SOCY 7717: Event History Analysis and Sequence Analysis

Handout 2: Introduction to Stata

Wen Fan

Jan 17, 2019

1 A few general principles for data analysis

- Automate: (a) Automate everything that can be automated; (b) Write a single script that executes all code from beginning to end.
- Version Control: Store code and data under version control.
- Directories: (a) Separate directories by function; (b) Separate files into inputs and outputs; (c) Make directories portable.
- Documentation: (a) Don't write documentation you will not maintain; (b) Code should be self-documenting.
- Management: (a) Manage tasks with a task management system; (b) E-mail is NOT a task management system.

Source: summarized from Code and Data for the Social Sciences: A Practitioner's Guide by Matthew Gentzkow and Jesse M. Shapiro, 2014.

Along the same vein, I highly recommend Scott Long's book *The Workflow of Data Analysis Using Stata*, even if you do not use Stata. At the very least, you should check out Long's summary of his book.

2 Recommended do-file structure

Make a habit of beginning each do-file with the following:

```
// program:  _name_.do
// task:
// project:
// author:  _who_\_date_

*** 0.  program and data setup
version 14.1
clear all
set linesize 80
macro drop _all

cd "ENTER WORKING DIRECTORY HERE"
capture log close
log using _name_, replace text
using "DATA.dta", clear
```

Sometimes newer versions of Stata change the way in which a statistic is computed. Therefore different versions may produce different results.

The "`log`" command tells Stata to start a log file. Log files record everything that happens during a given session, including the commands you entered and the results you obtained.

```
*** 1.
// Description of task 1

*** 2.
// Description of task 2

log close
exit
```

3 *A toy example*

```
  .  cd "\\appsstorage.bc.edu\Desktop"
  \\appsstorage.bc.edu\Desktop
  .  use gss2014, clear
  .  desc
```

The "cd" command tells Stata what directory you are working in. The name of the current working directory is listed beneath the variable window. You can display the current working directory by entering the "pwd" command.

"use" tells Stata what file to open. Notice that if we had not named a working directory earlier, we'd have to specify the full path name.

"describe" summarizes the data set in memory.

```
Contains data from GSS2014.dta
  obs:         2,538
 vars:           875
 size:     2,510,082
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
abany           byte    %8.0g      LABA       abortion if woman wants for any
                                                reason
abdefect        byte    %8.0g      LABA       strong chance of serious defect
abhlth          byte    %8.0g      LABA       womans health seriously endangered
abnomore        byte    %8.0g      LABA       married--wants no more children
abpoor          byte    %8.0g      LABA       low income--cant afford more
                                                children
...
--------------------------------------------------------------------------------
```

```
  .  codebook joblose
```

The "codebook" command lists information about your variables.

```
--------------------------------------------------------------------------------
joblose                                                  is r likely to lose job
--------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  JOBLOSE

                 range:  [1,4]                        units:  1
```

```
         unique values:  4                        missing .:  0/2,538
       unique mv codes:  3                       missing .*:  1,513/2,538

            tabulation:  Freq.   Numeric  Label
                            37         1  very likely
                            57         2  fairly likely
                           289         3  not too likely
                           642         4  not likely
                             5        .d  dk
                         1,505        .i
                             3        .n  na

  .  note:  The General Social Survey (GSS) is a sociological
survey used to collect data on demographic characteristics
and attitudes of residents of the United States.
```

Attach a note to the data set.

```
. note _dta:
  1.  The General Social Survey (GSS) is a sociological survey used to collect
      data on demographic characteristics and attitudes of residents of the
      United States.

  .  sum age joblose
```

"summarize" prompts Stata to calculate descriptive statistics.

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      2,529    49.01265    17.41187         18         89
     joblose |      1,025    3.498537    .7605044          1          4

  .  tab sex

respondents |
        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      1,141       44.96       44.96
     female |      1,397       55.04      100.00
------------+-----------------------------------
      Total |      2,538      100.00
```

The first "tab" command creates a one-way frequency table. The second "tab" command cross-classifies `sex` and `joblose`. The "row" and "col" options can be used to obtain percentages within rows and columns, respectively.

```
  .  tab sex joblose

respondent |           is r likely to lose job
     s sex | very like  fairly li  not too l  not likel |     Total
-----------+--------------------------------------------+----------
      male |        17         31        133        324 |       505
    female |        20         26        156        318 |       520
-----------+--------------------------------------------+----------
     Total |        37         57        289        642 |     1,025
  .  tab sex joblose, row column chi2

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

respondent |           is r likely to lose job
     s sex | very like  fairly li  not too l  not likel |     Total
-----------+--------------------------------------------+----------
      male |        17         31        133        324 |       505
           |      3.37       6.14      26.34      64.16 |    100.00
           |     45.95      54.39      46.02      50.47 |     49.27
           |      1.66       3.02      12.98      31.61 |     49.27
-----------+--------------------------------------------+----------
    female |        20         26        156        318 |       520
           |      3.85       5.00      30.00      61.15 |    100.00
           |     54.05      45.61      53.98      49.53 |     50.73
           |      1.95       2.54      15.22      31.02 |     50.73
-----------+--------------------------------------------+----------
     Total |        37         57        289        642 |     1,025
           |      3.61       5.56      28.20      62.63 |    100.00
           |    100.00     100.00     100.00     100.00 |    100.00
           |      3.61       5.56      28.20      62.63 |    100.00

          Pearson chi2(3) =   2.3494   Pr = 0.503
```

```
  .  histogram age
```

Creates a histogram of age.

```
  .  kdensity age
  .  kdensity age if sex == 2, addplot(kdensity age if sex
== 1)
```

Generates kernel density plots reflecting the distribution of age in the GSS sample. The second command tells Stata to overlay two seperate density plots, where the first plot pertains to women (`if sex == 2`) and the second pertains to men (`if sex == 1`).

```
  .  graph box educ
  .  graph box educ, over(sex)
```

Generates box plots of years of schooling, first for the entire sample and then separately by gender.

```
  .  recode educ (0/15 = 0) (16/20 = 1) (.d = .d), gen(college)
```

Generate a binary indicator of college attainment. Call the new variable “college”.

```
. label var college "Indicator of college degree"
. tab college
. label define college 0 "0 < college" 1 "1 >= college"
. label values college college
. log close
```

Check to make sure the recode worked properly.

Attach value labels to the college variable.

4 Common Stata commands

4.1 Working with your data set

`log`: Begins a log file, which maintains a full record of the output that appears on the screen. The log is stored to your working directory. To save it to another location specify the full path. When you specify a file name, make sure to include the text option, otherwise your log will be saved in Stata's own markup language (.smcl).

`help`: Stata brings you instructions for a certain command. If you prefer, instructions can be displayed in the results window (as opposed to the pop-up viewer) by typing `chelp`.

`search`: Stata goes online to find help for a command.

`use`: Loads data into memory. The clear option erases all data currently being held in memory.

`merge`: Allows you to merge two or more datasets.

`set memory`: Tells Stata how much of the computer's memory to use.

`set scrollbufsize #`: Sets how far back you can scroll in the Results window ($10,000 \leq \# \leq 2,000,000$).

`clear`: Clears a dataset from memory.

`save`: Saves your dataset. The `replace` option indicates that if the file already exists, Stata should overwrite it.

`pwd`: Displays the current working directory.

`cd`: Changes the working directory.

4.2 Learning about the variables

`describe`: Lists some information about variables specified.

`lookfor`: Searches variable names and labels for specified words.

`list`: Displays the data for the observations.

`codebook`: Lists various information about the variables. You can add "`, compact`" in the end.

`count`: Counts the number of observations.

4.3 Examining distributions and values

`summarize`: Provides the mean, SD, and range. Using the `detail` option at the end of this command will provide additional information, including skewness, kurtosis, the four smallest and largest values, and various percentiles.

`tabulate`: Creates frequency table; can do cross tabulations with two variables.

`dotplot`: Draws a plot showing a quick graphical summary of a variable, useful when checking your data.

`histogram`: Creates a histogram.

`graph box`: Draws box plot.

`qnorm`: Draws a plot of the quantiles of a variable against the quantiles of a normal distribution.

`pnorm`: Draws a plot of the standardized normal probability plot.

4.4 *Creating and altering variables*

`generate`: Creates a new variables. Memorize the following operators:

- `+`: Add
- `-`: Subtract
- `*`: Multiply
- `/`: Divide
- `^`: Take to a power
- `ln()`: Natural log
- `exp()`: Exponential
- `sqrt()`: Square root
- `==`: Equal to
- `!=`: Not equal to
- `>`: Greater than
- `>=`: Greater than or equal to
- `<`: Less than
- `<=`: Less than or equal to
- `&`: And
- `|`: Or

`label variable`: Assigns a label for a variable. Variable labels appear in the right-hand column in the variable window.

`label define`: Creates a set of value labels. A value label is a way to assign meaningful information to numbers in your data. Certain qualitative information (such as marital status) may be stored in your dataset as a number, and a value label tells you what each number represents (e.g., 1 "Single" 2 "Married" 3 "Divorced").

`label values`: Attaches a value label to a variable.

`label drop`: Drop value labels.

`rename`: Assigns a new name to a variable.

`recode`: Changes some values of a variable to new ones.

`replace`: Replaces values of a variable.

`alpha`: Calculates inter-item correlations for a set of variables; can be used to generate a new variable which represents a scale formed from them.

`sort`: Arranges the observations according to ascending order of variables.

`by`: Tells Stata to do a command for every value of that variable.

`bysort`: Combines the `sort` and `by` command.

`keep`: Stata will keep only the variables or observations listed.

`drop`: Stata will drop all the variables or observations listed.

4.5 More advanced...

`ado uninstall`: Uninstalls user-written packages.

`dir *.dta`: Lists all files in your working directory with the extension `.dta`.

`global`: Associates a name with a string of characters or a number. It can then be accessed by any do-file or command until you either exit Stata or drop the macro from memory.

`local`: Associates a name with a string of characters or a number. It can be accessed only within the do-file in which it is defined.

`foreach`: Lets you execute a set of commands multiple times.

`forvalues`: Lets you execute a set of commands multiple times.

5 Some general tips for working with Stata

Largely based on Eric Grodsky's course notes, Sociology 8811 Advanced Social Statistics, Spring 2008, Department of Sociology, University of Minnesota.

1. Preserve a copy of the original data. When you first get your data, always save a copy and call it `original_data` or something similar so that you have an unaltered version of the data set. There will be occasions when you accidentally (and painfully) change or drop variables that you didn't intend to. Keeping a copy of the original data file will save you the trouble of trying to download them again. Similarly, use new names for new variables; you never

know when you'll need the original variable.

2. Stay organized. You will be surprised at how many files you create during this class and in your work outside of this class. You can minimize the clutter and, more importantly, save yourself (and others you work with) the trouble of trying to recreate previous work by using only one do-file, to which you simply add additional commands as your analyses progress.

3. Give new variables and new files new names. For example, if you collapse the categories of your income variable (`inctot`) give the resulting variable a new name (`inctot2`) and retain the original. Likewise, if you drop a handful of variables from your data set (`incanalysis.dta`), save the resulting data set under a new file name (`incanalysis2.dta`).

4. Long command lines. Occasionally you may write comments or command lines that are very long and span multiple lines in your do-files. Unless you tell it differently, Stata will assume that a command ends every time you hit "Enter". If you want to type a command or a comment that fills more than one line, there are a couple of options: (1) Stata ignores anything that comes after /// and treats the next line as continuation of the current one. So if you need to spread a command over, say, two lines, simply place /// at the end of the first line and then pick up where you left off on the line immediately below it. (2) Tell Stata that you plan to end each command or comment with a semi-colon using the command "`#delimit ;`". Once you execute this command (typically at the beginning of a do-file), Stata will assume that all subsequent commands end with a semi-colon. You can tell Stata to go back to its default by using the command "`#delimit cr`", where `cr` stands for carriage return. (see pp. 36-37 in Long and Freese).

5. Add notes to your dataset or variables. You have the ability to put comments in your dataset with the `note` command. Typing `note _dta: some comment` will add some comments to your dataset. You can add notes to specific variables by typing `note varname: some comment`. Display the notes with `note list`.