

Supplement to Tutorial #1: Using STATA with Your Data

Frank Stafford and Matt Wyble, December 2008

This PSID user guide supplement provides a STATA version of the Excel-based User Guide Tutorial #1 (Cross Sectional Files.) This guide assumes that the user is using STATA 10 and that they have selected "ASCII Data with Stata Statements" as their Data Output Type during checkout. All procedures prior to checkout are exactly the same for both Microsoft Excel and STATA.

You will see a notice indicating that the Data Center is creating your dataset, and when the task is finished a note appears indicating that your dataset has been sent to your e-mail address. When retrieving the dataset, PC users should right click (other users may be required to use alternatives to right click) on the blue text of both STATA Statements and Rectangular ASCII Text Data, each time selecting either "Save Target As" (in Internet Explorer) or "Save Link As" (in Firefox). Make sure to save both files to an accessible and relevant folder. The file will be transferred to you.

Before you can load the data, you must edit the .do file to include the location of your .txt data file that you also downloaded. You can do this in two ways. The first is to open up Notepad (or a similar basic text editor), and then open the new .do file from the file menu within Notepad. You can also edit the file in STATA itself by typing the following command: `doedit "{path}\{filename}.do"` where {path} and {filename} are replaced with the appropriate values.

Once the file is open, scroll down to the line with the following text where XXXX is replaced with the filename of your Rectangular ASCII Text Data:

```
using [path]\XXXX.txt, clear
```

Replace [path] with the full path to the Rectangular ASCII Text Data file you saved to your computer. For instance, if you saved the file in your default user's My Documents folder, the location might be

```
C:\Documents and Settings\Default User\My Documents
```

Note that if your file path or filename contain any spaces, you will need to put quotation marks around the entire path and filename in order for it to load correctly. Once you have completed these edits, save and close the .do file. Next, double click on the newly edited .do file to load the downloaded data.

Once the data file has loaded in STATA, we want to see how many of the women in the study are 65 years old or younger. To do this, we type the following command into STATA:

```
count if ER33504<=65
```

This will return a value of 4994, indicating that 4994 women in the sample are 65 or younger. Note that the variable name that STATA uses is ER33504, which is the variable name on the PSID. A similar command with the other variables will produce a chart similar to the one below.

Table 2. Case Counts from Specific Data Partitions

N=5725: All females (head or wife only) and living in the FU at the time of 1999 interview.
--

Age <= 65	N=4994		
		0 <= Hours <= 112	Weight > 0
	0 <= Hours <= 112	4940	
	Weight > 0	2754	2784
0 <= Hours <= 112	N=5655		
		Age <= 65	Weight > 0
	Age <= 65	4940	
	Weight > 0	2754	3421
Weight > 0	N=3465		
		Age <= 65	0 <= Hours <= 112
	Age <= 65	2784	
	0 <= Hours <= 112	2754	3421

II. Using Stata 10 on your Output Subset

1. In order to make the variables easier to differentiate, you can rename them by right clicking on them and choosing the rename function.
2. In order to get an idea of the types of responses to a question, consider using this function: summarize [variablename], detail . It gives a set of basic metrics on a variable.
3. To look at raw data, navigate to Data>Data Browser (or Data Editor if necessary.)
4. One benefit to STATA is the ease with which one can create dummy variables for regressions. In this example, run the following two commands:
generate float male = 1 if ER13011==1
replace male=0 if ER13011==2
This command creates a variable called male that has two possible values: 0 if the respondent is female and 1 if the respondent is male. It is generally a good idea to name the variable such that a value of 1 equals the state described by the variable name.
5. It may be necessary to exclude a particular response value from a regression. For instance, you should remove values such as 99 when it is used to signify an inapplicable question. Consult the codebook for examples of this. In our example, variable ER33504 contains some problematic responses in the form of 999 used to signify that the question was inapplicable or that the respondent did not know his or her age. To remove such responses, run the following command:

drop if ER33504==999

This will remove all the unusable data points for the regression. The data points dropped are not

permanently removed from the data set unless it is saved before closing.

6. Interpreting basic regressions:

One of the advantages of using Stata is the ease with which one can run basic regressions. To run a linear regression, use the following command:

Reg [dependent variable] [independent variable 1] [independent variable 2] [independent variable 3] [etc.]

For example, let's examine the effect of the number of children and the age of the wife on the average number of housework hours per week.

reg ER14230 ER13013 ER33504

Here is the output of this regression:

Source		SS	df	MS	Number of obs	=	5725
Model		9845.2032	2	4922.6016	F(2, 5722)	=	0.45
Residual		62449313.8	5722	10913.8961	Prob > F	=	0.6370
Total		62459159	5724	10911.8028	R-squared	=	0.0002
					Adj R-squared	=	-0.0002
					Root MSE	=	104.47
ER14230		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ER13013		.6488269	1.15983	0.56	0.576	-1.62488	2.922534
ER33504		.0536022	.0602521	0.89	0.374	-.0645146	.1717191
_cons		17.29382	3.521887	4.91	0.000	10.38959	24.19806

Consult a statistics guide to determine the meaning of each number included in this output. Note that output tables such as this must use a "fixed width" font (such as Courier New) if copied into another document—otherwise the document spacing will be inconsistent and hard to read.