhc_pid       Household member pid
hhc_rela      Relationship to head
hhc_yrborn    Year Member was born
hhc_sex       Sex
hhc_edu       Member years of schooling
```

```
hhc_rela
0          Head
1          Spouse
2          Son/daughter
4          Son/daughter in law
11         Step/foster child
```

```
hhc_sex
0          Male
1          Female
```

This is version 1 of the data
And the data were collected in a mystery country.

Exercise: Document the data

Data Cleaning

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Having documented the data, do some basic data cleaning

- 1 Identify missing cases.
- 2 Are there any observations whose "relationship to head" codes are outside the range?
- 3 Are there any household heads who are younger than their children
- 4 We know that parents should be at least 12 years old. Are there any parents younger than this age?
- 5 When you are done, save the data. See help save on how to do it

Exploring data. Some useful commands

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- `.browse`: Look but do not touch
- `.edit`: Look and touch
- `.list`: My most used Stata command
- `.inspect`: Provides a quick summary of a numeric variable
- `.codebook`: to get a codebook describing the data
- `.describe`: Comes a close second to list. Very useful

Manipulating Variables

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Creating new Variables

- `.generate` creates a new variable
- `.encode` changes a string variable to a numeric variable
- `.decode` changes a numeric variable to a string variable
- `.egen` a powerful extension to generate.

Example

- Generate the age of household members
- Generate a string variable from the variable sex.
- Generate the square of age
- Generate two variables nmales and nmales which is a count of males and females per household
- Generate a dummy variable “children” which is 1 if the member is a child and 0 otherwise

Manipulating Variables

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Changing existing variables

- `.replace`: changes the contents of an existing variable
- `.recode` changes the values of numeric variables. Values that do not meet any of the conditions of the rules are left unchanged, unless an otherwise rule is specified.
- `.rename` changes the name of an existing variable; the contents of the variable are unchanged.

Example

- We just learnt that member sex codes 3 and 4 should be 1. Change them
- Change the name of `hhc_sex` to `mem_sex`

Manipulating files

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

`.append`

- Appends Data (adds observations / rows)
- You have different people entering data `.append`
- You have a time series 1990-2000, a new time series 2000-2006 has been released

`.merge`

- Merges data (adds new variables / columns)
- multilevel data (plot and household level, you want to add plot data to household data)

Append illustration

An

Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Round 1			
Year	Price	MPG	Weight
1990	4,749	17	3,350
1991	3,799	22	2,640
1992	4,816	20	3,250
1993	7,827	15	4,080

Round 2			
Year	Price	MPG	Weight
1994	5,788	18	3,670
1995	4,453	26	2,230
1996	5,189	20	3,280

Append the Round 2 to Round 1 to get

Final DataSet			
Year	Price	MPG	Weight
1990	4,749	17	3,350
1991	3,799	22	2,640
1992	4,816	20	3,250
1993	7,827	15	4,080
1994	5,788	18	3,670
1995	4,453	26	2,230
1996	5,189	20	3,280

Merge Illustration

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

One to Many (Spread, most common)

hhid	pid	los
1	1	4
1	2	6
1	3	5
2	1	5
2	2	8

hhid	beds
1	200
2	750

hhid	pid	los	beds
1	1	4	200
1	2	6	200
1	3	5	200
2	1	5	750
2	2	8	750

Merge and Append exercise

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- Append the two files
excel_entry_demo.xls
append_demo.xls
- Merge the two files
hhda1_a_id.xls
hhda1_b1_hhc.xls

Changing Data Structure

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

```
.collapse
```

- Aggregate data (make a dataset of summary statistics). Examples include
- Data on multiple crops per household and you just want one value per household
- Household aggregates (e.g, number of males, number of females, average years of schooling for the household)

```
.reshape
```

- changes data from wide to long shape and vice versa
- reshape is a very powerful command that many new users find rather perplexing, however, if you harness its power, many have sworn it can be life changing

Reshape illustration

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

Wide Format			
hhid	inc80	inc81	inc82
1001	700	900	1500
1002	210	500	100
1003	215	600	420

Long Format		
hhid	year	inc
1001	80	700
1001	81	900
1001	82	1500
1002	80	210
1002	81	500
1002	82	100
1003	80	215
1003	81	600
1003	82	420

■ Variables with a few categories

- `.inspect` How many different numbers occur and how are they distributed (`.inspect rep78`). How many categories are there and do you have any missing values?
- `.tabulate` Get a one way frequency table. This is where having labeled the categories would be helpful.
e.g. `.tabulate rep78`
- Using `.tabulate` with more than two variables will give you a two-way frequency table

■ Comparing distributions

- You can compare a distribution for different groups by using the prefix by
- `.by foreign: tabulate rep78`

Describing distributions . . . Graphs

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- The most commonly used graph for eye balling data is the Histogram. You can plot a histogram for both continuous and discrete data (add the option `discrete`)
- Examples `.histogram length` and `.histogram rep78, discrete`
- Look at the Graphics Manual for inspiration

More on Graphs

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to

Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- **Basic syntax:** `.graph type varlist, options`
- **Scatter plot:** `.twoway scatter mpg weight`
- **Overlay a line:** `twoway scatter mpg weight || lfit mpg weight`
- **Quadratic fit better** `twoway scatter mpg weight || qfit mpg weight`
- **For domestic cars:** `.twoway scatter price weight if foreign == 0`
- **By grouping variable** `.scatter price weight, by(foreign)`
- **Add a total plot** `.scatter price weight, by(foreign, total)`
- **Better?** `.scatter price weight, by(foreign, total row(1))`
- **Or?** `.scatter price weight, by(foreign, total col(1))`

Describing data using Statistics

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- Most often used are the mean and the standard deviation. Some commands
- `.summarize price`
- `.summarize price, detail`
- You can also use `.tabstat`, a generalization of `summarize` that allows you to choose additional statistics. see `.help tabstat`
- There are other commands, explore the GUI.

An example of Descriptive Statistics

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- **Qn.** What was the average price of a cars in the autodata
Ans. `. summarize`. Add detail and note the difference
- **Qn.** Was the average price of a foreign car significantly different from the price of a domestic car
Ans. `. ttest price, by(foreign)`

... Example continued

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- **Qn.** What was the percentage of foreign cars
Ans. `.tabulate foreign`
- **Qn:** What was the frequency of repair records for domestic and foreign cars (1= poor ... 5 = Excellent)
Ans. `.tabulate rep78 foreign` (you can get row or column percentages)
- **Qn.** What is the correlation mpg and weight?
Ans. `correlate mpg weight`

Basic OLS

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

Before

- Examine the data: Look for missing cases, strange values etc
- use `.codebook` and or `.summarize`
- you can also use graphical methods e.g histogram, box plots e.t.c
- examine relationships between variables `.summarize` or `.graph matrix price weight mpg length`

The Regression

- `.regress price weight mpg length`
- You can use options `.summarize price weight mpg length, robust`

After

- Carry out regression diagnostics `.help regress postestimation.`
- test hypotheses
- Export your regression results

Other types of regression analysis

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

■ Limited dependent variable models

- Binary: `.help logit`, `.help probit`

- Ordered categories `.help oprobit` or `.help ologit`

- Multiple unordered categories `.help mprobit` or
`.help mlogit`

- Truncated: `.help tobit`

■ Panel data `.help xt`

■ Survey data `.help survey`

Getting regression output out of Stata

An
Introduction
to Stata

Ronnie
Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics
Regression with
Stata

Getting
results out
of Stata

Summary

Stata's `.estimates` store and `.estimates` table are simple, straight forward. However, there are also user written programs

- `.outreg` : very popular but has not been supported until recently
- `.outreg2`: Birthed as a result of the lack of support for `outreg`
- `.estout`: Superb but some have found it a little complex. Start with its baby brother `.eststo`

Stata's estimates commands

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

```
. regress price mpg weight  
<--- output omitted ----->
```

```
* Store the estimates of the First model  
. estimates store model1, title(Model 1)
```

```
. regress price mpg weight foreign  
<--- output omitted ----->
```

```
* Store estimates for Second model  
. estimates store model2, title(Model 2)
```

```
* Print table of estimates
```

```
* Basic table
```

```
estimates table model1 model2
```

```
<--- output omitted ----->
```

```
* Better
```

```
. estimates table model1 model2, star label stats(r2) stfmt(%9.2f) b(%9.3f)
```

```
-----  
                Variable |      model1      model2  
-----+-----  
Mileage (mpg) |    -49.512      21.854  
Weight (lbs.) |     1.747**      3.465***  
Car type |                   3673.060***  
Constant |    1946.069     -5853.696  
-----+-----  
                r2 |         0.29         0.50  
-----
```

```
legend: * p<0.05; ** p<0.01; *** p<0.001
```

Exporting graphs

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background
Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- Basic: Copy and paste to MS word
- Better: use `.graph export` (wmf or emf, but can use other formats)

Solution

```
. .scatter mpg weight || lfit mpg weight  
. .graph export scatter_emf.emf  
. .graph export scatter_wmf.wmf
```

Look at `Graphs_in_word.doc`

Summary

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to
Work

Getting Data into
Stata

Data Man-
agement in
Stata

Summarizing
and
reporting
data

Descriptive
Statistics

Regression with
Stata

Getting
results out
of Stata

Summary

- **Stata is accessible:** It is a worthwhile investment.
- **There is so much to learn:** checkout the G.U.I
- **Walk before you fly:** Before dashing off to estimate fancy models invest sometime in knowing your data.
- **HELP:** Use help files to learn how the syntax works
 - Look at the menu, save what works
 - Use search, findit
 - Look at Stata's FAQ's
(<http://www.stata.com/support/faqs/>)
 - If none of the above helps, contact stata technical support

A few books to get you started

An

Introduction
to Stata

Ronnie

Babigumira

Introduction

Background

Advice on Data

Getting to

Work

Getting Data into

Stata

Data Man-

agement in

Stata

Summarizing

and

reporting

data

Descriptive

Statistics

Regression with

Stata

Getting

results out

of Stata

Summary

- Lots of good material written on Stata.
- Visit <http://stata.com/bookstore/statabooks.html>.
- Some of my favorites (these lecture notes draw from and are inspired by these books)
 - Juul, S., (2006) *An Introduction to Stata for Health Researchers*
 - Kohler, U., and Kreuter, F., (2005) *Data Analysis Using Stata*, Stata Press
 - Baum, C., (2006) *An Introduction to Modern Econometrics Using Stata*, Stata Press
 - Hamilton, L. C., (2006) *Statistics with Stata (Updated by version 9)*, Stata Press
 - Long, S.J., and Fresse, J., (2006) *Regression Models for Categorical Dependent Variables Using Stata*, Stata Press