

**Stata Lesson**  
**Thursday February 15, 2006**

**[1] Where to find the data sets**

<http://socs.berkeley.edu/~olney/spring06/econ154>

There are four versions of the dataset there: txt, excel, stata, and sas transport file. Download all four to your desktop.

**[1.1] To transfer an excel data set to stata format**

start translate program  
choose “input file type” format (here, excel or SAS transport)  
browse to find the data set  
choose “output file type” format (stata)  
it will automatically suggest a file name  
choose “all variables” and “all cases”  
select “transfer”

**[1.2] To open the data in txt (ascii) format**

start stata

know where your data set is

type

insheet using “*path*\io01.txt”

**[1.3] To read in data that are in fixed format**

Read the manual!

Use infix

Create a “dictionary file” that specifies the locations of the variables and their new names

## [2] To save your work

Create a log file:

File - Log - Begin

(or, click on the icon that looks like a scroll)

Choose either formatted log file (\*.scml; only readable within stata)

Or text log file (\*.log; readable within any word processor)

## [3] To look at the data

describe

list (Careful, this will go on and on; use red X to stop)

“Tabulate” tells you how many different occurrences there are of each observed value of a variable; useful for variables with just a few possible values; not useful for continuous variables. Including two variables gives you a cross-tab table.

tab *variable-name* (gives a list of all the values and their frequencies)

e.g.

tab occup How many teachers, principals, superintendents?

tab occup sex By gender, how many teachers, principals, superintendents?

“Summarize” tells you a variety of summary statistics.

summarize *variable-name*

or sum *variable-name* (gives brief summary statistics)

sum *variable-name*, detail (gives longer list of summary statistics)

### [3.1] To look at the dataset sorted by some variable; two steps

Step one

sort *variable-name*

e.g.

sort sex

Step two

by *variable-name*: sum *variable name*

e.g.

by sex: sum totear

#### [4] To start naming the variables

label variable *variable-name* “*variable label*”

e.g.

Label variable pob “Place of Birth”

#### [5] To name the values of variables; two steps

Step one:

label define *labelname value* “*name*” *value* “*name*”

e.g.

Label define sexlabel 1 “male” 2 “female”

Step two:

label values *variablename labelname*

e.g.

Label values sex sexlabel

#### [6] To create new variables

generate *newvariable* = *some function of existing variables*

gen *newvariable* = *some function of existing variables*

e.g.

Gen num\_months\_paid = totearn / earnmo

#### [7] To create dummy variables

generate *dummyvar* = (*existing variable* == *value for dummy to equal 1*)

e.g.

generate male = (sex == 1)

#### [7.5] To replace values of existing variables

replace *varname* = *newvalue* if *varname* == *oldvalue*

e.g.

gen female = sex

replace female = 0 if female == 1

replace female = 1 if female == 2

## [7.6] To replace missing values of existing variables

replace *varname* = . if *varname* == *missingvalue*

e.g.

Replace age = . if age == -9

## [8] To graph the data

Click “graphics” along the top menu

Choose the type of graphs (start with “easy graphs”)

Choose a particular graph (try “scatter plot” to start)

Specify x (horizontal) axis variable and y (vertical) axis variables

Specify any “if” restrictions

Click “submit” and the graph will pop up in a couple of seconds

Edit your specifications as you wish; click “submit” again

when you have it as you like it, **save your graph** by right-clicking anywhere on graph

## [9] To run a linear regression

regress *depvariable independentvariables*

e.g.

Regress totear male

## [10] To run a probit

Probit *depvariable independentvariables*

e.g.

Generate ownhome = (ownhm == 2)  
Probit ownhome male totear age

To get probit results that are directly interpretable

Dprobit *depvariable independentvariables*

e.g.

Dprobit ownhome male totear age

**[11] To include only part of the data set, add if statements**

e.g. restrict probit to include only teachers:  
Dprobit ownhome male totear age if occ == 1

**[12] Creating interaction variables**

Suppose you think that gender matters, but not as a constant shift factor. You think instead that the effect of gender is to alter the return to college education. In that case you need to create “an interaction term”

*Gen male\_grad = male \* grad*

Now include this as a variable in your regression, along with just “grad.” The coefficient on “male\_grad” tells you if men have a different return to being a college grad than women do.

**[13] Now, let’s play with the data set** (see also data exercise page on class webpage)

Estimate a relationship between total earnings and its determining variables. Think first about:

- a) how will you take age into account
- b) does gender matter? Does it shift the earnings function, or change some of the returns, or both?
- c) do you want separate equations for teachers, superintendents, principals, or do you just want a shift factor for different occupations?

**[14] “do” files**

When you have a lot of commands, write a “do” file which contains all of the commands and then have stata run the “do” file

Save the “do” file in ascii format, with the suffix .do. For instance “monday.do”

Then type

do monday, nostop

(Nostop is an option that tells stata to keep going even if it encounters an error)

**[x] A good place to look for help, data sets to play with, do files, etc.**

[www.aw.com/stock\\_watson](http://www.aw.com/stock_watson)