Introduction to Stata

14.74 Recitation Handout #1
February 4, 2005
Nancy Qian

This handout provides an introduction to Stata for Windows. You can access Stata in the Undergraduate Lab on the 3rd floor of E52. You need an economics department username and password for these machines. If you do not already have one from a previous class, please email me your Athena username and full name at nqian@mit.edu.

- **I. 3 components of your program**
- **II. Practice program**
- **III. Other Basic Commands and Syntax**
- **IV. Misc.**

Commands are in bold. So, if you want more details, you can look up that command name in the Stata manuals or Stata Help on the tool bar.

I. Components of Your Stata Program

a. .dta Files

i. This is the file containing your Stata format dataset.

ii. Let's say that our data set's name is "Sample.dta". In your .do file, you will tell Stata to use Sample.dta by typing

use sample, clear

"clear" tells Stata to erase the previous dataset.

iii. Stata also reads non-.dta formats with ***infile*** or ***insheet*** commands.

b. .do Files

i. These files are your programs.

ii. You write your programs in wordpad, textpad, or the Stata program editor. Just be sure to save your files with the *.do* extension.

iii. You can also type and run commands once at a time in the Stata command window. This is a good way to try commands and see what they do. However, for problem sets, you will have to write .do files. Do files allow you to save the commands your use and run them whenever you like.

iv. If you have a .do file written, you can run it by typing the following in the Stata command window:

do filename.do, clear

c. .log Files

i. This is your output file. If you are running a lot of commands at once, it is also a good way to review your program to see what each command did.

ii. At the beginning of your .do file, type

log using filename.log, replace

"replace" tells Stata to write over the previous log while whe you run the program anew.

iii. At the end of the .do file, type

log close

iv. You can open your .log file with Wordpad or Textpad.

II. Practice Program

(A dataset will be put up on the course website for you to practice running this program on. Some small modification in variables names may have to be made.)

a. Let's run a practice program and then discuss the Syntax

i. The dataset we are using is called practice.dta

ii. It is a dataset of wealth indicators for different countries.

iii. The variable name is in bold. Countries and continents are identified by numerical IDs.

Country	**Continent**	**GNP**	**GDP**	**Population**…….
1	1	3000	2000	10000000000
2	1	2000	1000	230000
3	2	1000	900	200005670
.				
.				
.				

iv. In the data set above it is important to note the difference between observations and variables. Each observation is a country, with many variables. It is very important to be clear about the structure of the dataset when you work with it.

b. The text enclosed in /* */ is to document what each command means. This will help you remember what you were trying to do when you come back to a program after not seeing it for a while

clear

*/*this clears the memory so that a new dataset may be inputted*/*

```
capture log close
/*this closes any log files that are still open*/

log using practice.log, replace
 /*this opens my log file*/

use sample, clear
 /*this tells Stata to use the dataset called “sample”*/

describe
 /*this lists all the variable names and their labels*/

sum
 /*this gives basic summary statistics for all the variables. You can also
type “sum GNP GDP” so that only statistics for those two variables are
produced */

sum, detail
 /* produces additional statistics including skewness, kurtosis, and the
four smallest and four largest values, along with various percentiles.*/

tab GDP
 /* this lists all the GDP values in the dataset and the frequency of
occurrence for each GDP value*/

tab1 GDP GNP
/* tab1 does what tab does for more than one variables*/

sort continent
/*sort the data according to continent. The data must be sorted before you
can use the by command*/

by continent: sum GDP GNP
/* this gives the mean and std deviation for the GDP and GNP of each
continent.*/

tab continent GDP, column
/*this produces a table of continent on the y axis and average GDP for
that continent on the X axis. “column” tells it to give you the
percentages*/

gen GDPPC=GDP/population
/*this generates a new variable called GDPPC that is the GDP per
capita*/

label var GDPPC “GDP per capita”
```

```
    /*creates an explanatory data label*/

    gen poor=0
    /*creates a new variable called poor and assigns it a 0 value*/

    replace poor=1 if GDPPC<400
    /*assigns the value 1 to poor for every observation where GDP per
    capita<400*/
    /*the variable "poor" is called a dummy variable*/

    sum GDPPC poor

    tab continent poor, column
    /*creates a table that has continent on the y-axis and the frequency and
    percentagel of poor on the x-axis*/

    tab poor continent, column
    /*how is the interpretation different?*/

    corr GDP GNP
    /* computes the correlation between GDP and GNP*/

    cov GDP GNP
    /*computes the covariance*/

    reg GDP population
    /* regresses GDP on population—computes Ordinalry Least Squares*/

    predict yhat
    /*creates a variable yhat that contains the predicted values of GDP based
    on the regression just run*/

    predict xhat, resid
    /*creates a variable xhat that contains the residual—(y-yhat)-based on the
    regression just run.)

log close
/*closes the log file*/
```

c. Open your log file in Wordpad to view the results.
d. As you can tell, Stata does not need any notation to end each command line or to end the program.

III. Other frequently commands

Stata has a very easy to use help section that you access from the toolbar. For more details about these commands, just type the command in the search box.

When trying the following commands, keep in mind that: *keep* is different from *keep if* b/c observations are different from variables. Open the data editor to look at the data after each command to see how each command works.

```
keep GDP GNP
/* throws out all the variables except GDP and GNP for all observations*/

use sample, clear
/*reads in the full original dataset */

keep if GDP>10000
/*throws out  observations that have GDP<=10000*/

use sample, clear

drop GDP
/*drops the variable GDP for all observations*/

use sample, clear

drop if GNP <=900
/*drops the observations which have GNP<=900*/
```