Managing and analyzing PEN data using Stata

Ronnie Babigumira

PEN Workshop, 11-12/January/2008
Outline

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
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)

Data Management Experience

- Basic: 11
- None: 5
- Alot: 3
- Above Average: 1

PEN Mini Survey

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

Data Management Software

- Excel: 11
- MS Access: 8
- SPSS: 5
- Fox Pro: 1

PEN Mini Survey

Frequency

Ronnie Babigumira
An Introduction to Stata
Data Analysis Experience

- Basic: 15
- Alot: 2
- Above Average: 2
- None: 1

PEN Mini Survey

None
Above Average
Alot
Basic

frequency

None
Above Average
Alot
Basic

5
10
15
0
15
frequency

PEN Mini Survey

None
Above Average
Alot
Basic

We the people: The PEN mini-survey results (n=20)
An Introduction to Stata

Ronnie Babigumira

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

Data Analysis Software

<table>
<thead>
<tr>
<th>Software</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSS</td>
<td>10</td>
</tr>
<tr>
<td>Excel</td>
<td>6</td>
</tr>
<tr>
<td>Stata</td>
<td>3</td>
</tr>
<tr>
<td>SAS</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>MS Access</td>
<td>2</td>
</tr>
<tr>
<td>Jump</td>
<td>1</td>
</tr>
<tr>
<td>Statistica</td>
<td>1</td>
</tr>
</tbody>
</table>

PEN Mini Survey

Ronnie Babigumira

An Introduction to Stata
### We the people: The PEN mini-survey results (n=20)

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>13</td>
</tr>
<tr>
<td>Yes</td>
<td>7</td>
</tr>
</tbody>
</table>
We the people: The PEN mini-survey results ($n=20$)

Used Stata

PEN Mini Survey

Yes: 3

No: 17

frequency

Ronnie Babigumira
An Introduction to Stata
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

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

Stata Proficiency

PEN Mini Survey

Frequency

Above Average

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

Other Software

- **None**: 14
- **R**: 3
- **SAS**: 2
- **Minitab**: 1
- **C++**: 1
- **SPSS**: 1

PEN Mini Survey

<table>
<thead>
<tr>
<th>Software</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>14</td>
</tr>
<tr>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>SAS</td>
<td>2</td>
</tr>
<tr>
<td>Minitab</td>
<td>1</td>
</tr>
<tr>
<td>C++</td>
<td>1</td>
</tr>
<tr>
<td>SPSS</td>
<td>1</td>
</tr>
</tbody>
</table>

Ronnie Babigumira
Who is Stata

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

- 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.

---

**Why Stata**

...(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...  

<table>
<thead>
<tr>
<th>Pervasive Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Command language.</td>
</tr>
<tr>
<td>Reproducible research.</td>
</tr>
<tr>
<td>Platform independent.</td>
</tr>
<tr>
<td>Backward compatible.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What sayeth the Mitchell report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuitive data management and manipulation, wide range of Statistics</td>
</tr>
<tr>
<td>Specialized statistical capabilities at no additional cost (e.g. xt: Time series, svy: Survey, st: Survival, Robust S.E’s e.t.c.)</td>
</tr>
<tr>
<td>Fully programmable publication quality graphics</td>
</tr>
<tr>
<td>Inexpensive (c.f. SPSS SAS) and widely used</td>
</tr>
</tbody>
</table>

---

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

### 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/](http://www.cdc.gov/EpiInfo/)
  EpiData [http://www.epidata.dk/](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)

<table>
<thead>
<tr>
<th>hhid</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>56</td>
</tr>
<tr>
<td>1002</td>
<td>83</td>
</tr>
<tr>
<td>1003</td>
<td>26</td>
</tr>
</tbody>
</table>
Problem excel sheets

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>This is a header record</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>This is another unnecessary line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>data</td>
<td>company</td>
<td>var1</td>
<td>var2</td>
<td>sex</td>
<td>bad name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1/1/1990</td>
<td>IBM</td>
<td>12</td>
<td>123</td>
<td>m</td>
<td>990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1/1/1991</td>
<td>ATT</td>
<td>43</td>
<td>43</td>
<td>f</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1/1/1992</td>
<td>Microsoft</td>
<td>556</td>
<td>67</td>
<td>m</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2/3/1995</td>
<td>Citicorp</td>
<td>65</td>
<td>76</td>
<td>f</td>
<td>324</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5/23/1998</td>
<td>Chase</td>
<td>788</td>
<td>88</td>
<td>m</td>
<td>546</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12/31/1997</td>
<td>WalMart</td>
<td>n/a</td>
<td>96</td>
<td>f</td>
<td>776</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9/15/1998</td>
<td>McDonald</td>
<td>456</td>
<td>290</td>
<td>m</td>
<td>879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

http://dss.princeton.edu/online_help/stats_packages/stata/excel2stata.htm
Preparing excel data for entry in Stata

- 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

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/).
Start up Stata: Start > Programs > Stata 10 > Intercooled Stata

The GUI

Previous Commands

Variables in Memory

Command Window

Ronnie Babigumira
An Introduction to Stata
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.
Exploring the GUI

- 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

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

Solution

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

\[ \text{prefix: } \text{command [varlist][=exp][if][in][weight][using filename ], [options] } \]

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

**Prefix [command before the main command]:** Some commands precede a Stata command and modify its behaviour

**Varlist [a list of one or more variables]:** 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 Management in Stata

Summarizing and reporting data
Descriptive Statistics
Regression with Stata

Getting results out of Stata

Summary

Ronnie Babigumira
An Introduction to Stata
"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

- For a few variables and observations such as

<table>
<thead>
<tr>
<th>hhid</th>
<th>pid</th>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>

- 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

**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

- `.infile`
- `.infix`
- `.odbc`
How Stata organizes data

<table>
<thead>
<tr>
<th>Variables</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Make</td>
<td>Price</td>
<td>MPG</td>
<td>Weight</td>
</tr>
<tr>
<td>AMC Pacer</td>
<td>4,749</td>
<td>17</td>
<td>3,350</td>
</tr>
<tr>
<td>AMC Spirit</td>
<td>3,799</td>
<td>22</td>
<td>2,640</td>
</tr>
<tr>
<td>Buick Century</td>
<td>4,816</td>
<td>20</td>
<td>3,250</td>
</tr>
<tr>
<td>Buick Electra</td>
<td>7,827</td>
<td>15</td>
<td>4,080</td>
</tr>
<tr>
<td>Buick LeSabre</td>
<td>5,788</td>
<td>18</td>
<td>3,670</td>
</tr>
<tr>
<td>Buick Opel</td>
<td>4,453</td>
<td>26</td>
<td>2,230</td>
</tr>
<tr>
<td>Buick Regal</td>
<td>5,189</td>
<td>20</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Columns are variables and
Rows are observations
Data Types

**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

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
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

- `.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

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 `nfemales` 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

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
### .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

### Round 1

<table>
<thead>
<tr>
<th>Year</th>
<th>Price</th>
<th>MPG</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>4,749</td>
<td>17</td>
<td>3,350</td>
</tr>
<tr>
<td>1991</td>
<td>3,799</td>
<td>22</td>
<td>2,640</td>
</tr>
<tr>
<td>1992</td>
<td>4,816</td>
<td>20</td>
<td>3,250</td>
</tr>
<tr>
<td>1993</td>
<td>7,827</td>
<td>15</td>
<td>4,080</td>
</tr>
</tbody>
</table>

### Round 2

<table>
<thead>
<tr>
<th>Year</th>
<th>Price</th>
<th>MPG</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>5,788</td>
<td>18</td>
<td>3,670</td>
</tr>
<tr>
<td>1995</td>
<td>4,453</td>
<td>26</td>
<td>2,230</td>
</tr>
<tr>
<td>1996</td>
<td>5,189</td>
<td>20</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Append the Round 2 to Round 1 to get

### Final Data Set

<table>
<thead>
<tr>
<th>Year</th>
<th>Price</th>
<th>MPG</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>4,749</td>
<td>17</td>
<td>3,350</td>
</tr>
<tr>
<td>1991</td>
<td>3,799</td>
<td>22</td>
<td>2,640</td>
</tr>
<tr>
<td>1992</td>
<td>4,816</td>
<td>20</td>
<td>3,250</td>
</tr>
<tr>
<td>1993</td>
<td>7,827</td>
<td>15</td>
<td>4,080</td>
</tr>
<tr>
<td>1994</td>
<td>5,788</td>
<td>18</td>
<td>3,670</td>
</tr>
<tr>
<td>1995</td>
<td>4,453</td>
<td>26</td>
<td>2,230</td>
</tr>
<tr>
<td>1996</td>
<td>5,189</td>
<td>20</td>
<td>3,280</td>
</tr>
</tbody>
</table>
Merge Illustration

One to Many (Spread, most common)

<table>
<thead>
<tr>
<th>hhid</th>
<th>pid</th>
<th>los</th>
<th>beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hhid</th>
<th>pid</th>
<th>los</th>
<th>beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>200</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>750</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
<td>750</td>
</tr>
</tbody>
</table>
Merge and Append exercise

- 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

- **.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

<table>
<thead>
<tr>
<th>Wide Format</th>
<th>Long Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>hhid</td>
<td>inc80</td>
</tr>
<tr>
<td>1001</td>
<td>700</td>
</tr>
<tr>
<td>1002</td>
<td>210</td>
</tr>
<tr>
<td>1003</td>
<td>215</td>
</tr>
<tr>
<td>1002</td>
<td>80</td>
</tr>
<tr>
<td>1002</td>
<td>81</td>
</tr>
<tr>
<td>1002</td>
<td>82</td>
</tr>
<tr>
<td>1003</td>
<td>80</td>
</tr>
<tr>
<td>1003</td>
<td>81</td>
</tr>
<tr>
<td>1003</td>
<td>82</td>
</tr>
</tbody>
</table>
Describing data

- 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`
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

- 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))`
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

- **Qn.** What was the average price of a car 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

- **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

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

- **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`
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

```
. 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)

<table>
<thead>
<tr>
<th>Variable</th>
<th>model1</th>
<th>model2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mileage (mpg)</td>
<td>-49.512</td>
<td>21.854</td>
</tr>
<tr>
<td>Weight (lbs.)</td>
<td>1.747**</td>
<td>3.465***</td>
</tr>
<tr>
<td>Car type</td>
<td>3673.060 ***</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1946.069</td>
<td>-5853.696</td>
</tr>
</tbody>
</table>

r2    | 0.29     | 0.50

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

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

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

Look at Graphs_in_word.doc
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

- Lots of good material written on Stata.
- Some of my favorites (these lecture notes draw from and are inspired by these books)