Some Stata Commands

Last modified: January 2, 2006 9:51AM

General Plotting Commands

1. Plot a histogram of a variable:
   `histogram vname`
2. Plot a histogram of a variable using frequencies:
   `histogram vname, freq`
   `histogram vname, bin(xx) norm`
   where `xx` is the number of bins.
3. Plot a boxplot of a variable:
   `graph box vname`
4. Plot side-by-side box plots for one variable (vone) by categories of another variable
   vtwo. (vtwo should be categorical):
   `graph box vone, over(vtwo)`
5. A scatter plot of two variables:
   `scatter vone vtwo`
6. A matrix of scatter plots for three variables:
   `graph matrix vone vtwo vthree`
7. A scatter plot of two variables with the values of a third variable used in place of points
   on the graph (vthree might contain numerical values or indicate categories, such as
   male ("m") and female ("f")):
   `scatter vone vtwo, symbol([vthree])`
8. Normal quantile plot:
   `qnorm vname`

General commands

1. To compute means and standard deviations of all variables:
   `summarize`
   or, using an abbreviation,
   `summ`
2. To compute means and standard deviations of select variables:
   `summarize vone vtwo vthree`
3. Another way to compute means and standard deviations that allows the `by` option:
   `tabstat vone vtwo, statistics(mean, sd) by(vthree)`
4. To get more numerical summaries for one variable:
   `summ vone, detail`
5. See help tabstat to see the numerical summaries available. For example:
   `tabstat vone, statistics(min, q, max, iqr, mean, sd)`

6. Correlation between two variables:
   `correlate vone vtwo`

7. To see all values (all variables and all observations, not recommended for large data sets):
   `list`
   Hit the space bar to see the next page after "-more-" or type "q" to "break"

   (stop/interrupt the listing).

8. To list the first 10 values for two variables:
   `list vone vtwo in 1/10`

9. To list the last 10 values for two variables:
   `list vone vtwo in -10/l`
   (The end of this command is "minus 10" / "lowercase letter L").

10. Tabulate categorical variable `vname`:
    `tabulate vname`
    or, using an abbreviation,
    `tab vname`

11. Cross tabulate two categorical variables:
    `tab vone vtwo`

12. Cross tabulate two variables, include one or more of the options to produce column, row or cell percents and to suppress printing of frequencies:
    `tab vone vtwo, column row cell`
    `tab vone vtwo, column row cell nofreq`

**Generating new variables**

1. General.
   a. Generate index of cases 1,2, ...,n (this may be useful if you sort the data, then want to restore the data to the original form without reloading the data):
      `generate case= _n`
      or, using an abbreviation,
      `gen case=_n`
   b. Multiply values in `vx` by `b` and add `a`, store results in `vy`:
      `gen vy = a + b * vx`
   c. Generate a variable with values 0 unless `vtwo` is greater than `c`, then make the value 1:
      `gen vone=0`
      `replace vone=1 if vtwo>c`
   d.

2. Random numbers.
a. Set numbers of observations to \( n \):
   ```stata
   set obs n
   ```
b. Set random number seed to \( XXXX \), default is 1000:
   ```stata
   set seed XXXX
   ```
c. Generate \( n \) uniform random variables (equal chance of all outcomes between 0 and 1):
   ```stata
   gen vname=uniform()
   ```
d. Generate \( n \) uniform random variables (equal chance of all outcomes between \( a \) and \( b \)):
   ```stata
   gen vname=a + (b - a)*uniform()
   ```
e. Generate \( n \) discrete uniform random variables (equal chance of all outcomes between 1 and 6)
   ```stata
   gen vname=1 + int(6*uniform())
   ```
   (These commands simulate rolling a six-sided die.)
f. Generate normal data with mean 0 and standard deviation 1:
   ```stata
   gen vname= invnorm(uniform())
   ```
g. Generate normal data with mean \( \mu \) and standard deviation \( \sigma \):
   ```stata
   gen vname= \mu + \sigma * invnorm(uniform())
   ```

Regression

1. Compute simple regression line (\( vy \) is response, \( vx \) is explanatory variable):
   ```stata
   regress vy vx
   ```
2. Compute predictions, create new variable \( yhat \):
   ```stata
   predict yhat
   ```
3. Produce scatter plot with regression line added:
   ```stata
   graph twoway lfit vy vx || scatter vy vx
   ```
4. Compute residuals, create new variable \( residuals \):
   ```stata
   predict residuals, resid
   ```
5. Produce a residual plot with horizontal line at 0:
   ```stata
   graph residuals, yline(0)
   ```
6. Identify points with largest and smallest residuals:
   ```stata
   sort residuals
   list in 1/5
   list in -5/l
   ```
   (The last command is "minus 5" / "lowercase letter L".)
7. Compute multiple regression equation (\( vy \) is response, \( vthree \), \( vtwo \), and \( vvthree \) are explanatory variables):
   ```stata
   regress vy vone vtwo vthree
   ```

Important Notes on the "\texttt{stem}" command
In some versions of Stata, there is a potential glitch with Stata's `stem` command for stem-and-leaf plots. The `stem` function seems to permanently reorder the data so that they are sorted according to the variable that the stem-and-leaf plot was plotted for. The best way to avoid this problem is to avoid doing any stem-and-leaf plots (do histograms instead). However, if you really want to do a stem-and-leaf plot you should always create a variable containing the original observation numbers (called `index`, for example). A command to do so is:

```stata
generate index = _n
```

If you do this, then you can re-sort the data after the stem-and-leaf plot according to the `index` variable:

```stata
sort index.
```

Then, the data are back in the original order.

### Summary of These and Other Commands

Here is a list of the commands demonstrated above and some other commands that you may find useful (this is by no means an exhaustive list of all Stata commands):

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>anova</code></td>
<td>general ANOVA, ANCOVA, or regression</td>
</tr>
<tr>
<td><code>by</code></td>
<td>repeat operation for categories of a variable</td>
</tr>
<tr>
<td><code>ci</code></td>
<td>confidence intervals for means</td>
</tr>
<tr>
<td><code>clear</code></td>
<td>clears previous dataset out of memory</td>
</tr>
<tr>
<td><code>correlate</code></td>
<td>correlation between variables</td>
</tr>
<tr>
<td><code>describe</code></td>
<td>briefly describes the data (# of obs, variable names, etc.)</td>
</tr>
<tr>
<td><code>diagplot</code></td>
<td>distribution diagnostic plots</td>
</tr>
<tr>
<td><code>drop</code></td>
<td>eliminate variables from memory</td>
</tr>
<tr>
<td><code>edit</code></td>
<td>better alternative to <code>input</code> for Macs</td>
</tr>
<tr>
<td><code>exit</code></td>
<td>leave Stata</td>
</tr>
<tr>
<td><code>generate</code></td>
<td>creates new variables (e.g., <code>generate years = last - first</code>)</td>
</tr>
<tr>
<td><code>graph</code></td>
<td>general graphing command (this command has many options)</td>
</tr>
<tr>
<td><code>help</code></td>
<td>online help</td>
</tr>
<tr>
<td><code>histogram</code></td>
<td>create a histogram graphic</td>
</tr>
<tr>
<td><code>if</code></td>
<td>lets you select a subset of observations (e.g., <code>list if radius &gt;= 3000</code>)</td>
</tr>
<tr>
<td><code>infile</code></td>
<td>read non-Stata-format dataset (ASCII or text file)</td>
</tr>
<tr>
<td><code>input</code></td>
<td>type in raw data</td>
</tr>
</tbody>
</table>

http://www.stat.uchicago.edu/~collins/resources/stata/stata-commands.html
<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>insheet</code></td>
<td>read non-Stata-format spreadsheet with variable names on first line</td>
</tr>
<tr>
<td><code>list</code></td>
<td>lists the whole dataset in memory (you can also list only certain variables)</td>
</tr>
<tr>
<td><code>log</code></td>
<td>save or print Stata output (except graphs)</td>
</tr>
<tr>
<td><code>lookup</code></td>
<td>keyword search of commands, often precursor to help</td>
</tr>
<tr>
<td><code>oneway</code></td>
<td>oneway analysis of variance</td>
</tr>
<tr>
<td><code>pcorr</code></td>
<td>partial correlation coefficients</td>
</tr>
<tr>
<td><code>plot</code></td>
<td>text-mode (crude) scatterplots</td>
</tr>
<tr>
<td><code>predict</code></td>
<td>calculated predicted values (y-hat), residuals (ordinary, standardized and studentized), leverages, Cook's distance, standard error of predicted individual y, standard error of predicted mean y, standard error of residual from regression</td>
</tr>
<tr>
<td><code>qnorm</code></td>
<td>create a normal quantile plot</td>
</tr>
<tr>
<td><code>regress</code></td>
<td>regression</td>
</tr>
<tr>
<td><code>replace</code></td>
<td>lets you change individual values of a variable</td>
</tr>
<tr>
<td><code>save</code></td>
<td>saves data and labels in a Stata-format dataset</td>
</tr>
<tr>
<td><code>scatter</code></td>
<td>create a scatter plot of two numerical variables</td>
</tr>
<tr>
<td><code>set</code></td>
<td>set Stata system parameters (e.g., obs and seed)</td>
</tr>
<tr>
<td><code>sebarr</code></td>
<td>standard error-bar chart</td>
</tr>
<tr>
<td><code>sort</code></td>
<td>sorts observations from smallest to largest</td>
</tr>
<tr>
<td><code>stem</code></td>
<td>stem and leaf display</td>
</tr>
<tr>
<td><code>summarize</code></td>
<td>produces summary statistics (# obs, mean, sd, min, max) (has a detail option)</td>
</tr>
<tr>
<td><code>tabstat</code></td>
<td>produces summary statistics of your choice</td>
</tr>
<tr>
<td><code>tabulate</code></td>
<td>produces counts/frequencies for categorical data</td>
</tr>
<tr>
<td><code>test</code></td>
<td>conducts various hypothesis tests (refers back to most recent model fit (e.g., <code>regress</code> or <code>anova</code>) (see help function for info and examples))</td>
</tr>
<tr>
<td><code>ttest</code></td>
<td>one and two-sample t-tests</td>
</tr>
<tr>
<td><code>use</code></td>
<td>retrieve previously saved Stata dataset</td>
</tr>
</tbody>
</table>